Exploring the genomic complexity of bacterial infection in 3D

Cyril Matthey-Doret

Sorbonne Université

École doctorale Complexité du Vivant

Institut Pasteur | SORBONNE UNIVERSITÉ

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: **Romain Koszul**

Présentée et soutenue publiquement le 16 décembre 2021

Jury

Angela Taddei	Directrice de recherche	Présidente
Francesco Ferrari	Directeur de recherche	Rapporteur
Matthieu Legendre	Chargé de recherche	Rapporteur
Matthias Horn	Professeur	Examinateur
Olivier Espeli	Directeur de recherche	Examinateur
Romain Koszul	Directeur de recherche	Directeur de thèse

Exploring the genomic complexity of bacterial infection in 3D

Thèse de doctorat en Bioinformatique

Rapporteurs: Francesco Ferrari et Matthieu Legendre

Examinateurs: Matthias Horn et Olivier Espeli

Présidente: Angela Taddei

Directeur de thèse: 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 another 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 of many studies aiming to understand their "hijacking" mechanisms. Recent advances 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 "effector" proteins into the host cytoplasm which interact with its proteins and affect its genetic expression program. Recently, studies in *Legionella pneumophila*, an experimental model for intracellular bacteria, have shown it was able to alter the epigenetic state of its host. Such modifications allow rapid physiological changes and are intimately linked to the spatial organisation of the genome. 3D genome organisation plays an important part in many biological processes, for example by modulating gene expression through long range interactions in the sequence. Throughout this work, we develop computational tools to explore and measure spatial changes occuring in the genome, and exploit them to investigate the changes taking place during infection by intracellular bacteria. 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 de 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 utilisant ce mode de fonctionnement et font l'objet de nombreuses études qui visent à comprendre leurs mécanismes de "piratage". Les récents progrès en génomique ont largement stimulé la recherche dans ce domaine en offrant la possibilité de déchiffrer la séquence des gènes exprimés durant l'infection. Plusieurs 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 d'altérer la régulation épigénétique de son hôte. Ce genre de modifications permet des changements physiologiques rapides et est intimement lié à l'organisation spatiale du génome. L'organisation 3D du génome joue un rôle important dans de nombreux processus biologiques, par exemple en modulant l'expression génétique par la formation d'interactions entre des éléments éloignés dans la séquence d'ADN. A travers ce travail, nous développons des outils computationels pour explorer et mesurer les changements spatiaux du génome, et nous les exploitons pour étudier les changements qui ont lieu pendant l'infection par des bactéries intracellulaires. Nous utilisons en particulier les modèles *Salmonella enterica* et *Legionella pneumophila* pour explorer les changements de structure qui surviennent dans les chromosomes de leurs hôtes et leurs liens avec l'expression génétique.

Acknowledgements

First, I would like to thank Romain Koszul, my thesis supervisor, to whom I am very grateful for entrusting me with several projects and giving me the freedom to explore my scientific interests during these three years. This has helped me evolve and allowed me to discover a variety of research questions as well as supervise students and take part in exciting collaborations. I would also like to thank members of my PhD committee and reviewers for taking the time to review the present manuscript and giving me precious advice.

Thanks to all members of the RSG lab for the exceptional atmosphere and the great times we enjoyed despite the global pandemic which changed our lives for a good part of my stay in the lab. Special mentions to Axel, Axel and Lyam for all the stimulating whiteboard discussions and long debates taking place during Chromosight's development, Pierrick and Agnès for their motivation and spreading good humor the lab, Nadège for frequently cheering me up with geeky jokes and moral support as a fellow PhD student, Charlie for entrusting me with a pipette, Théo for all these interesting late night discussions, Amaury and Jacques for being such a pleasure to work with and inspiring good computational practices in the lab.

Thanks to all our collaborators, especially Carmen and Pedro who taught me a great deal about the biology of *Legionella*.

J'aimerais aussi remercier mes parents pour leur soutien et leur support tout au long de cette période. 感谢家人在这个困难时期欢迎我进屋。 Finalement, merci Fleur d'avoir rendu ces trois années aussi mémorables et de m'avoir supporté en toutes circonstances.

Ce manuscript est dédié à mon grand papa, qui a éveillé mon intérêt pour la science et a toujours stimulé ma curiosité.

Contents

List of Figures

List of Tables

Abbreviations

3C Chromosome Conformation Capture 14

API Application Programming Interface 34, 37

BAC Bacterial Artificial Chromosome 25

BMM Bone Marrow Macrophage 116

ChIPseq Chromatin Immuno-Precipitation Sequencing 12

CLI Command Line Interface 34, 37

CNR Contrast-to-Noise Ratio 70–72

CTCF CCCTC-binding binding factor 16

ER Endoplasmic Reticulum 7

HGT Horizontal Gene Transfer 4, 5, 24, 28

indel Insertion / Deletion xi, *Glossary:* Insertion / Deletion

LCV *Legionella* Containing Vacuole 5, 6

NGS Next Generation Sequencing 25, 33

o/e observed over expected 17, 19

PCA Principal Component Analysis 17

PCR Polymerase Chain Reaction 14

PFGE Pulsed Field Gel Electrophoresis 10

RFLP Restriction Fragment Length Polymorphism 10

SCV *Salmonella* Containing Vacuole 8

SMC Structural Maintenance of Chromosomes 16

SNP Single Nucleotide Polymorphism xii, *Glossary:* Single Nucleotide Polymorphism

SV Structural variant xii, *Glossary:* Structural variant

TAD Topologically Associating Domain 13, 17, 19, 20, 22

WGS Whole Genome Sequencing 11

Glossary

chromatin An association of DNA and various DNA-binding proteins forming chromosomes. 13

contig Contiguous subsequence of a DNA molecule generated by combining multiple reads. 26

fitness The reproductive success of an organism or individual. It is equal to the average contribution to the gene pool of the next generation of the population. 2

heterochromatin The inactive portion of chromatin 13

Insertion / Deletion Short mutations in a DNA sequence consisting of a few added or missing nucleotides. xi

k-mer A sequence of length k molecules. For DNA, there are 4^k possible k-mers. 25

Muller's ratchet The irreversible accumulation of slightly deleterious mutations in genomes with lack of recombination. 3

parthenogenesis Mode of reproduction where females generate offspring asexually. 2

polishing The use of short accurate reads to correct a genome assembly generated with error prone long reads. 26

read Contiguous subsequence of a DNA molecule as read by a sequencing technology. 25

restriction enzyme Protein that cleaves DNA at specific recognition sites. 10

Single Nucleotide Polymorphism Punctual mutations in a DNA sequence consisting of a nucleotide substitution. xii

Structural variant Large scale alterations in a genome, including deletions, insertions and insertions xii

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 and 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 other species. Such 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 are often classified according to their perceived impact on the *Fitness* of their members: we traditionally refer to parasitism for interactions with one-way benefits, and to 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 regarded as a continuum, depending on the fitness cost and benefit of the relationship to the host (Fig. I.A).

These biological interactions shape the evolutionary trajectories and genomic landscapes of the species involved. 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, benefiting the bacterium which can only be transmitted through females. In some species, infection by *Wolbachia* has even become essential to reproduction. While the bacterium takes advantage of its host reproduction, it also provides numerous advantages such as

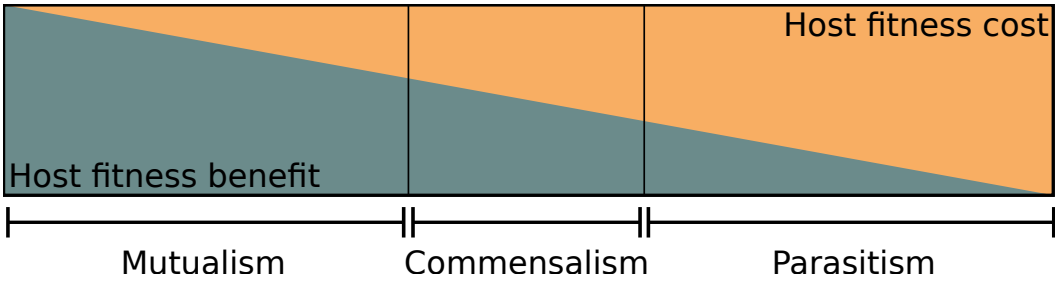


Fig. I.A: Parasitism - Mutualism spectrum: A spectrum of host fitness cost underlies common terms used to described a biological interaction.

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 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 to recombine 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 is under constant selective pressure and must evolve novel strategies (i.e. weapons) to improve its own fitness at the expense of the other, a manifestation of the Red Queen hypothesis[1] [9, 10]. This is the case for intracellular bacteria such as *Legionella* or *Salmonella*, which secrete a large arsenal of effector proteins into their host's cytoplasm. These proteins manipulate host signalling and metabolic 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 [11].

Perfectly redundant genes within the bacterial genome should be subject to low selective pressures, making them susceptible to genetic drift and therefore unstable

[1] The hypothesis states: "For an evolutionary system, continuing development is needed just in order to maintain its fitness relative to the systems it is co-evolving with". It is named after the quote from Lewis Carroll's novel "Through the Looking Glass": "It takes all the running you can do, to keep in the same place".

[12]. 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 [11]. Selective pressure would therefore be applied on these specific characters. This is likely an important phenomenon for parasites with a broad range of hosts, or encountering multiple 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 involved in HGT at some point in their history [13]. More recently, HGT from bacteria to eukaryotes have also been detected in eukaryotic genomes. Although much less frequent (0.04-6.49% of genes in microbial eukaryotes [14]), 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 [14].

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 Amoebae as a host model

Free living amoebae are ubiquitous unicellular organisms found in soil and various bodies of water, such as rivers, lakes [15] or even puddles [16]. 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.

Amoebae offer a great experimental model, as many species are relatively 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 amoebae remain still largely unknown. For example the *Acanthamoeba castellanii* genome has evidence for highly variable ploidy levels [17] and horizontally acquired genes [18]. These peculiar genomic features are likely important in their interactions with endosymbionts. For instance, high ploidy levels have been proposed as a mean for asexual amoebae to escape Muller's ratchet through homologous recombination between haplotypes [17].

Similarly, the amoeba *Paulinella chromatophora* has photosynthetic organelles whose genome benefits from HGT from endosymbionts, as they counteract Muller's ratchet. This exciting observation provides an interesting track to investigate the conservation of horizontal gene transfers in the genome of free living amoeba *A. castellanii* [18].

Their long coevolution with endosymbionts make free living amoebae an interesting model for evolutionary biology and ecology. In addition, they are also highly relevant to public health concerns, as they are the reservoir of several human pathogens such as *Legionella pneumophila*. Besides, many free living amoebae 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 a public health standpoint, since intracellular bacteria infecting amoebae are able to survive water chlorination or antibiotic treatments using the encysted amoebae as shelters.

1.3 *Legionella pneumophila*

L. pneumophila is an important model for studying intracellular bacteria. It infects a range of 15 species of amoebae and ciliated protozoa in the wild [19], and can also infect lung macrophages of humans and other mammalians. In humans, this can cause a severe pneumonia known as *Legionnaire's disease* [20]. Human to human transmission of *L. pneumophila* is extremely rare [21], 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 resulting in 182 cases, 29 of them fatal. Since then, outbreaks are associated to *Legionella* every year with over 32,000 cases reported between 1995 and 2005 [22].

Unlike other bacteria on which phagocytic cells prey (Fig. I.Ba), when engulfed by a predatory cell *Legionella* evades the lysosomal degradation pathway and survives in a special vacuole, the *Legionella* Containing Vacuole (LCV) (Fig. I.Bb). It does so by using a type IV secretion system to secrete ∼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 [23]. Through their secretion, the bacterium is able to create a niche inside of the host cell with stable conditions and ample 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 either spread planktonically as a free living organism using its flagella to reach new hosts, or by associating with biofilms [24, 25]. This extracellular phase is called the "transmissive form", as bacteria will search for new host cells but will not replicate [26]. In contrast, when entering a host cell, the bacterium enters the "replicative form". In that stage, 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 [25]. 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* in 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 [27]. 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 mostly by nutrients present in the environment [28].

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 [29].

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 [27]. Those nutrients are transferred from the host cytoplasm to the LCV by transporters on the LCV membrane [30]. 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 [31]. Other effectors block host protein translation to increase the pool of free amino acid available for consumption by *L. pneumophila* [32].

The host trafficking system is also hijacked, resulting in the recruitment 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 [33]. Some *Legionella* effectors directly affect the host actin cytoskeleton, which is important in many cellular processes including vesicle trafficking [34, 35].

L. pneumophila also ensures successful infection by promoting host cell survival. The effector SdhA interferes with host cell apoptosis by inhibiting caspases [36]. All these interference with the host cell signalling pathways are likely bound to affect its expression program. 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 [37].

There is still much to learn about the interaction between *Legionella* effectors and its host regulation, but that the bacteria is able to modify directly nucleosomes of the host unveiled a new level of intimacy between bacterial endosymbionts and their host, with fascinating perspectives. Besides, epigenetics and gene expression are tightly connected with spatial genome organization in eukaryotes [38, 39], providing a new angle to approach the study of host-pathogen interactions.

1.4 *Salmonella enterica*

Unlike *L. pneumophila*, *S. enterica* infects not only mammals but also birds and reptiles [40]. 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 [41].

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 [42]. Most of the current knowledge on *Salmonella* infection biology was built on the non-typhoidal serovar *S. enterica* subsp. *enterica* serovar Typhimurium [41]. 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 [43].

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 [41]. Although no effector of *Salmonella* is known to directly affect the host epigenetic state, a global rewriting of histone modifications and DNA methylation [44, 45] is observed in *Salmonella*-infected cells. Furthermore, histone modifications was associated with susceptibility to Salmonella infection in chickens [46], further highlighting the importance of investigating chromatin changes during bacterial infection.

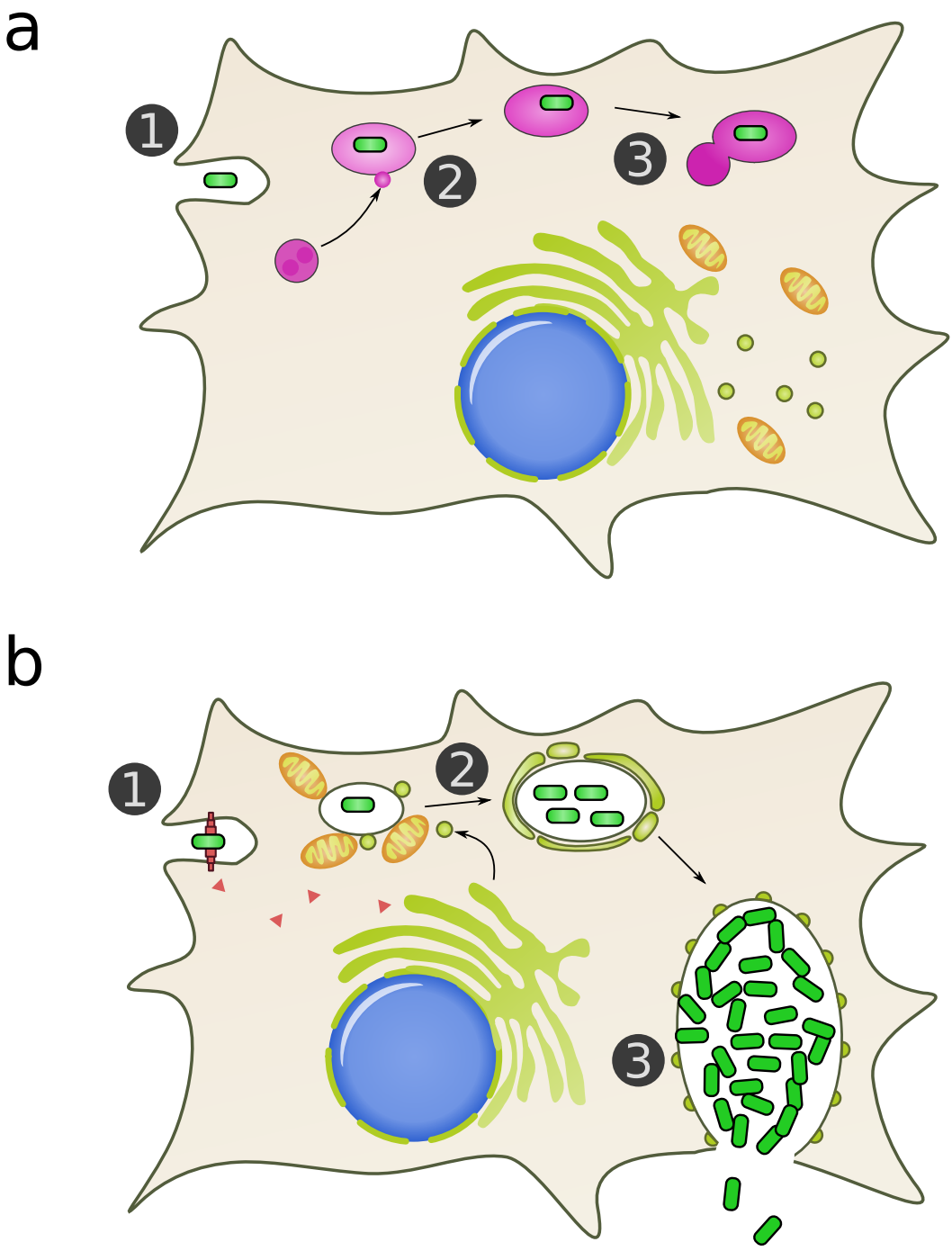


Fig. I.B: Infection by *Legionella*: **a** Non infectious bacteria (green) are phagocytized by amoebae or macrophages (1), the early and late endosomes (pink) acidify the compartment (2), and it finally merges with the lysozome (3) where the bacteria is degraded. **b** Upon phagocytosis, *Legionella* 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

Discoveries in biology are typically associated with progresses and advances in technological development, and the toolbox to detect and investigate bacterial infection traditionally included biochemical assays and microscopy. The recent advances in DNA sequencing have spurred a rapid extension of sequencing-derived, genomewide methods. Here we introduce the different ways these genomics approaches can provide biological insights into the biology of bacterial pathogens.

2.1 Pathogen characterization

A key task related to infection in biomedical research is the detection and characterization of infectious agents. This is of public health relevance, as it allows testing patients who present suspicious symptoms for the presence of known pathogens, or determine the pathogenicity of a particular strain.

Genotyping, i.e. the determination of one sample's genetic make-up from DNA analysis, can be achieved using molecular biology techniques such as Restriction Fragment Length Polymorphism (RFLP) or Pulsed Field Gel Electrophoresis (PFGE) [47]. These techniques use gel electrophoresis, which relies on the negative charges carried by DNA molecules. When put in a polymer gel submitted to an electromagnetic field, these acidic molecules migrate along the electrical current towards the positive pole of the field. The migration distance depends on the density of the gel polymer meshwork that impairs progression of DNA, and is proportional to the size of DNA molecules. After migration is complete, the gel can be treated with chemical reagent such as ethidium bromide, a fluorescent agent that intercalates between bases pairs. These treatments allow highlighting the position of DNA molecules. A large DNA molecule can be fragmented using a *Restriction enzyme* into smaller segments of sizes able to migrate in the gel. Those will generate discrete bands of similar-length DNA fragments once revealed by chemical reagents. Together these bands form a bar-code of the larger molecule, and can be interpreted by the scientist to draw conclusions about its presence or nature. In the case of RFLP, the entire genome is digested by restriction enzymes beforehand. The digestion will result in a series of discrete fragments whose lengths can be seen on the gel. Bacterial genotypes 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. In theory, Whole Genome Sequencing (WGS), the process of determining the nucleotide sequence of an entire genome at a single time, provides accurate information on an organism's nucleotidic or structural polymorphisms compared to the genome sequences of related strains, 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 relatively 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. Multiple levels of regulation are affected upon infection, from signalling to epigenetic modifications [48]. Over the years, a vast arsenal of NGS techniques has been developed to characterize and investigate these regulatory states.

The most widely used approach consists in gene expression analysis (RNA-seq). The total transcribed RNAs present in a biological sample made of cells (infected or not) can be extracted, and reverse-transcribed into cDNA. That cDNA can be then sequenced and the relative abundance of each gene's transcript determined. This allows the quantification of the expression of all genes in the genome, known as the transcriptome. Typical transcriptome analysis consists in comparing different conditions, to find out which genes undergo perturbations (increase or decrease of expression levels) during infection.

Many levels of regulation allow eukaryotes to fine tune their gene expression (Fig. I.C). Regulatory elements encoded in the sequence, such as enhancers, can trigger or facilitate the recruitment of protein complexes to modulate gene expression [49]. Regulation can also apply at the post-translational level, for example by degrading proteins [50] or applying chemical modifications such as phosphorylation or acetylation to modulate their activity [51]. Epigenetic changes, in the form of chemical modification of histone proteins offer yet another way to regulate gene expression in eukaryotes. These chemical modifications are thought to collectively form a "histone code" [52] - a combinatorial set of instructions dictating the regulatory state of DNA sequences. Although the role of many histone modifications is still partially or completely unknown, there are many examples of histone marks affecting chromatin

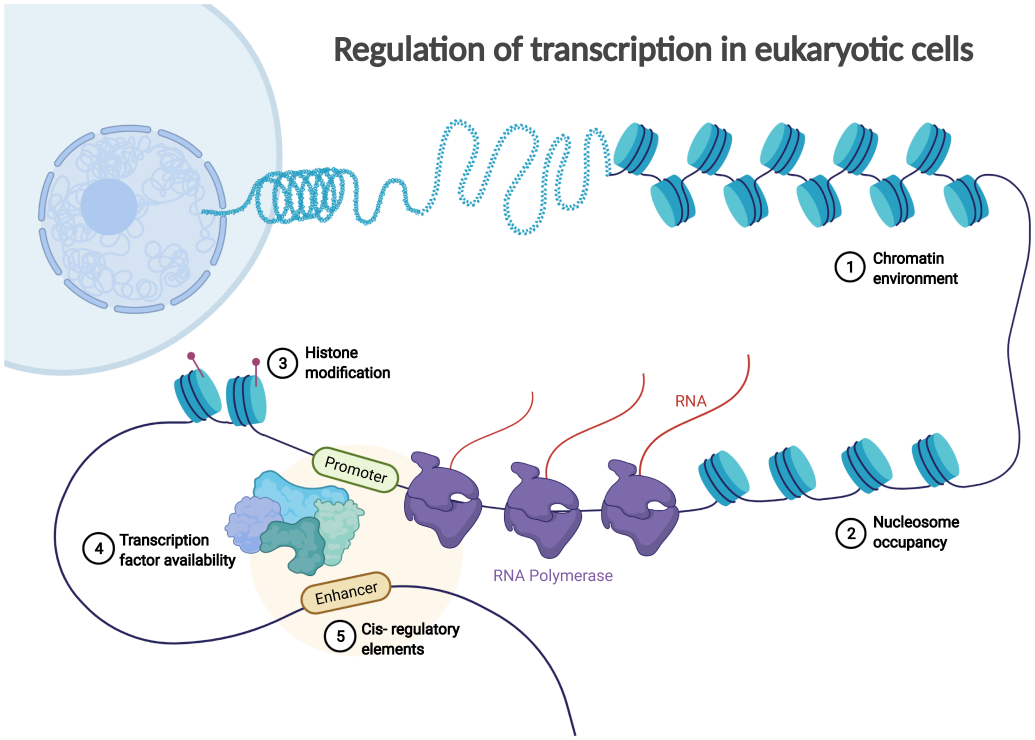


Fig. I.C: Regulation of transcription in eukaryotic cells. Visual summary of the different levels at which transcription can be regulated. At the largest scale (1), the chromatin environment can form structures affecting transcription. The open space between nucleosome can also affect accessibility of 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 availability of those factors (4) and the proximity of regulatory sequences such as enhancers (5) provide another level of transcriptional regulation. Reprinted from "Regulation of Transcription in Eukaryotic Cells", by BioRender.com (2020). Retrieved from https://app.biorender.com/biorender-templates

structure [53, 54] and gene transcription (Tab. I.A). The amount of epigenetic marks can be quantified along the genome using another NGS-derived technique known as Chromatin Immuno-Precipitation Sequencing (ChIPseq). In ChIPseq, the chromatin sample is crosslinked with formaldehyde, a fixative molecule that will generate covalent bonds between proteins and DNA. The sample is then sonicated to break the DNA into smaller fragments. Beads coated with antibodies targeted against a protein of interest (e.g. an epigenetic mark) are then used to precipitate and isolate DNA molecules bound to the protein of interest from the pool of total DNA. The crosslink is then reversed and the DNA fragments purified. This allows one to retrieve all genomic regions that were bound to the protein of interest.

Mark	Regulatory state	Type	Sources
H3K4Me1	primed enhancers	active	[55]
H3K4Me3	active promoters	active	[56, 57]
H3K9Me2	facultative heterochromatin	repressive	[58]
H3K9Me3	constitutive heterochromatin	repressive	[59]
H3K27Me3	repressed genes	repressive	[60]
H3K27Ac	active enhancers	active	[61]
H3K36Me3	transcribed gene bodies	active	[62]

Tab. I.A: Examples of commonly studied histone modifications of histone subunit 3 and their impact on transcriptional regulation.

2.3 Capturing chromosome conformation

Most eukaryotic chromosomes are made of a linear DNA molecule which is not randomly organized into the nuclear space. This long polymer can fold back on itself, resulting in three-dimensional structures which have several useful properties. One of the of most obvious, direct advantages of folding consists in compactness: for example, the human chromosome 1 is made of 250 millions nucleotides, each spaced by 0.34nm [63]. If straightened, the chromosome would be 85mm long, yet the whole genome fits into a nucleus of $\sim$10$\mu\mathrm{m}$ in diameter. Another benefit of genome folding lies in its potential to contribute to the regulation of gene expression through the formation of multi-scale structures [64]. Compacting large regions of the genome by spreading of *Heterochromatin* can repress their activity [65]. Smaller scale structures, such as chromatin loops, appear also involved in the fine tuning of gene regulation [64]. For example, it is suspected that such *Chromatin* loops play a role in bridging enhancers and promoters, even though these sequences can be separated by large genomic distances [66–68]. Chromatin can also organize into compact self-interacting neighbourhoods forming local "domains", with distinct domains being isolated from each other. Here too, the significance of such local structures remains relatively elusive, and is being investigated in a number of Eukaryotic species. In mammals, Topologically Associating Domain (TAD)s are also associated with large scale, cohesin-dependent loops [69], while in other species such as the budding yeast and fruit fly, self-interacting domains could be more delimited by supercoiling, as they display highly expressed genes at their extremities [70, 71]. Transcription *per se* could therefore be a direct player of chromosome folding, but the reciprocal interplay between transcription and folding remains also elusive and investigated. Nevertheless, characterizing the regulation of these different levels of chromosome folding is an essential step towards understanding their potential interplay with

chromosome function, including the coordination of the gene expression program with other cellular processes.

2.3.1 3C technologies

The use of genomics to investigate the 3D folding of genomes started with the invention of the Chromosome Conformation Capture (3C) technique [72]. This technique allowed researchers to quantify the relative frequencies of physical contacts between pairs of DNA segments in a genome (Fig. I.D, left). This is done by crosslinking the genome with formaldehyde, a small chemical fixative molecule that 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. Performed over a population of cells, this will result in DNA-protein complexes, where chromatin fragments from genomic regions that were on average spatially closer to each other will be over-represented compared to DNA segments that were on average more distant from each other. A DNA ligase is then added to the mix, resulting in the ligation of DNA segments, with a strong bias towards segments that have been trapped together within the same DNA-protein complex. As a result, on average, DNA segments that were closer to each other will be preferentially ligated together compared to segments distant from each other in the population of cells. The crosslink is then reversed. In the original 3C protocol, the relative frequency of religation events between segments of interest (i.e. presumed to reflect their relative contact frequency, hence spatial proximity) was assessed using semi-quantitative PCR. Primers designed to hybridize on two regions of interest were used to perform a semi-qPCR. Quantified onto a gel, the amount of amplified product, once normalized, reflected the relative contact frequency of two known genomic loci. Some limitations of the technique are the requirement to select regions to monitor and the design of qPCR oligonucleotides, which could also lead to a number of biases and caveats. Nevertheless, 3C was successfully used in the seminal study by Dekker et al. [72] to characterize the overall conformation of budding yeast chromosome III from a 12 x 12 contact map (reflecting the number of qPCR primers designed). This contact map was further converted into a distance matrix, itself useful to generate a 3D representation of the chromosome. These remarkable data remained fully consistent over the next two decades with results obtained using improved protocols offering increasingly high resolutions.

From then on, many derivatives of the 3C technique were developed. The most significant improvement was enabled by the possibility to perform paired-end high-throughput sequencing, which led to the development of a genome-wide application of 3C, called Hi-C (Fig. I.D, right). This method shares the main steps of 3C, the

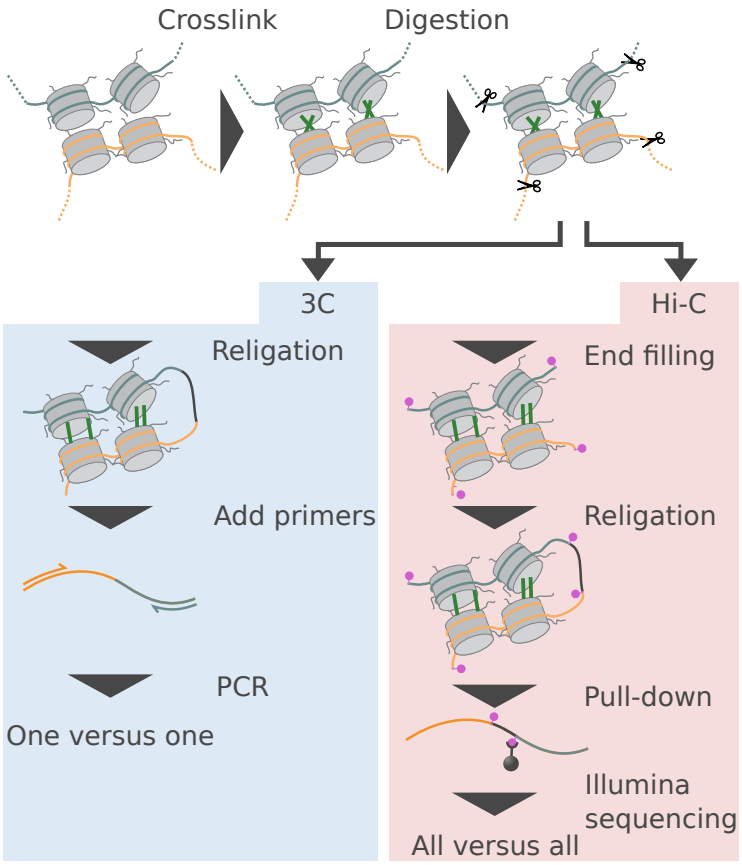


Fig. I.D: Chromosome conformation capture protocol: Chromosome conformation capture protocols share common steps (top): The chromatin is first crosslinked to form covalent DNA-protein bonds and then digested using a restriction enzyme. The Hi-C protocol subsequently differs from the original 3C protocol. In 3C (left), fragments are religated, the crosslinked is reversed and specific primers are added to amplify a pair of known loci. This allows the quantification of interactions 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.

main difference being that pair-end sequencing is used instead of qPCR. In Hi-C, fragment ends are filled with biotin prior to religation [73] and religated products are enriched through pull-down using streptavidin beads (which have high affinity for biotin). After ligation, sequencing primers are directly plugged to the extremities of the 3C enriched library. Paired-end sequencing is performed, and each read of the pair is then aligned onto the reference genome to determine its original position. This procedure allows the quantification contact frequencies of all versus all loci in the genome instead of using specific primers for a single pair of loci.

The information generated by Hi-C experiments therefore consists in counting how many times pairs of restriction fragments were found ligated together. This results in a list of contacts between all (in theory) pairs of restriction fragments in a genome. These contacts are most commonly visualized and interpreted using matrices, also called contact maps (Fig. I.Ea), which are two-entry tables represented as color-

coded heat maps. The color of each value in the matrix corresponds to its value relative to the others, reflecting the contact frequency between the associated pair of fragments. Those contact maps are an indirect representation of the presumed tri-dimensional folding of chromosomes. When processed and associated with other "omics" data, or performed in various mutant contexts, they are rich in information regarding chromosome regulation.

The various folding structures formed by chromatin can result from the direct or indirect action of DNA binding proteins. In mammals (and most metazoans), a typical example is the CCCTC-binding factor CCCTC-binding binding factor (CTCF), a transcription factor that also appears to act as an "architectural protein" structuring chromatin. Molecular motors such as the members of the Structural Maintenance of Chromosomes (SMC) complexes (cohesin, condensin...) and other proteins families 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 its ring-shaped structure, a process known as loop extrusion [74]. 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.

Spatial structural features of chromatin are reflected on Hi-C contact map by specific patterns (Fig. I.Eb). At the largest scale, chromosomes are relatively independent from each other in the nucleus, reflecting both their polymer nature, as intra (cis) contacts are favored over extra (trans) contacts, as well as their propensity to occupy distinct non-random "chromosome territories". Such large-scale disposition was unveiled in a broad variety of species using fluorescent in situ hybridization (FISH) of individually labelled chromosomes [75–78]. In genome-wide Hi-C contact maps, chromosomes therefore appear as squares of darker intensities along the diagonal, reflecting the fact that, on average, each chromosome makes more contacts within itself than with any other chromosome (Fig. I.Ea).

In several plants, mammals, as well as Drosophila [73, 79, 80], chromosomes are segmented 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 [73, 81]. 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 [82]. This spatial segregation results in enriched contacts between regions segregating within the same type of compartment, which are reflected on Hi-C contact maps by a plaid-like pattern.

Within each chromosome, chromatin forms "self-interacting domains" where DNA appears more prone to interact with other DNA segments in the same region than outside, Topologically Associating Domain (TAD) in mammals [83–85]. Genes and regulatory elements sharing the same domain are in close proximity, while being isolated from genes in neighbouring domains. Although genes within the same domain have more similar expression [84], the global impact of domains on the maintenance of expression is modest [86]. On contact maps, 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). These contact patterns are an ensemble average structure from millions of cells observed through bulk Hi-C. Although, well conserved across experiments and functionally important, domains are variable across single cells [87]. One emerging thought is that TADs should be more precisely defined based on their mechanism of formation rather than their appearance on contact maps [88–90]. 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 [74, 91].

2.3.2 Processing and analysis of Hi-C data

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 presumably linked to chromosome compaction and can vary depending on the environmental, cellular or genetic conditions (mutants, cell cycle stage, etc.)

Due to the high intensity of this $P(s)$ gradient, patterns of biological relevance, that deviate from polymer expectations/predictions, are often obscured on the contact map (especially at short scales). A common preprocessing step to account for this variation in the contact frequencies, and ponder 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) [92, 93]. Lower intensity patterns such as domains, compartments and chromatin loops become then 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 by decomposing the normalized contact map using Principal Component Analysis (PCA) [94], a dimension reduction technique

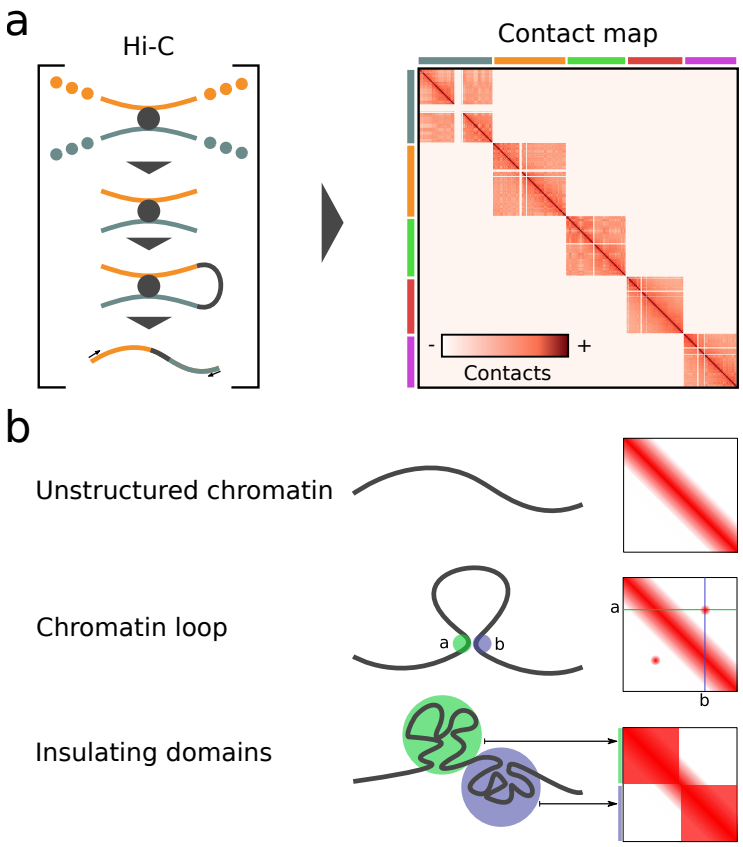


Fig. I.E: Interpretation of Hi-C contact maps: **a**: The Hi-C protocol (left) generates millions of read pairs representing contacts between genomic loci in a population of cells. Those contacts can be stored into an all-versus-all contact matrix (right) averaging all contacts in the population. Each 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.

which produces unit vectors retaining as much variability in the data as possible. The vector (i.e. principal component) explaining the most variance (Fig. I.Fb) can then be retrieved to get the compartment signal. In some cases where the compartment signal is weak (e.g. noisy datasets), it may not be contained in the first eigenvector. A robust approach is to select the principal component with the strongest absolute correlation to an external signal known to be associated with active chromatin, such as gene expression or GC content [95] (Fig. I.Fc). The sign of the principal component 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).

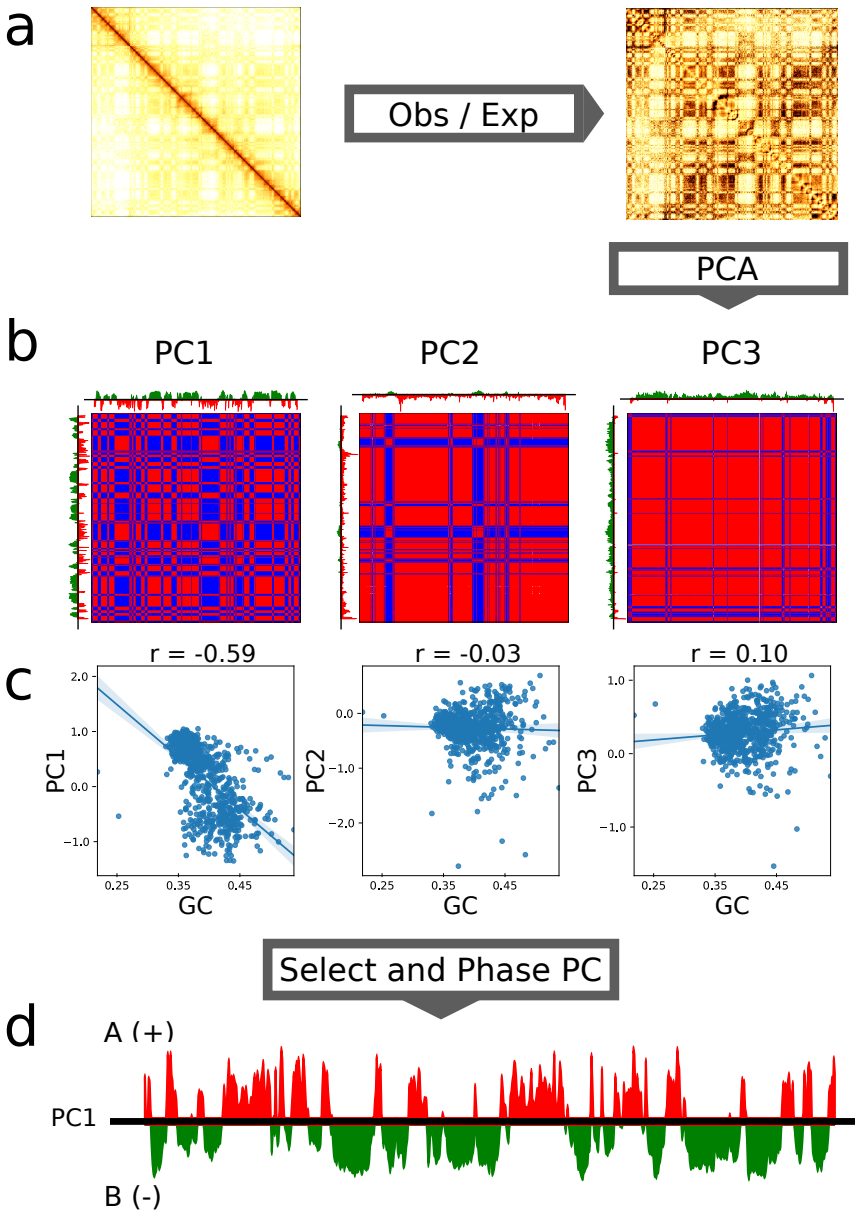


Fig. I.F: Representation and analysis of chromatin compartments in Hi-C. **a:** observed over expected (o/e) normalization is applied to the balanced contact map to remove the distance-decay gradient. Higher frequency of interactions within the same compartment result in a plaid-like pattern on chromosome contact maps. **b:** PCA is applied to the o/e normalized contact map and the first few principal components (PC) are retained. For visualization, each PC is shown alongside its outer product, 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.

Many eukaryotic genomes are segmented into insulating domains containing frequently interacting loci, named TAD in mammals. TADs form a multi-scale hierarchical organization and often contain genes and their associated regulatory elements

[96]. Regions in separate TADs present various degree of insulation from each other, and the strength of this relative insulation can be quantified using an "insulation score". The insulation score of a given region can be simply defined as the intensity of contacts across that region (upstream with downstream) within a pre-defined distance [97] and can be represented as a numerical track along the genome. Improved metrics such as the relative insulation score have since been developed to improve the detection of TADs [98]. The relative insulation score (RI) at a locus *s* between bins *k* and *k+1* with a predetermined window size *w* is defined as:

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

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

$$B(w,s) = \sum_{i=-w+1}^{0} \sum_{j=1}^{w+i} M_{k+i,k+j} \tag{2.3}$$

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

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, chromatin loops contain valuable information regarding interactions between regulatory elements such as enhancers and promoters, and their characterization has unveiled unknown layers of regulation complexity. Accurately calling the positions of the loops is therefore of interest, and many tools have been developed for this purpose. These computational approaches typically search for local enrichment of contacts, appearing as dots away from the diagonal, in mammalian, or large metazoan genomes contact maps [99, 100]. 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 and strengths of chromatin loops between regions, or be compared between mutants or experimental conditions [101].

Identifying and quantifying contact changes proved important in the study of many biological processes, such as differentiation or cell cycle progression [102–104].

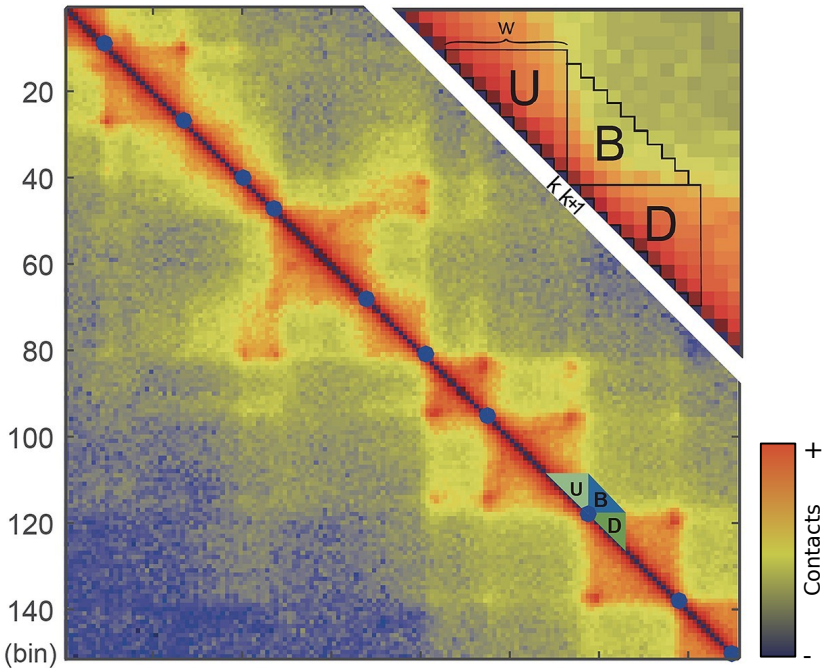


Fig. I.G: Visual illustration of the relative insulation score. Computing the relative insulation score at bin k involves computing the average interactions between upstream (U) and downstream (D) regions, denoted as B, as well as the average contacts within U and B. The key parameter when computing insulation is the window size (w), determining the size of the U, B and D. Figure adapted from [98]

In that regard, a direct, global and quite approximate comparison is to compute a similarity metric between pairs of samples. This metric recapitulates very general patterns, and is very imprecise at identifying discrete changes, but can be convenient as a quality control, notably to estimate technical (replicates) and biological (conditions) variability. Different comparison metrics have been used, such as the sum of differences between Hi-C matrices [105], correlation coefficients [106] or distance between the matrix eigenvectors [107].

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 [108–110]. 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 consistent across replicates. This approach relies on solid statistical grounds, but it often does not address the question at hand. When analyzing Hi-C data, one is sometimes more interested in finding specific structures appearing or disappearing rather than identifying discrete simple contact change at a region. The development of methods to discover relevant changes in chromatin conformation patterns is still an active area of research.

2.4 Combining layers of biological information

The central dogma of biology - "DNA → RNA → 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 [111], noncoding RNAs recruiting proteins to repress transcription [112] or directly repressing translation by preventing ribosome binding [113, 114]. More generally, gene expression is affected by transcription factors binding to surrounding regulatory DNA sequences. The arrangement of those sequences along the genome form what has been coined the *cis*-regulatory code [115] and regulates gene expression in coordination with transcription factor concentrations. Spatial interactions have been proposed as an additional player in the regulatory code, through the delineation of "gene domains", consisting of insulated domains (i.e. small TADs) which connect genes with the appropriate regulatory sequences [116, 117].

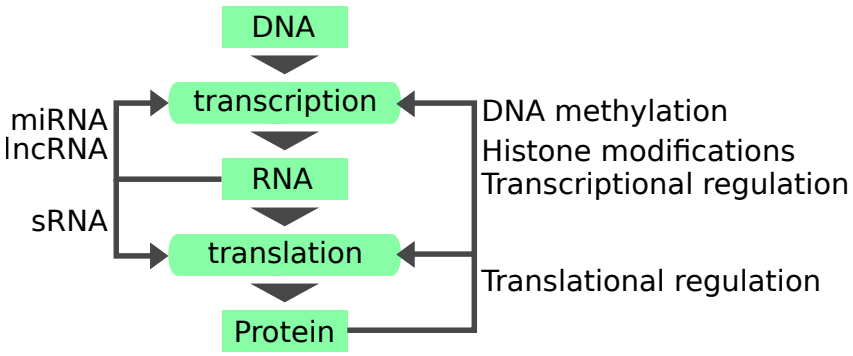


Fig. I.H: Central dogma of molecular biology. Products and reactions from the central dogma are shown in green, with grey arrows showing some of the regulatory interactions between the different biomolecules.

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, gene expression, protein binding, histone modifications or protein abundance and identify relationships with phenotypes of interest.

One of the main challenges is to find efficient ways to combine this information to extract meaningful biological information. More often than not, they are analysed separately to find regions of deregulation common to the different layers. Another challenge is the difficulty to combine different datasets due to technical heterogeneity or biological variations, such as different strains, heterozygosity, experimental

conditions. Nevertheless, there have already been attempts at integrating these levels of information [118, 119].

3 The importance of genome assembly

Most of the genomic techniques presented earlier cannot be performed without having a high quality sequence of the genome of the species investigated at hand. Ideally, a complete reference would consists to a telomere-to-telomere genome, as downstream analyses will rely on the relative positions of different biological elements on the genome sequence to draw biological conclusions.

The availability of several sequenced genomes closely related to the species of interest is also crucial for comparative analyses. Among other things, it enables identification of HGT events, comparative genomics analysis, and investigation of the evolution of molecular processes. However, for many species, full reference genomes remain incomplete or nonexistent. Several worldwide efforts have been undertaken to generate catalogs of the genomes of all existing species, broadening the number of species with their full genome being sequenced. Until recently, the only eukaryotic genomes with telomere-to-telomere sequences, and (almost) no gaps within the chromosomes, consisted in fungi with compact genomes species such as *Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*, the worm *Caenorhabditis elegans* [120, 121]. Here we describe in more detail the process of genome assembly and its relevance to infection genomics.

3.1 From reads to chromosomes

Genome assembly consists in reconstructing the linear sequence of the genome from the readings of DNA by different 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 and Gilbert methods were used to read DNA sequences [122, 123].These are low throughput, but highly accurate sequencing methods. These technologies allowed to unveil the complete genome sequences of viruses [124–126] followed by chromosome III of *S. cerevisiae*, the first eukaryotic chromosome to be sequenced [127]. A common practice at the time, was to clone small genomic regions of ~10-30 kb into a plasmid resulting in a bacterial artificial chromosome, amplify it in bacteria and extract it [128]. The positions of those clones and relative order on the chromosome were then determined by

digesting them into restriction fragments and hybridizing them to identify overlaps and construct a physical map of the chromosome [129]. Each clone was then randomly fragmented and sequenced. 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 enabled the automation of the sequencing process to tackle the assembly 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. Overlapping 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 [127, 130, 131], but technological advancements progressively reduced the cost and time required. A decisive change was the development of shotgun sequencing [132], 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 *Reads*. 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 the 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 [133]. This lead to the development of de Bruijn-based assembly algorithms, which many modern assemblers still use [134, 135]. de Bruijn assemblers split reads into short *K-mers* 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 [136]. Whenever a repeated sequence is longer than the read itself, the graph can not be solved and heuristics have to be

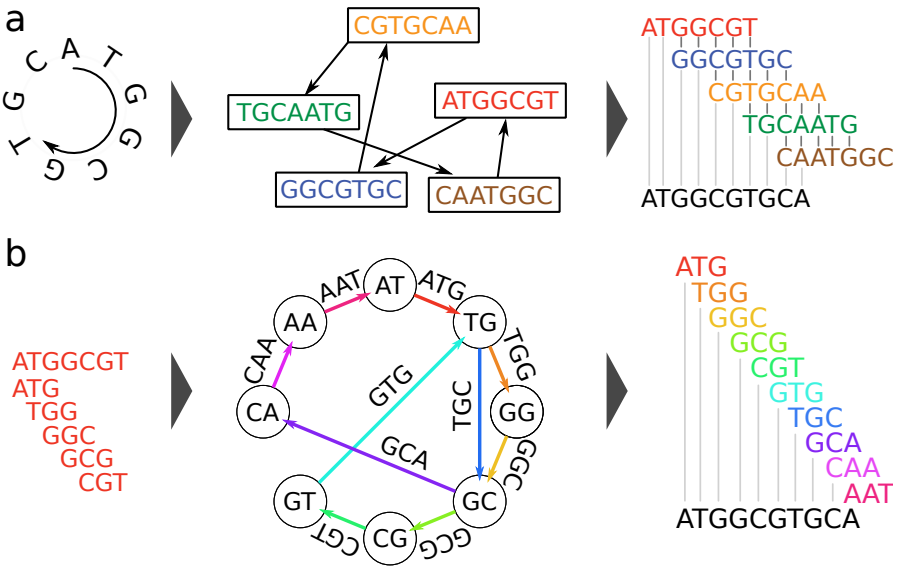


Fig. I.1: Graphs in genome assembly: A small circular genome is sequenced and the resulting reads are shown in color **a**: Early assembling techniques computed all pairwise alignments between reads to represent them as nodes in an overlap graph and their overlaps as edges. The genome sequence can be 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 [133]

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 can span most repeated regions. Recently, these technologies were used to generate telomere-to-telomere assemblies of several human chromosomes [137, 138]. Third generation sequencing techniques still suffer from their lower base calling accuracy resulting in assemblies with high point error rates (>10%) and indels [139, 140]. To remove these errors, some methods have been developed to correct long reads before assembly, either by correcting long reads among themselves [141], or using a separate set of short accurate reads to erase sequencing errors in long reads [142]. 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 [143]. One major drawback of read correction methods is their high computational cost, as they require to align high number of reads to each other. An alternative strategy is to use the uncorrected reads to assemble the genome and perform error 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 [144]. Additionally, they can correct larger scale misassemblies such as indels by using the pair-end information and alignment discrepancies [145]. Some polishers have obtained better polishing accuracy by combining the information in short and long reads [146].

The last sequencing technology in date is the HiFi platform from Pacific Biosciences which produces fairly long (10-25kb) but very accurate (<1% error rates) reads, thus offering a comfortable middle ground between Nanopore and Illumina reads [147]. This trade-off has made HiFi reads popular for genome assembly [148, 149]. Their low error rates have also enabled algorithmic developments with drastically lower computational costs for the assembly of large metagenomes [150].

Recently, the emergence of specialized technologies aimed at scaffolding have allowed the generation of even more contiguous and correct genomes at reduced costs. One example is the recent rebirth of optical mapping to introduce fluorescent probes into chromosomes at specific sites [151]. 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).

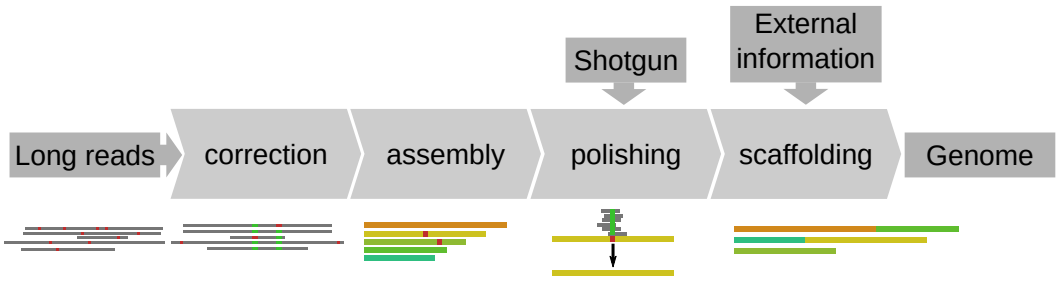


Fig. I.J: Example of a typical assembly pipeline using third generation sequencing. The error prone long reads are first corrected by pairwise comparisons. The corrected reads are assembled into contigs using their overlaps. The remaining sequencing errors in the assembly are removed by polishing with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

3.2 Phylogenetic representation

A common way to analyse 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 [152]. For this reason, detection of recent events requires genomes of closely related organisms as a comparison point.

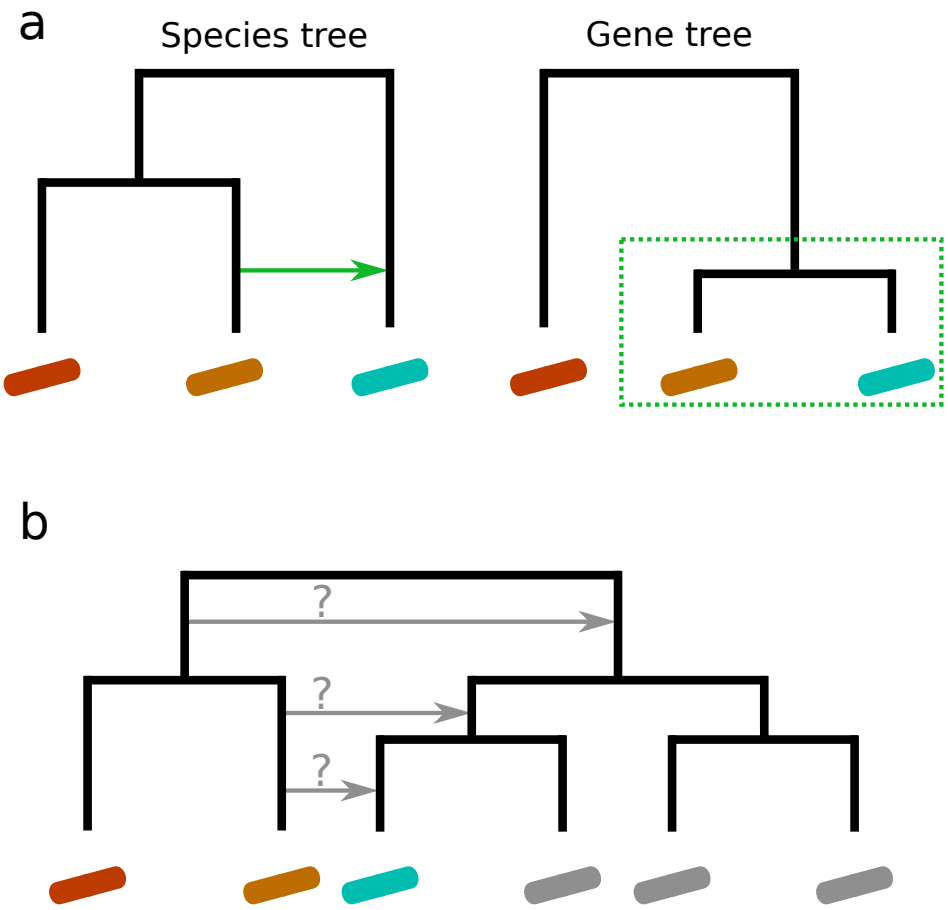


Fig. I.K: Phylogenetic representation of an HGT event. **a:** An HGT event between two species (shown with a green arrow) can be detected through discrepencies between the species (left) and gene (right) trees. **b:** In cases where genomes of closely related species are unavailable (greyed out organisms), 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 set of genes they contain, known as pangenome. This also allows the identification of genes specific to a subset of these genomes, 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 [153–155] 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 [156]. 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 genomes incurs mapping bias due to ambiguous alignments of redundant regions between references [157]. Similarly, genome graphs also allow systematic alignment to different alleles in polyploid organisms, solving the issue of allele-specific mapping bias in linear references [158].

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. Genomics is a fast changing field and 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*. In the last chapter we will focus on murine bone macrophages infection by *S. enterica* and the genomic alterations it entails. 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

> *Unfortunately, no one can be told what The Matrix is. You'll have to see it for yourself.*
>
> — **Morpheus**
> The Matrix

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 data. The information contained in these data is not always readily accessible, and in addition most of it is often not directly relevant for the problem at hand. One of the main challenges emanating from genomics data is therefore process the data to extract meaningful information, and then distill this information and extract only the relevant signal.

In the case of Hi-C and other related derivative genomics techniques, the resulting signal is a collection of contacts recorded between pairs of genomic segments. These contacts reflect the average genome structure from a population of cells. However, in addition to reflecting the average of a population of cells, the data themselves are subject to various biases intrinsic to their generation.

The specific contacts revealing spatial features and changes of interest are therefore sometimes hard to detect and, importantly, to quantify and validate statistically. They can be faint, reflecting events that occur only in a fraction of the cells in the population, or masked by experimental noise. This is especially true regarding data generated from species that are not heavily investigated and for which optimization of the protocols was not performed. The quality of the data has indeed strongly improved over the last 10 years, allowing the reduction of bin sizes from 1 Mb for a human contact map in 2009 [73], to 1 kb in more recent papers. In yeast, bacteria, or Archaea, several years passed before protocols reaching a decent resolution (e.g. ~20 kb bin) were developed. Only recently contact maps of bacteria reached bin sizes of 1 kb [159]. Detecting and quantifying the changes in these contact maps of variable quality therefore requires a set of bias correction and signal detection methods which are still under continuous development, drawing from innovation in computer science and algorithmic fields.

In this section we review the recent methodological developments that allow correcting the Hi-C signal and present new methods to extract, quantify and assess the relevance of biological features from these datasets. These developments proved necessary to tackle the questions raised in the following chapters.

1.1 Streamlined and reproducible Hi-C processing

Pre-processing of Hi-C data, which consists in converting 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 [94]. Chimeric reads become more problematic when increasing the read size relative to restriction fragment length.

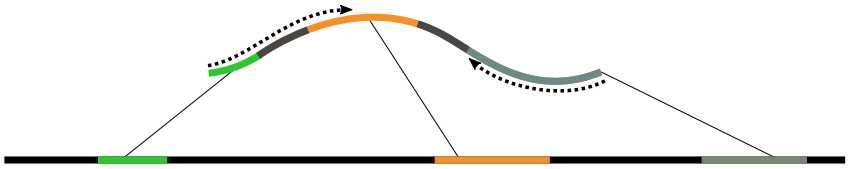


Fig. II.A: Chimeric reads in Hi-C: Example of a Hi-C fragment resulting in a chimeric read. The Hi-C fragment contains 3 different regions (green, orange and grey) which have been religated together. The paired-end sequencing reads are shown as dotted line. The sequencing read spanning the green and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

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) [160]. The various interaction types can be separated based on the strand of origin of their individual reads. In theory, and in practice at long ranges, one would expect religations to be strand agnostic and to have an equal abundance of all four possible combinations ($++$, $--$, $+-$, $-+$). In reality, this is never the case at short range contacts, due to the enrichment of dangling ends (or uncut fragments, $+-$) and self-circles (or loops, $-+$) (Fig. II.Bb, c) [160].

These biases must be accounted for 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.D) where reads are truncated and iteratively

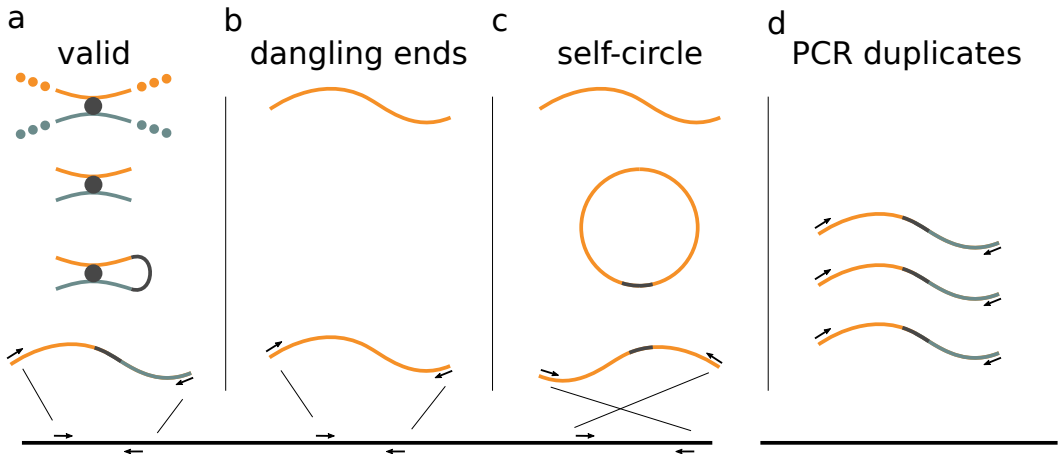


Fig. II.B: Types of interactions generated from Hi-C experiments: **a:** Valid interaction resulting from the religation of two distant loci in physical contact. **b:** Spurious event caused by the sequencing of a single restriction fragment, or undigested sequence. **c:** Spurious event resulting from 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.

extended until they align unambiguously. The Hi-C pairs are then assigned a numerical index according to the restriction fragment they originate from (Fig. II.E) and 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 (Fig. II.F). To allow compatibility with various programs, it can generate sparse matrices in 3 possible formats: COO, bedgraph2d and cool. COO (COOrdinate format) and bedgraph2D are text-based sparse matrix representations, while cool [161] is a specification based on the binary HDF5 (hierarchical data format) to represent Hi-C data. The cool format is seeing widespread adoption in the research community and offers several advantages, including low storage space, ease of use and fast random access compared to text formats.

Hicstuff is meant to be easily accessible [162], 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 program is written in python and is accessible 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 through 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 or organism.

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

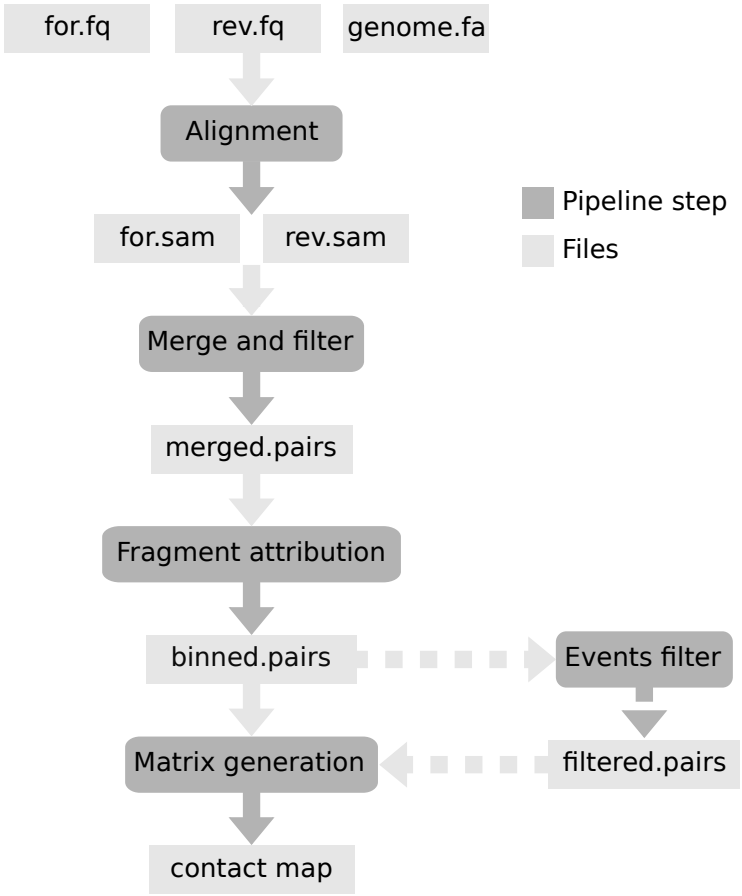


Fig. II.C: Overview of the hicstuff pipeline: Consecutive steps towards the generation of a contact map from sequencing reads, along with the intermediate files are shown as a directed acyclic graph.

The project has already fostered a modest community of users which are offering their contributions, suggest features or report issues they encounter. The Hicstuff pipeline is distributed through the python package index (PyPI) and its source code is available on github: `https://github.com/koszullab/hicstuff`.

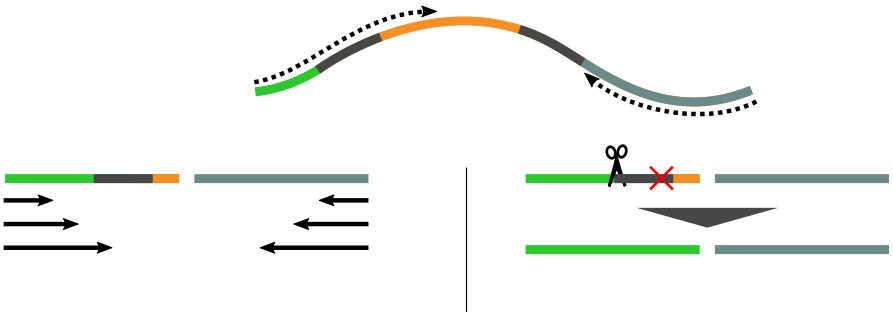


Fig. II.D: Iterative alignment of a Hi-C pair: The Hi-C fragment consists of 3 regions religated together (top). One sequencing read spans two regions (orange and green). Iterative alignment is used to uniquely align the resulting chimeric read (left). The read is truncated to a short length (e.g. 20bp) 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.

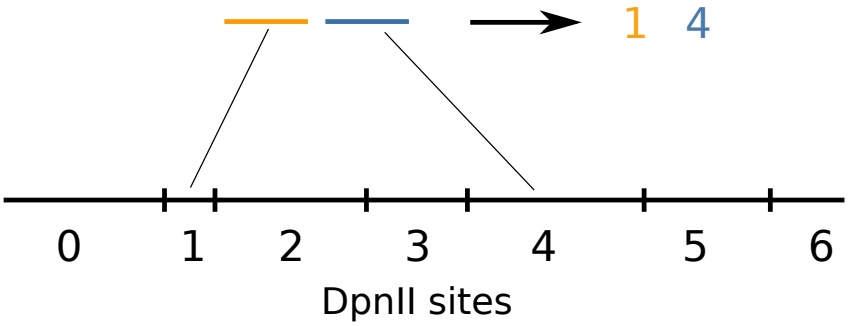


Fig. II.E: Fragment attribution of Hi-C contacts: The genome is segmented into discrete bins according to the positions of restriction sites. Hi-C reads are assigned an index according to the restriction fragment to which they aligned.

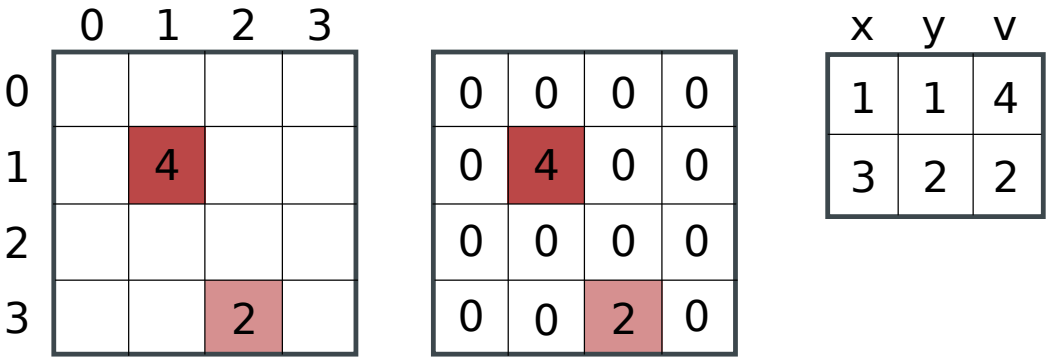


Fig. II.F: Dense and sparse matrix representations: . In Hi-C, matrices are very sparse (i.e. mostly contain 0s), (left). In dense matrix representation, we store all values explicitely. The information stored is highly redundant (middle). Such matrices can be stored efficiently using a sparse representation 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 [95, 163, 164]. Each of these methods use a set of specific rules to detect one particular type of pattern. For example, HICCUPS [164] 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 reliance on parameter values, 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. We developed a python package named *Chromosight*, which performs pattern detection on Hi-C matrices in cool format.

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 the user 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 (most 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*, an entire chromosome's contact can consist of to 30,000 x 30,000 bins of 10 kbp. One of the main drawbacks of sparse representation is that most algorithms are slower and harder to implement on such structures. 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.

During the development of Chromosight, we put special attention on good software practices 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 publishing publicly accessible tutorials and examples on a dedicated readthedocs website (`https://www.chromosight.readthedocs.io/`). Furthermore, the program is cov-

ered 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.

Chromosight's 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 Appendix A.1. Additionally, a case study demonstrating Chromosight capabilities is shown in appendix B.

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

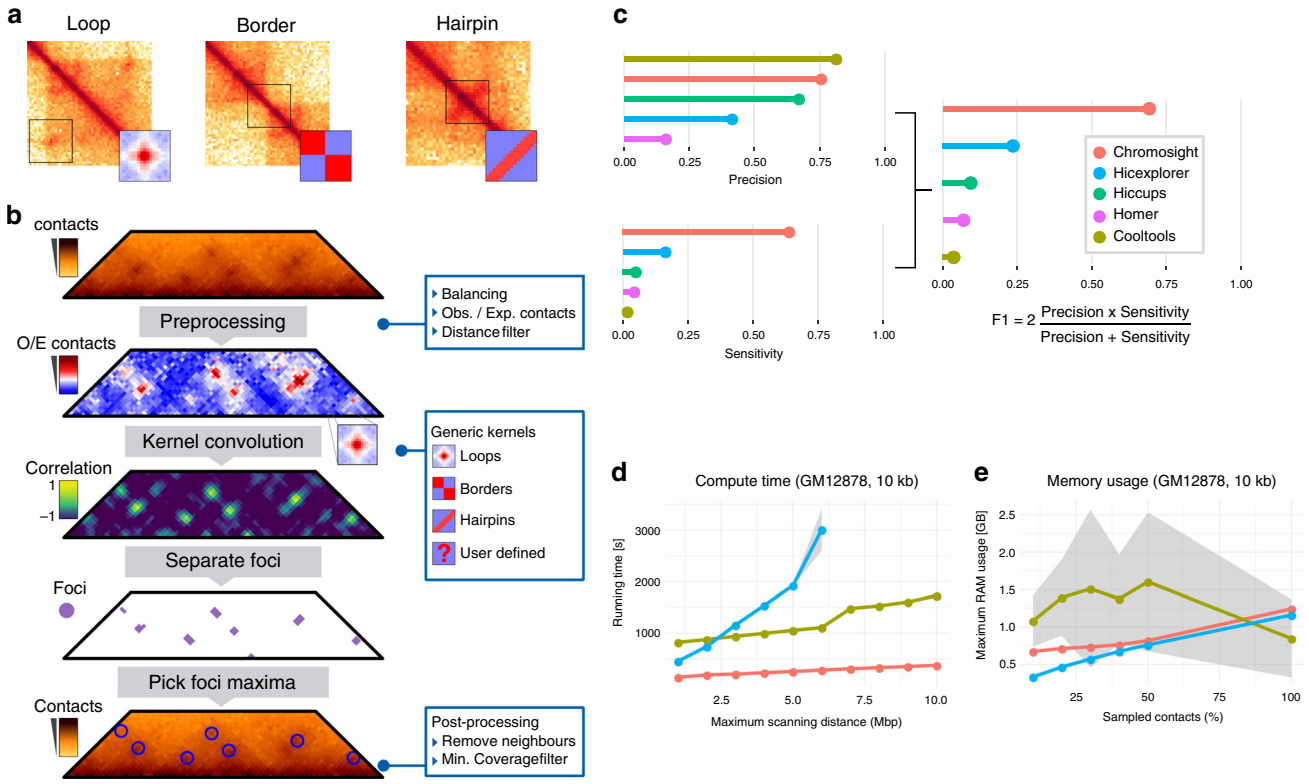


Fig. 1 Chromosight workflow and benchmark. a Examples of distinct patterns visible on contact maps (loop, border and hairpin) and the corresponding chromosight kernels. **b** Matrix preprocessing involves normalisation balancing followed by the computation of observed/expected contacts. Only contacts 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

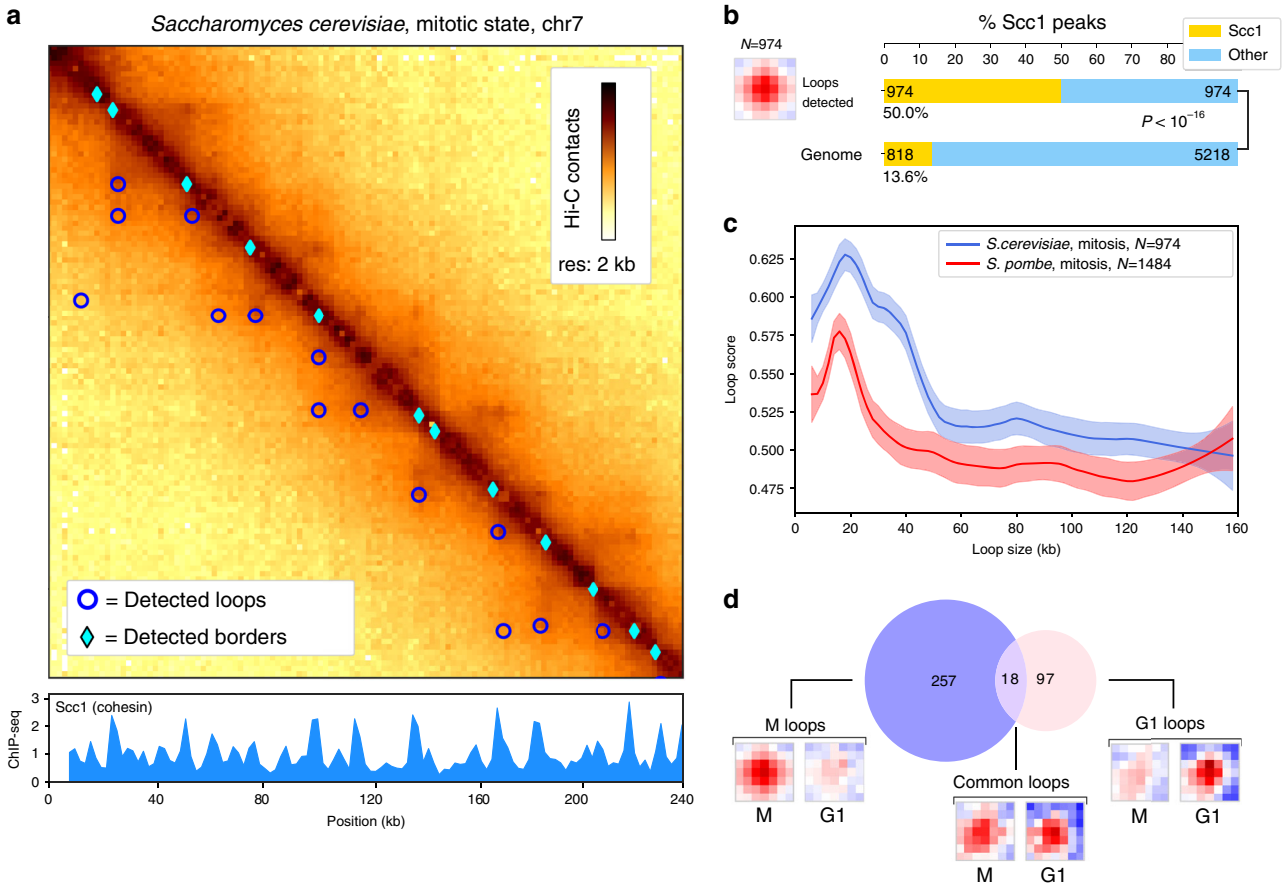


Fig. 2 Applications on yeast genomes. a Zoom-in of the contact map of chromosome 5 of *S. cerevisiae* with synchronised ChIP-Seq signal of Scc1 protein (cohesin) at 2 kb resolution with detected loops and border patterns,[8]. The darker, the more contacts. **b** Pileup plots of windows centered on detected 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.

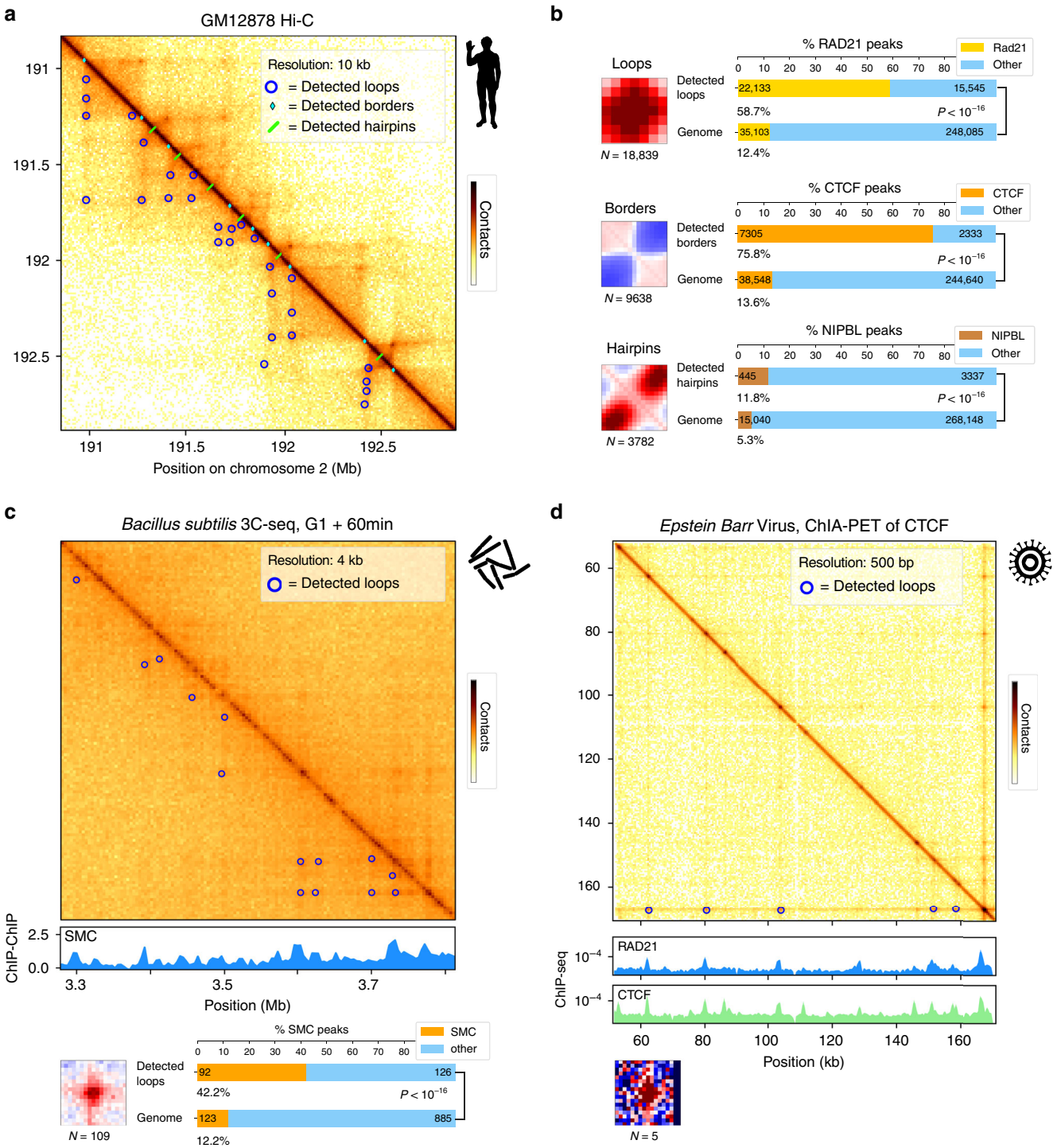


Fig. 3 Applications to various genomes. a Zoom-in of contact map for chromosome 2 of *Homo sapiens* at 10 kb resolution[36] with Chromosight detection of loop, border and hairpin patterns. The darker, the more contacts. **b** Left: pileup plots of windows centered on detected loops, borders and hairpins 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

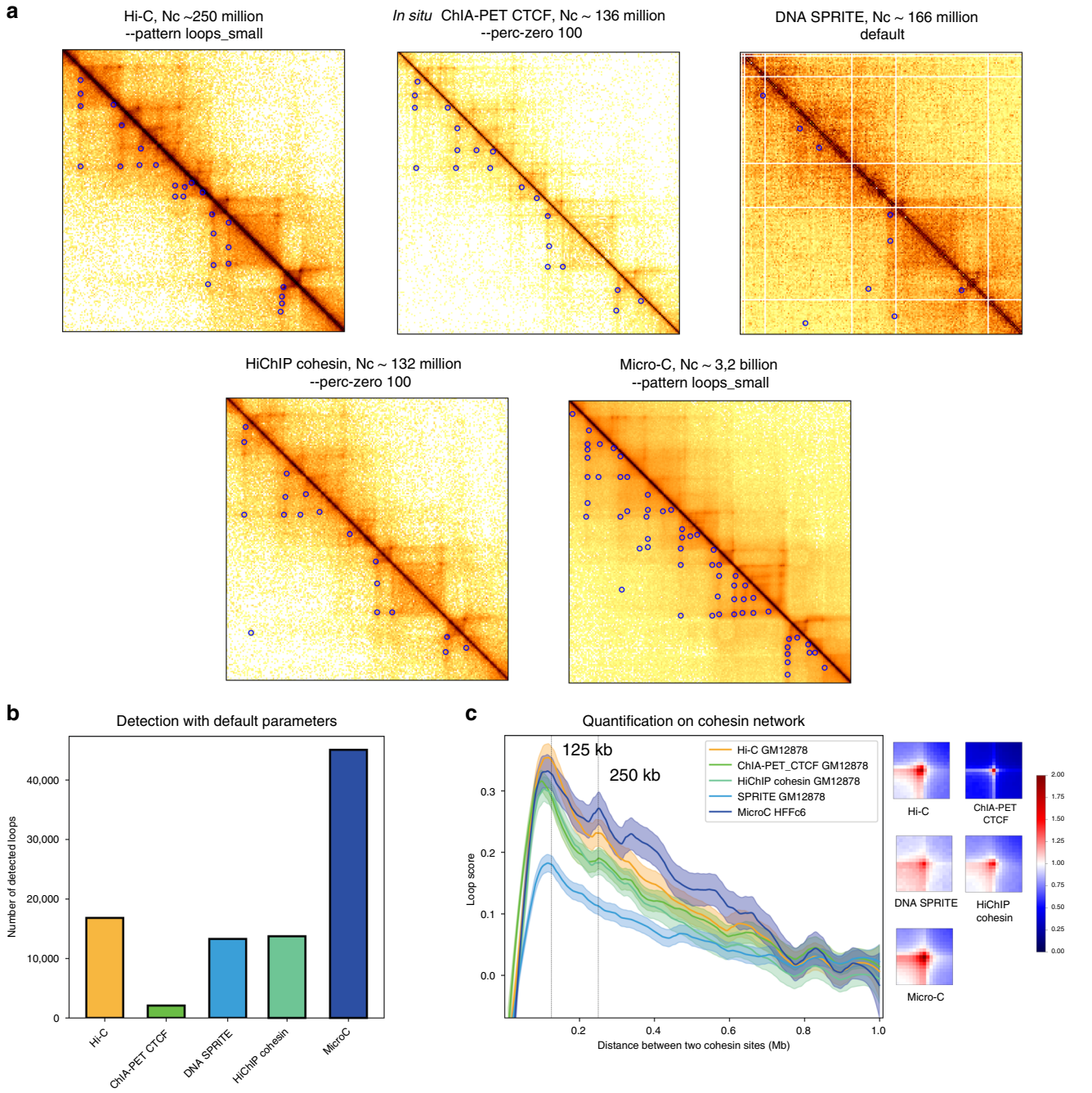


Fig. 4 Analyses with data from alternative contact technologies. a Magnification of *Homo sapiens* chromosome 2 contact maps generated with five different experimental methods (around STAT1 gene; bin:10 kb): Hi-C[36], In situ ChIA-PET of CTCF[42], DNA SPRITE[43], HiChIP of cohesin[44], Micro-C[45]. All 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

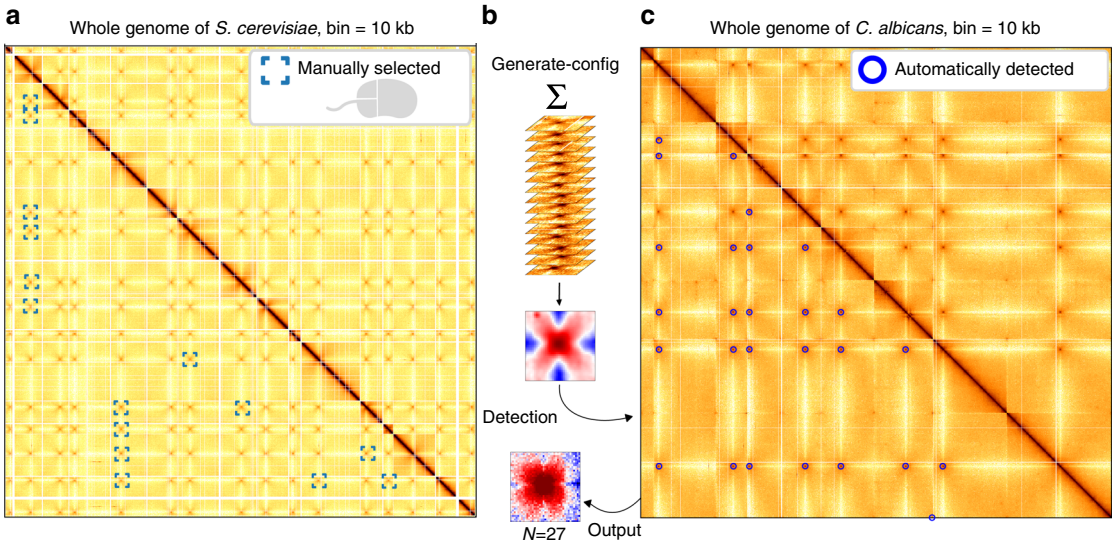


Fig. 5 Point and click mode. a Whole-genome contact map of *S. cerevisiae*[8] with 15 inter-centromere patterns that were selected by hand. Darker means more contacts. **b** Chromosight generates a new kernel by summing all the selected patterns and applying a Gaussian filter. **c** Chromosight detection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

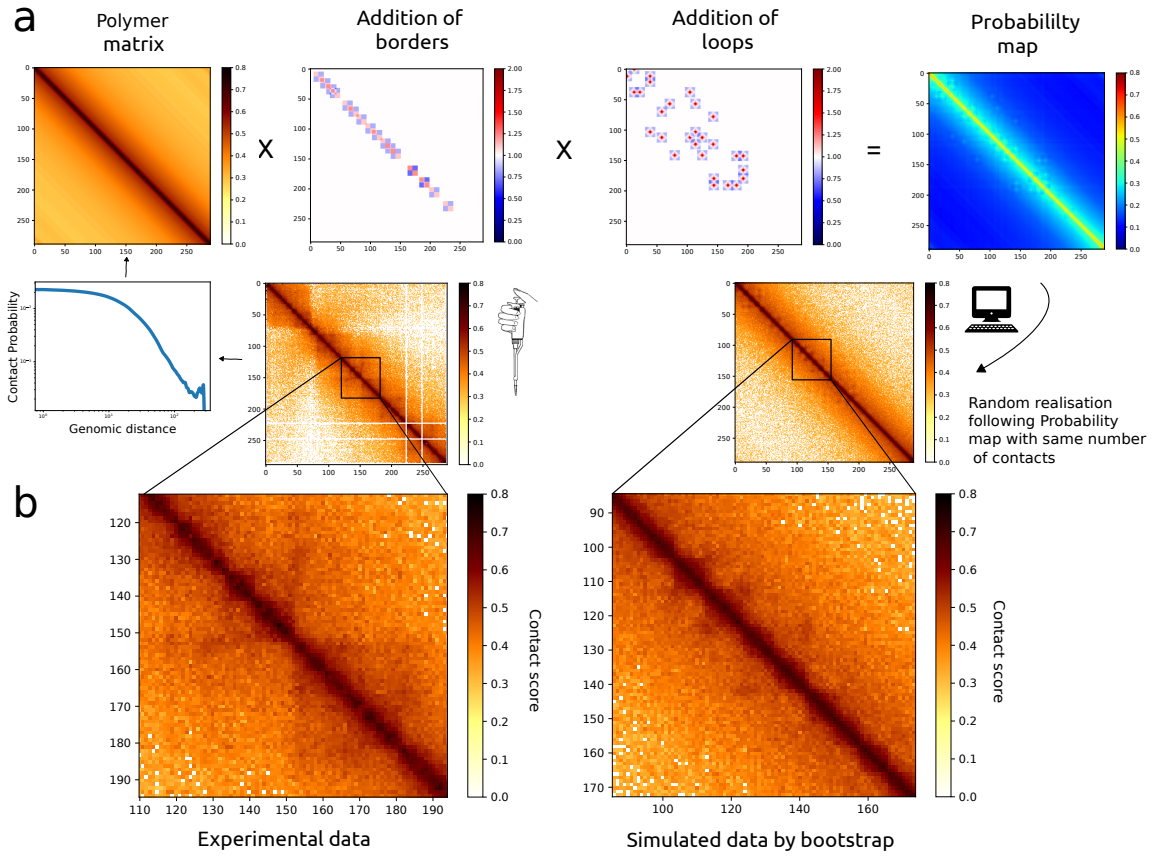
Given any template $\mathrm{IMG}_{\mathrm{TMP}}$, it is possible to approximate it by retaining only the K largest singular values in the SVD of $A = \mathrm{IMG}_{\mathrm{TMP}}$, such that:

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

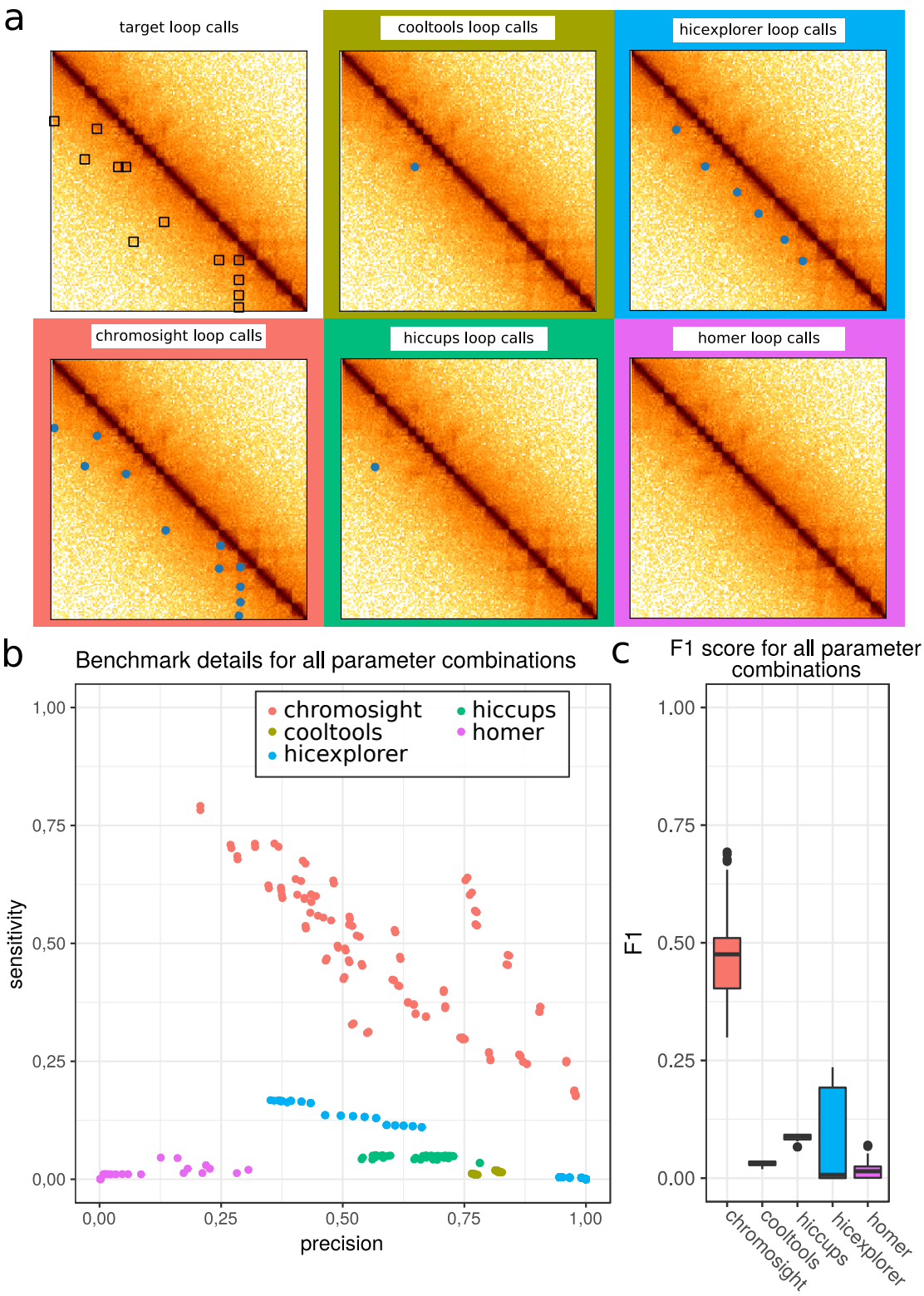
$$\mathrm{V}_{\mathrm{TMP}}[k,:] = \sqrt{D[k,k]}\,V[k,:] \tag{9}$$

$$\tag{10}$$

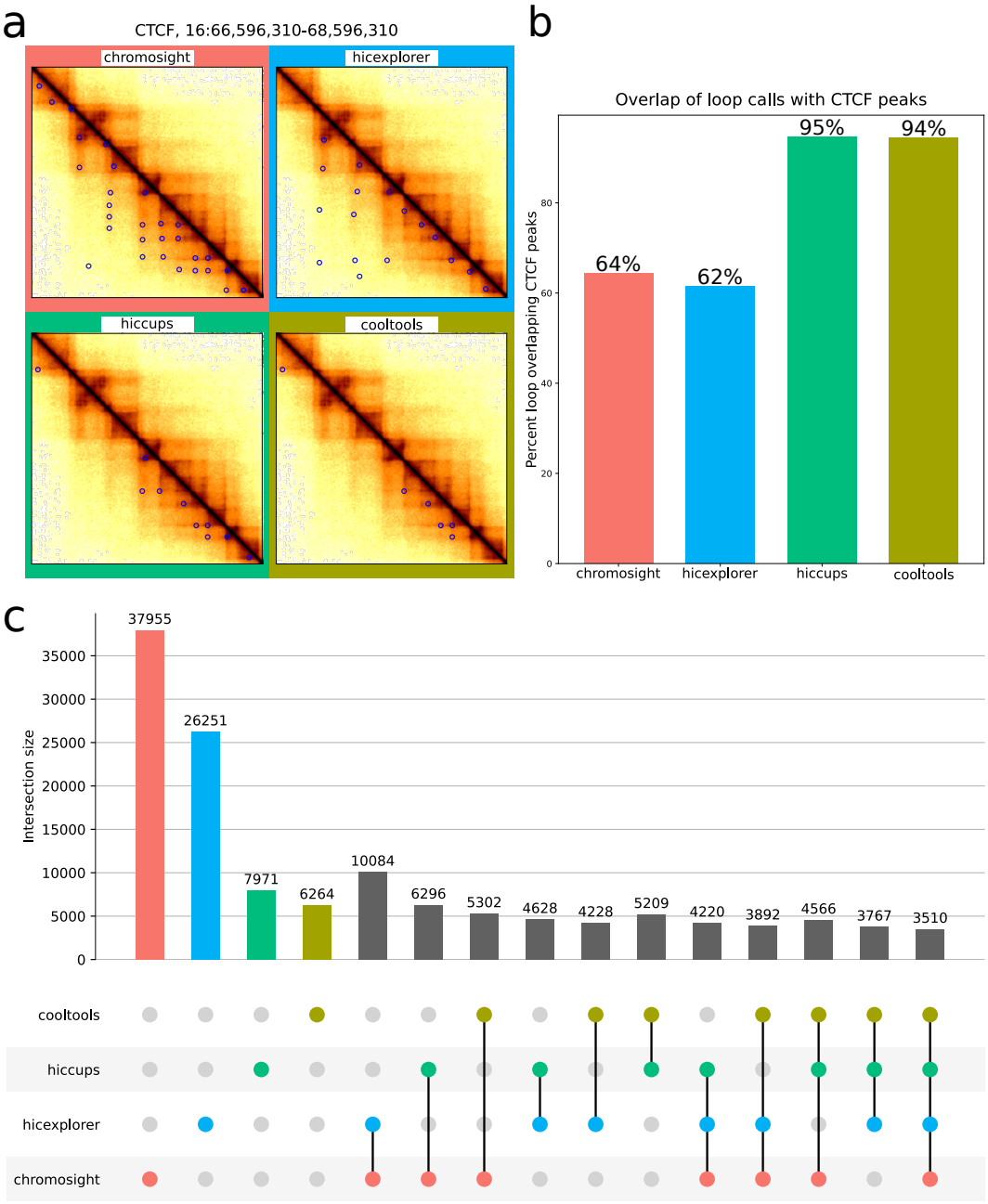
Let's give a toy example of the operations spared by using a SVD approach. Suppose that $M_{\mathrm{TMP}} = N_{\mathrm{TMP}} = 17$, the standard convolution would require $2 \times 17 \times 17 = 578$ operations per point. In contrast, if we use a SVD convolution with $K = 1$, the number of operations 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.



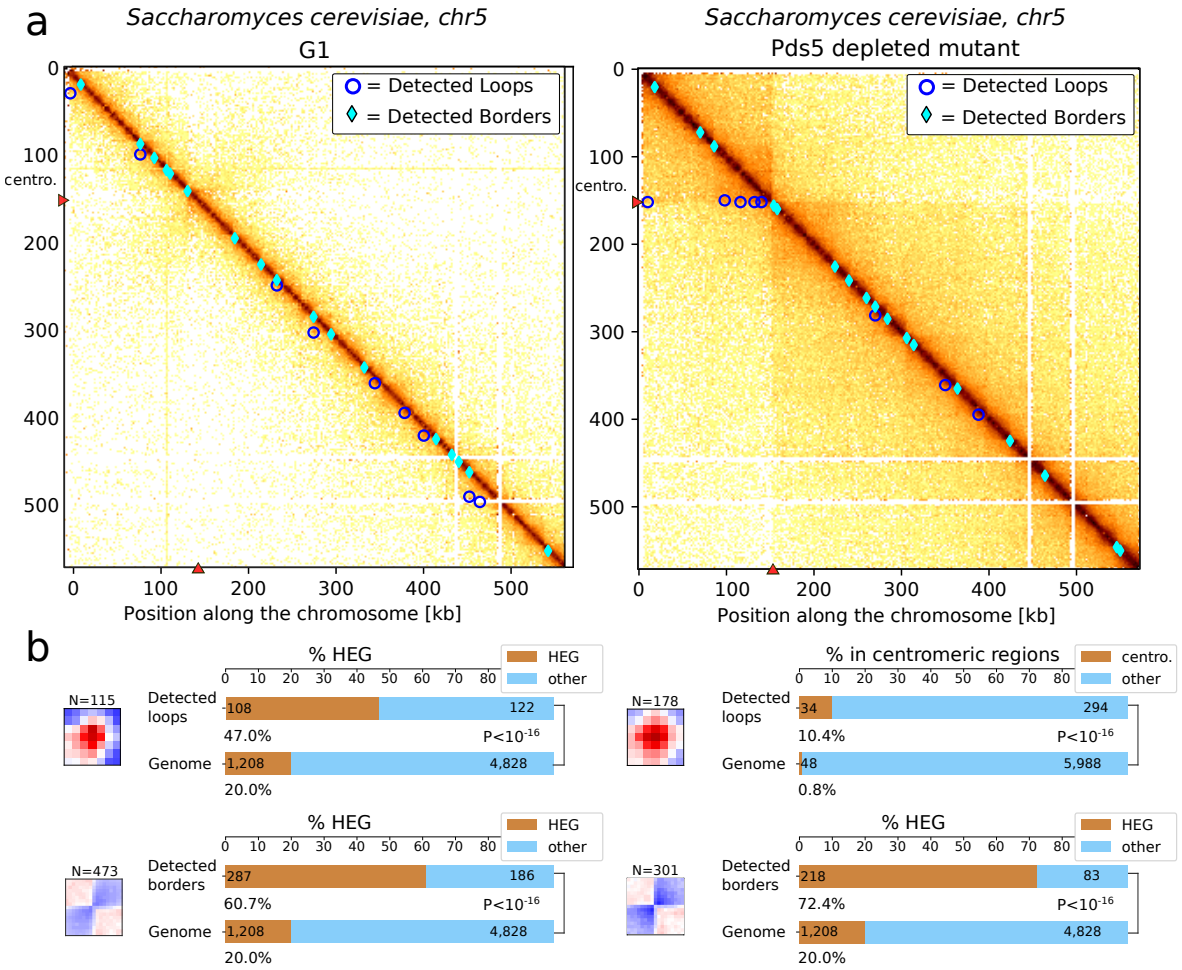
Supplementary Figure 1: Strategy for the generation of simulated contact data for benchmark tests of different loop calling algorithms a, The simulated data were generated with a bootstrap approach based on contact data generated for yeast *S. cerevisiae* in mitotic phase [2]. Three main features of 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.)



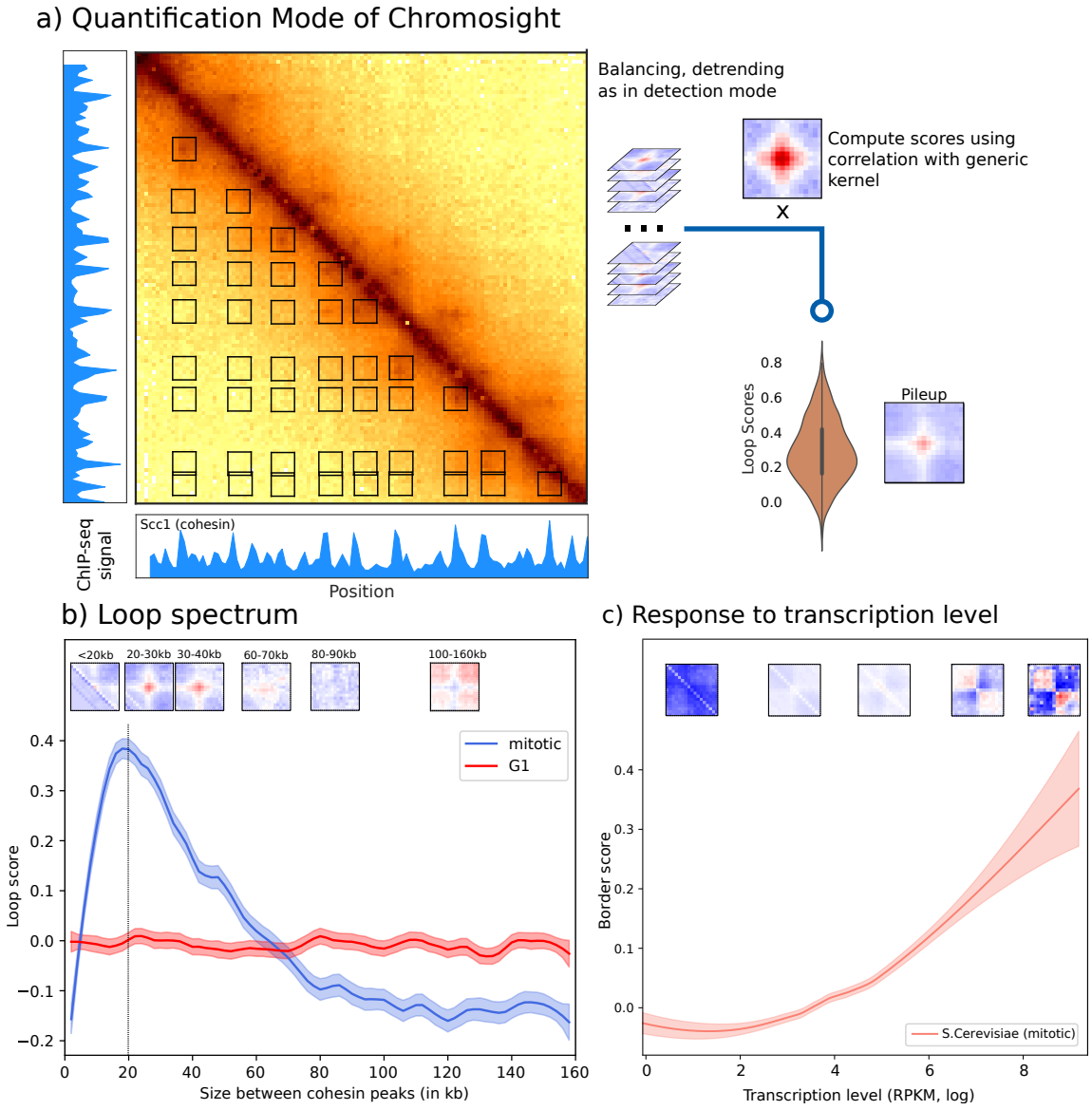
Supplementary Figure 2: Comparison of different loop callers on simulated data. **a**, Example region from a synthetic matrix with real loop calls (top left) and loops detected by all algorithms used in the benchmark using the combinations of parameters which yielded the highest F1 score. **b**, Precision 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).



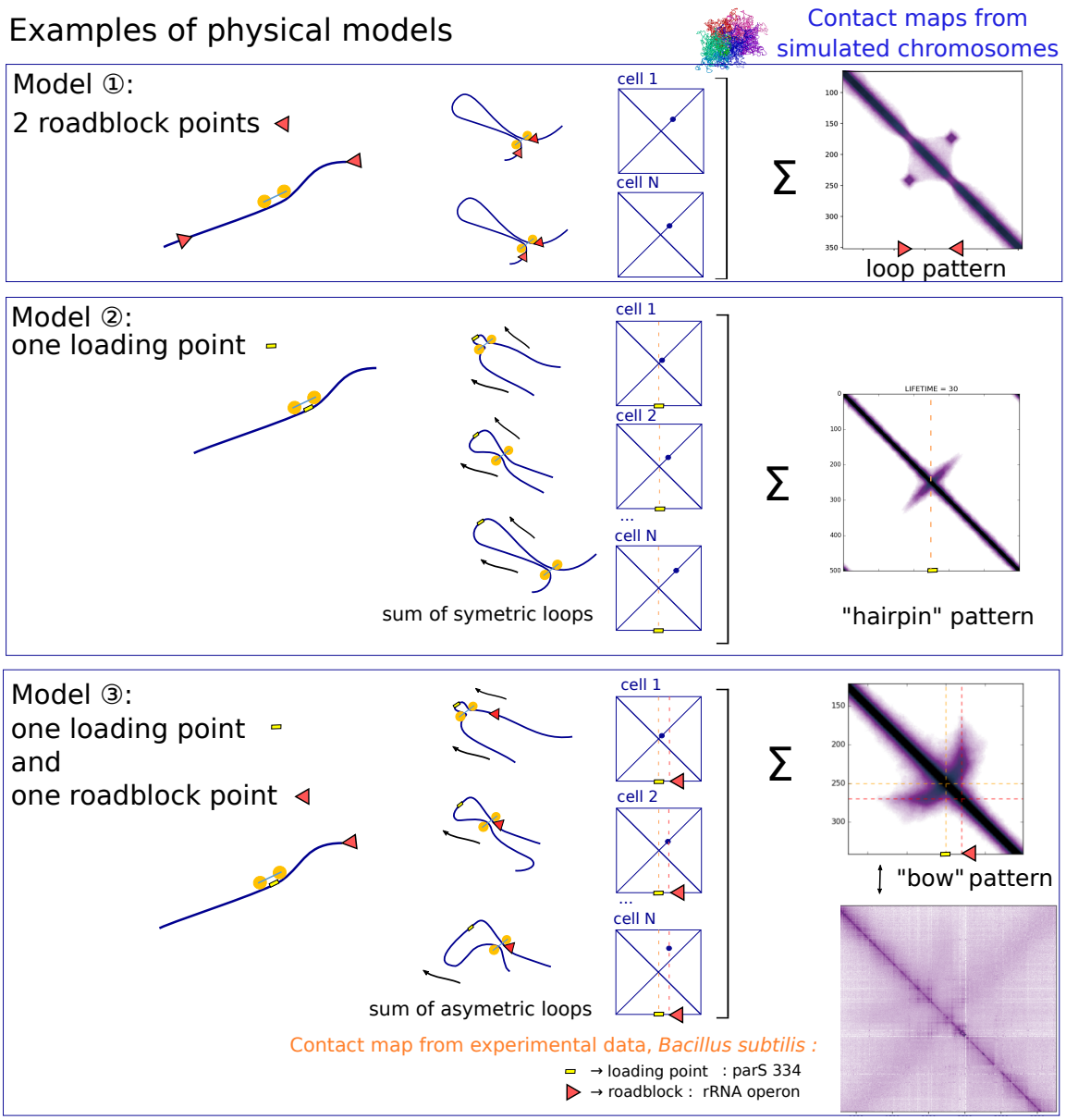
Supplementary Figure 3: Comparison of different loop callers on experimental data. **a**, Contact maps representating a region of +/- 1Mb around the CCCTC-binding factor (CTCF) gene of GM12878 (GSE63525, [4]), at 10kb resolution with coordinates of detected loops for different loop calling softwares, 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.



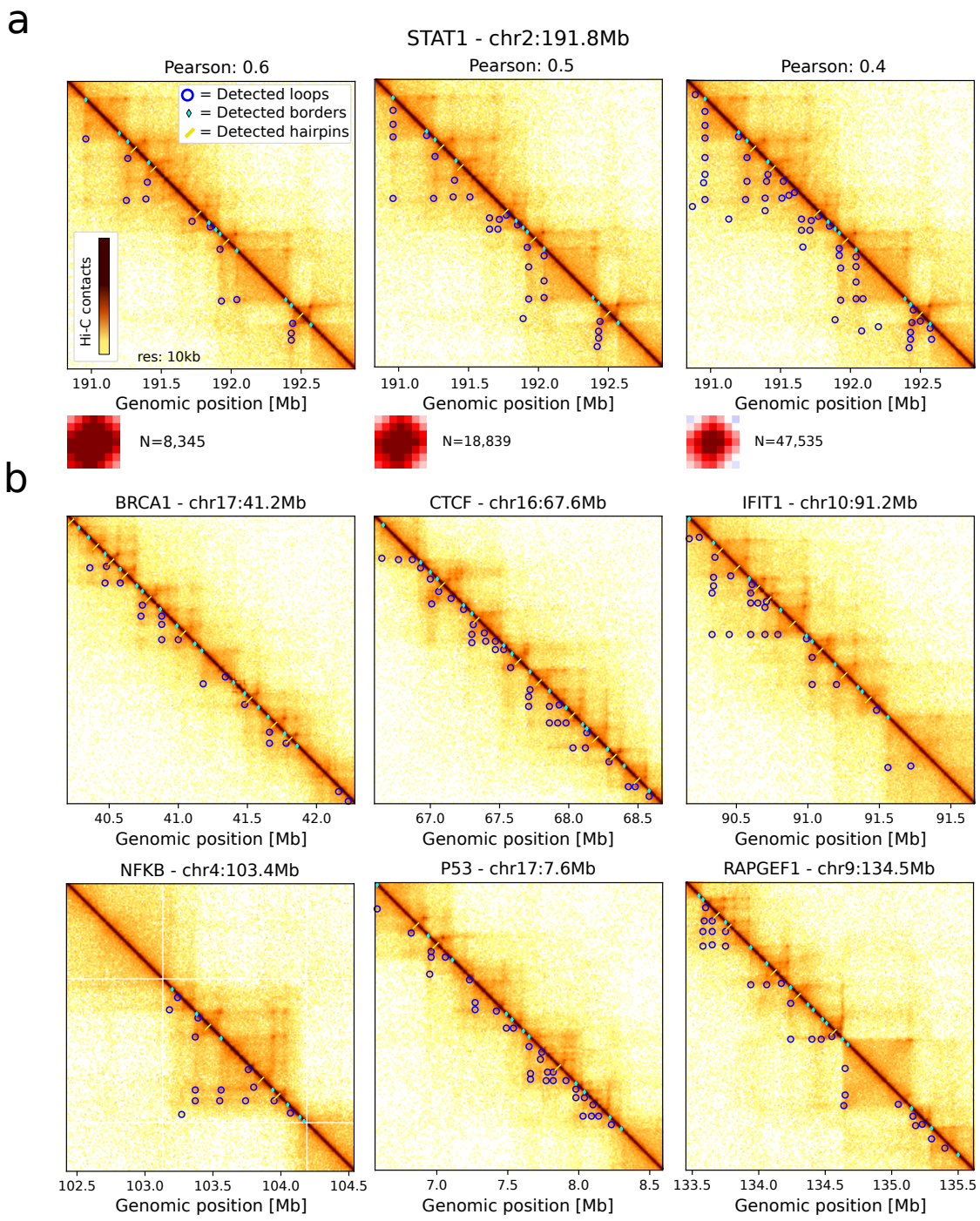
Supplementary Figure 4: Detection of loop and border patterns in yeast contact data a, Detection of loops and borders in Hi-C data of *S. cerevisiae* synchronised in G1 and for a mutant depleted in the protein Pds5 [2]. **b**, Bar plots showing enrichment in highly expressed genes (HEG) for detected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



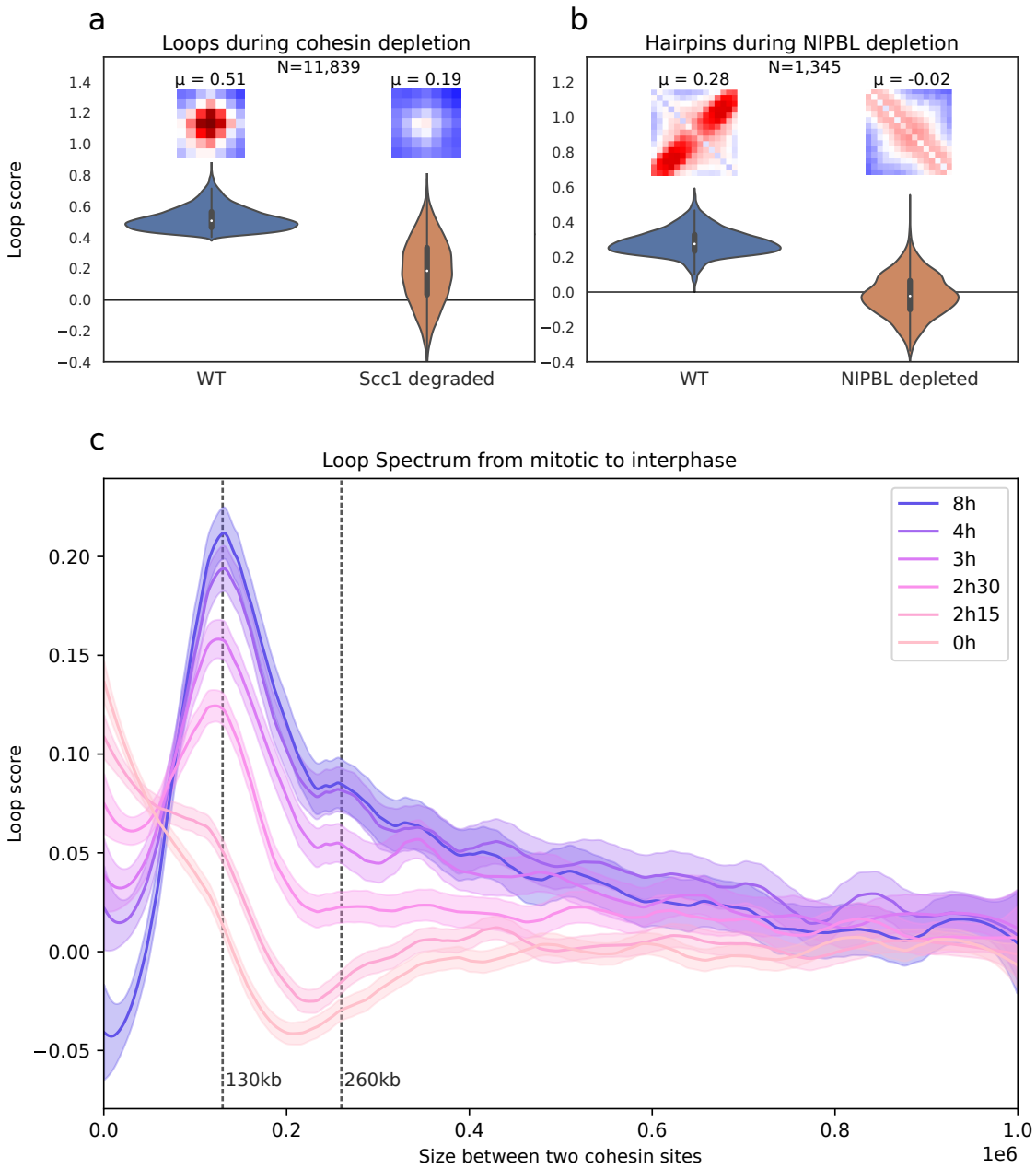
Supplementary Figure 5: Applications of quantification mode on yeast contact data a, Chromosight quantification mode workflow: sub-matrices from certain 2D genomic positions are extracted from balanced and detrended matrices (as in detection mode). Correlation with the kernel is then computed for each 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.



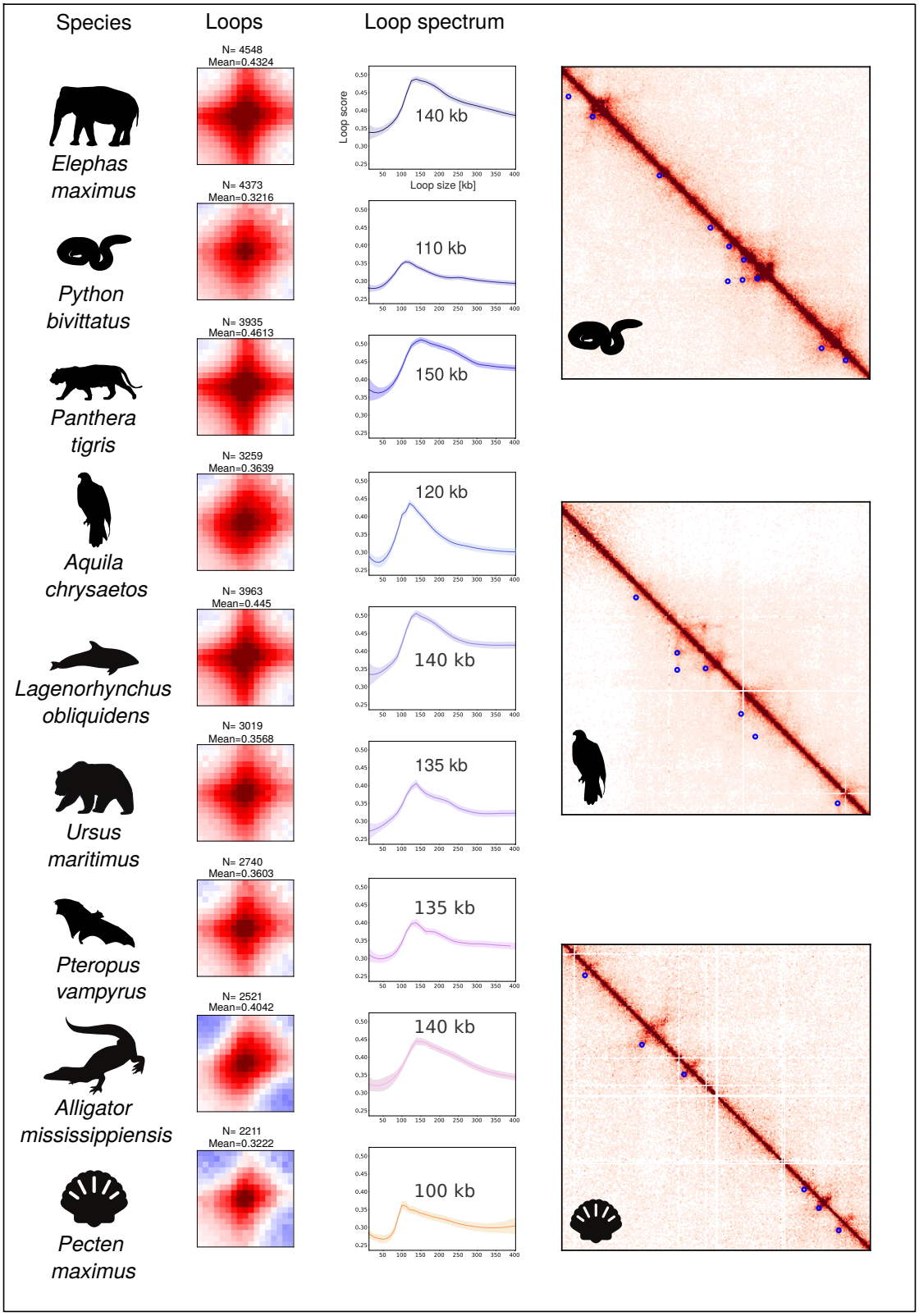
Supplementary Figure 6: Toy models that can link visual patterns and physical models. **Model 1**: loop extruding motors with two roadblock points leading to a loop pattern. **Model 2**: loop extruding motors with a specific loading point leading to a hairpin pattern. **Model 3**: loop extruding motors 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].



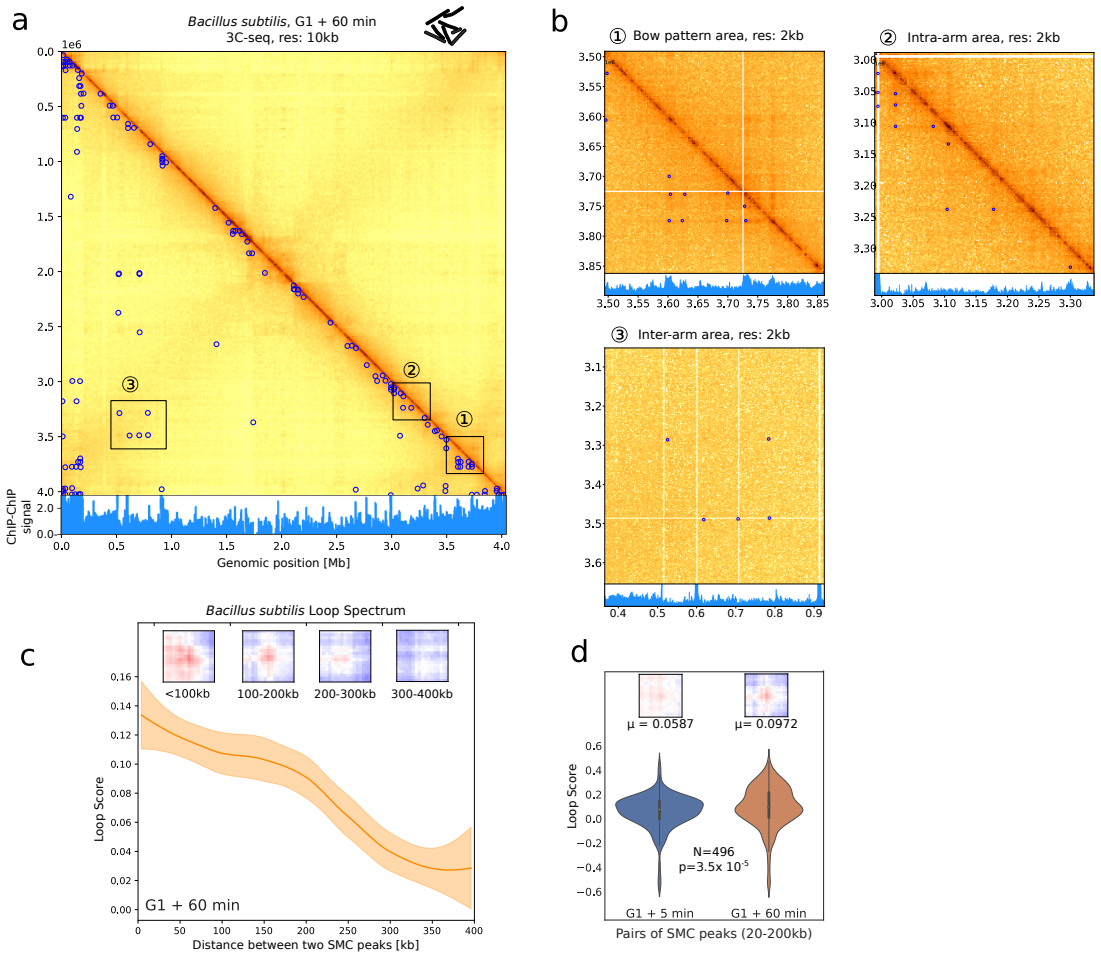
Supplementary Figure 7: Detection of loops, borders, hairpins in human Hi-C data. a, Effect of decreasing Chromosight's Pearson coefficient detection threshold on loop detection. Contact map in the vicinity of STAT1 gene in Hi-C data of lymphoblastoids [11] and total number of detected loops are shown 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 classic problem in the field of signal processing and remote sensing. Given two or more input signals such as images, one wants to identify the portions that differ between them. This problem also applies to Hi-C contact maps, where it natural to want to detect regions of contact maps that differ between biological conditions, indicating changes in the organization of chromosomes. Importantly, the significance of these differences have to be assessed, given the sparsity of many areas of the contact maps (especially between DNA segments separated by long distances in *cis*).

Several computational solutions have been developed to tackle this challenge. Some of them, such as diffhic [108], 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, however it only finds non-specific variations in the amount of contacts. These variations can reflect specific spatial interactions, but also differential compartment switches or insulation changes, which could be caused by a number of different phenomena.

When I started working on Hi-C data, there was therefore a need for a program that would detect significant changes in contact maps with a focus on specific chromatin features. We developed pareidolia[1], a software package for change detection with an a priori 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 convolution engine 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 similarity, 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 $t = \{t_0, t_1\}$, each with multiple biological samples (replicates) $r = \{r_1, r_2, ..., r_R\}$. The whole genome contact matrix from each sample $H_{r,t}^{n \times n}$ is first processed using Chromosight's convolution algorithm as described in

[1] Pareidolia: From Ancient Greek παρα (para, "alongside, concurrent") + εἴδωλον (eídōlon, "image"): the tendency to interpret a vague stimulus as something known to the observer, such as interpreting marks on Mars as canals, seeing shapes in clouds, or hearing hidden messages in music.

1.2.4 to generate a matrix of similarity $M_{r,t}^{n\times n}$ with kernel K representing the pattern of interest. In the resulting matrix, the value at position (i, j), thereafter denoted $M_{r,t}[i,j]$, is a Pearson correlation coefficient with K, therefore $M_{r,t}[i,j] \in [-1,1]$.

Change detection is applied using an algorithm inspired by median filtering-based background formation [165] (Fig. II.G). First, we generate a background matrix B_t for each condition (timepoint), whose values are defined as the median of all replicates' Chromosight correlation maps from that condition (Eq. 1.1). The change matrix D is then obtained by taking the difference between condition backgrounds (Eq. 1.2). Note that, although values in B_t and D are computed independently at each position [i, j], the spatial dependency between values is taken into account through the convolution operation used to produce $M_{r,t}$.

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

$$D[i,j] = B_{t_1}[i,j] - B_{t_0}[i,j] \tag{1.2}$$

We then compute the matrix of standard errors S between each replicate and their condition's median background (Eq. 1.3). This technical variability (among replicates) is then used to filter out noisy regions. This is done by generating a Contrast-to-Noise Ratio (CNR) map C (Eq. 1.4) and applying a threshold to it. This matrix represents the ratio of contrast (inter-condition difference) over noise (intra condition variability).

$$S = \frac{1}{T}\sum_{t=0}^{T}\sqrt{\frac{1}{R}\sum_{r=1}^{R}(M_{r,t} - B_t)^2} \tag{1.3}$$

$$C = \frac{|B_{t_1} - B_{t_0}|}{S} \tag{1.4}$$

Three coordinates sets are then defined to select regions of interest for change detection:

- Positions with a local contact density above T_d. If a kernel K of size $m_K \times n_K$ is used to compute M, the local contact density L is defined as the proportion of nonzero contact values within a window of $m_K \times n_k$ centered on $H[i,j]$ (Eq. 1.5). Each position must be above a threshold in all biological samples to be considered (Eq. 1.6). This set is used to select for sufficient coverage.

- Positions for which at least one biological sample from either condition has a Chromosight score above threshold T_p (Eq. 1.7). This set is used to select for clear patterns.

- Pixels in regions with a CNR above threshold T_c (Eq. 1.8). This set is used to select for high contrast between conditions.

The required thresholds T_c, T_p and T_d are provided with default values but can also be set by the user. The intersection F of the 3 resulting coordinate sets is then computed (Eq. 1.9) and the change matrix D is filtered to retain only coordinates present in F (Eq. 1.10), effectively retaining change values passing all conditions and setting others to 0.

$$L_{r,t}[i,j] = \sum_{x=1}^{m_K} \sum_{y=1}^{n_K} (1 - \delta(0, H_{r,t}[i + x - \frac{m_K + 1}{2}, j + y - \frac{n_K + 1}{2}])) \tag{1.5}$$

where $\delta(a, b)$ is the Kronecker delta:

$$\delta(a,b) = \begin{cases} 0, & a \neq b \\ 1, & a = b \end{cases}$$

$$F_d = \{(i,j) \mid \min_{r \in R} L_{r,.}[i,j] \geq T_d\} \tag{1.6}$$

$$F_p = \{(i,j) \mid \max_{r \in R} M_{r,.}[i,j] \geq T_p\} \tag{1.7}$$

$$F_c = \{(i,j) \mid C[i,j] > T_c\} \tag{1.8}$$

$$F = F_c \wedge F_p \wedge F_d \tag{1.9}$$

$$D_f[i,j] = \begin{cases} D[i,j], & (i,j) \in F \\ 0, & otherwise \end{cases} \tag{1.10}$$

Pareidolia can either return the pattern intensity changes D_f at a set of predetermined (i, j) positions provided as input, or perform a *de-novo* detection of differential loops on the Hi-C map. For *de-novo* detection, pareidolia applies Chromosight's implementation of the connected component labelling algorithm for sparse matrices, described in 1.2.4. This operation isolates contiguous foci of non-zero differential changes in D_f and retrieve the local absolute maximum in each focus. The list of coordinates along with their respective differential intensity score is returned.

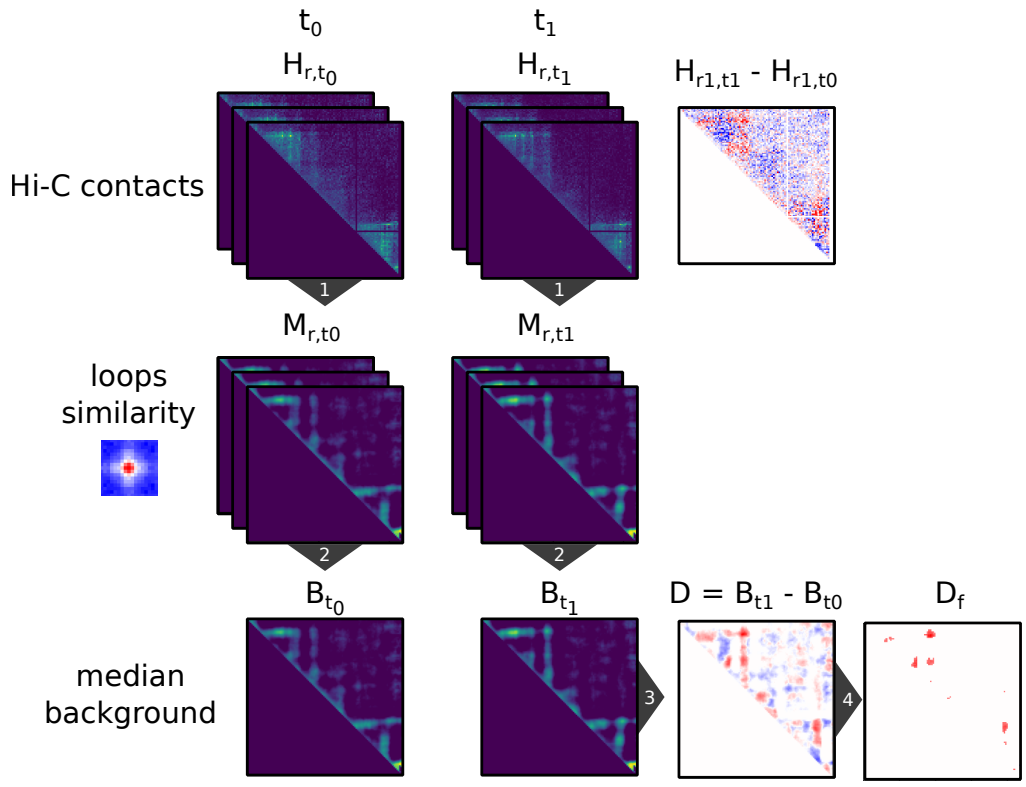


Fig. II.G: Pareidolia algorithm. From top to bottom: The Hi-C contact maps of several replicates (r) in two conditions (t_0, t_1) are shown, as well as the difference between conditions. Chromosight's convolution algorithm is used on each sample **1** to generate a map of correlation coefficients (M) with the kernel of interest (loops in this case). For each condition, a median background (B) is computed among replicates **2**. The difference between these two backgrounds (D) is then extracted **3** and filtered **4** using a combination of CNR, contact density and pattern intensity thresholds to obtain the final change matrix (D_f).

The inner steps of Pareidolia are further detailed with visual representations of the intermediate matrices and filters on experimental data in Appendix C.

1.3.2 Results on experimental data

To showcase the use of Pareidolia, we used it to measure loop changes upon depletion of CTCF in murine cells using published data [166]. 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.

These looping interactions have been shown to be weakened or disappear in the absence of cohesin [86] or CTCF [166]. Here we show an example use of Pareidolia to quantify these 3D changes.

The dataset consists of CTCF-AID mutant mouse 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 2 replicates per condition.

With default parameters, pareidolia identifies a total of 2,997 disappearing differential loops and 845 appearing loops (Fig. II.H).

1.3.3 Perspectives and potential improvements

Pareidolia is one of the few available computational methods for differential Hi-C analysis which leverages replicates [108, 109, 167–169]. While these methods focus on differential contact enrichment, often through fitting a Poisson or negative binomial model, Pareidolia detects change relative to a predefined pattern. In the case of Hi-C data, we estimated that this was a proper solution, given the emphasis of most if not all investigations on limited number of specific chromatin features such as loops, stripes, etc. On the other hand, one may miss some unexpected changes in some conditions, but these cases remain rare. Pareidolia is computationally efficient due to the use of sparse matrices throughout the program and maximal use of vectorized operations. Although a formal benchmark has yet to be performed, its core functionality relies on Chromosight which has been thoroughly evaluated, and gives promising results on empirical tests.

Although Pareidolia currently supports only two condition, the code was written with the intent of being extended to multiple conditions or timepoints. This would most likely involve a different change metric than background accumulation, such as a regression, but could be implemented with relatively few modifications to the code. A limitation of Pareidolia is the reliance on threshold values to filter noise and filter differences. It could be possible to solve this issue by automatically selecting thresholds based on the distribution of contacts and similarity scores in the input samples.

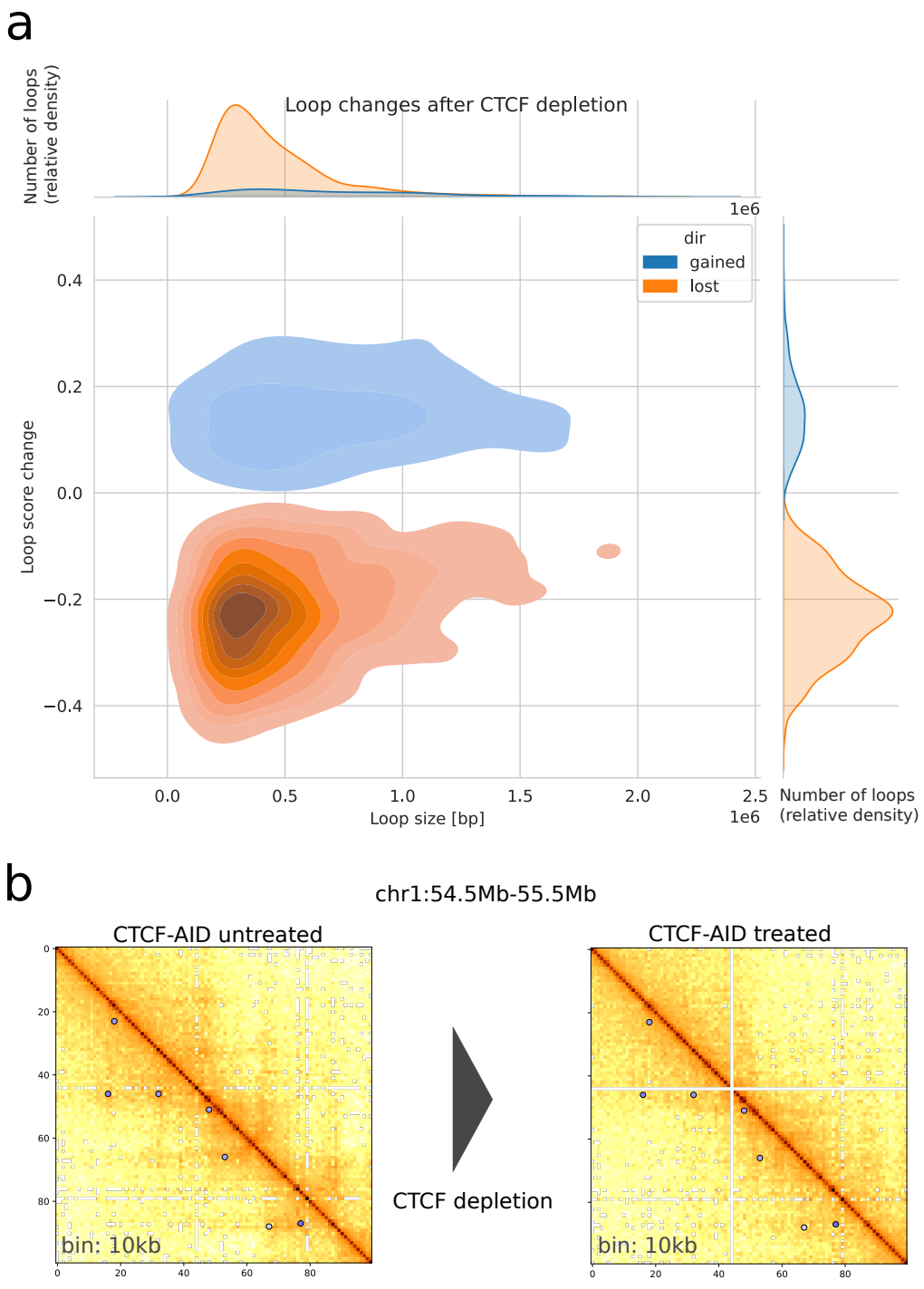


Fig. II.H: Pareidolia results on CTCF degradation experiments from [166]: **a:** Distribution of chromatin loop change and size upon CTCF depletion as detected by pareidolia. **b:** Zoom on a region of the Hi-C map from mouse chromosome 1. Disappearing loops detected by pareidolia are highlighted in 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*

Legionella pneumophila alters its host signal transduction, metabolism and gene regulation upon infection [170, 171]. In addition to all these changes, it also affects host histone marks [37], which are known to be related to gene regulation and genome architecture [64]. In this chapter, we investigate the genome structure of the amoeba *A. castellanii*, a natural host of *L. pneumophila*, and how it is affected during infection by the bacterium.

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

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 3192 contigs merged into 384 scaffolds which do not represent chromosomes.

This prompted us to generate a chromosome-level genome assembly for two strains of *A. castellanii*: The Neff strain [172], which is the most widely used strain for genomic analyses in that species, and the C3 strain [173], which is generally used for infection experiments with *L. pneumophila* due to higher intracellular bacterial replication. In the following manuscript, we describe and compare the genomic landscape of both strains, as well as the spatial organization of their genomes. We also investigate changes happening in the 3D genome organization in response to infection by *L. pneumophila*.

RESEARCH

Chromosome-scale assemblies of Acanthamoeba castellanii genomes provide insights into Legionella pneumophila infection-related chromatin re-organization

Cyril Matthey-Doret[1,2†], Morgan J. Colp[3†], Pedro Escoll[4], Agnès Thierry[1], Bruce Curtis[3], Matt Sarrasin[5], Michael W. Gray[3], B. Franz Lang[5], John M. Archibald[3*], Carmen Buchrieser[4*] and Romain Koszul[1*]

[*]Correspondence:
john.archibald@dal.ca;
cbuch@pasteur.fr;
romain.koszul@pasteur.fr
[3]Department of Biochemistry and Molecular Biology and Institute for Comparative Genomics, Dalhousie University, Sir Charles Tupper Medical Building, 5850 College Street, B3H 4R2 Halifax, Nova Scotia, Canada
[4]Institut Pasteur, Université de Paris, CNRS UMR3525, Unité Biologie des Bactéries Intracellulaires, F-75015, Paris, France
[1]Institut Pasteur, Université de Paris, CNRS UMR3525, Unité Régulation Spatiale des Génomes, F-75015, Paris, France
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

The unicellular amoeba *Acanthamoeba castellanii* is ubiquitous in aquatic environments, where it preys on bacteria. The organism also hosts bacterial endosymbionts, some of which are parasitic, including human pathogens such as *Chlamydia* and *Legionella* spp. Here we report complete, high quality genome sequences for two extensively studied *A. castellanii* strains, Neff and C3. Combining long- and short-read data with Hi-C, we generated near chromosome-level assemblies for both strains with 90% of the genome contained in 29 scaffolds for the Neff strain and 31 for the C3 strain. Comparative genomics revealed strain-specific functional enrichment, most notably genes related to signal transduction in the C3 strain, and to viral replication in Neff. Furthermore, we characterized the spatial organization of the *A. castellanii* genome and showed that it is reorganized during infection by *Legionella pneumophila*. Infection-dependent chromatin loops were found to be enriched in genes for signal transduction and phosphorylation processes. In genomic regions where chromatin organization changed during *Legionella* infection, we found functional enrichment for genes associated with metabolism, organelle assembly, and cytoskeleton organization, suggesting that changes in chromosomal folding are associated with host cell biology during infection.

Keywords: amoeba; genome organization; evolution; assembly; Hi-C

Introduction

The first amoebae were isolated in 1913 [1], and the genus *Acanthamoeba* was established in 1931 by Volkonsky [2]. It comprises different species of free living, aerobic, unicellular protozoa, present throughout the world in soil and nearly all aquatic environments [3]. The life cycle of *Acanthamoeba* includes a dormant cyst with minimal metabolic activities under harsh conditions and a motile trophozoite that can feed on small organisms and reproduce by binary fission in optimal conditions [4]. *Acanthamoeba* is perhaps most widely known from its role as a human pathogen, acting to cause the vision-threatening eye infection *Acanthamoeba* keratitis, but it can also cause serious infections of the lungs, sinuses, skin, and a central nervous system disease called granulomatous amoebic encephalitis [5]. The species *Acanthamoeba castellanii* was first isolated in 1930 by Castellani as a contaminant of a yeast culture [6].

In their natural environment, *Acanthamoeba* spp. are characterized by the ability to change their shape through pseudopode formation and are considered professional phagocytes as they feed on bacteria, but may also phagocytose yeasts and algae. However, some bacteria are resistant to degradation and live as endosymbionts in these protozoa, and others even use the amoeba as a replication niche. Thus *Acanthamoeba* are also reservoirs of microorganisms and viruses, including human pathogens, which have adapted to survive inside these cells and resist digestion, persist or even replicate as intracellular parasites. At least 15 different bacterial species, two archaea and several eukaryotes and viruses have been shown to interact with *Acanthamoeba* in the environment and may even co-exist at the same time within the same host cell [7].

Although it was observed early on that bacteria could resist digestion of free-living amoebae [8], it was not until the discovery that *Legionella pneumophila* replicated in amoebae that researchers began studying the bacterium-amoeba relationship in depth [9]. *L. pneumophila* is the agent responsible for Legionnaires' disease, a severe pneumonia that can be fatal if not treated promptly. In addition, many species of amoebae have the ability to form highly resistant cysts in hostile environments, providing shelter for their intracellular parasites [10]. Indeed, it is thought that *L. pneumophila* may survive water disinfection treatments and contaminate water distribution systems by encystation [11, 12, 13]. From these contaminated water sources, *L. pneumophila* can reach the human lungs via aerosols contaminated with the bacteria and replicate within the alveolar macrophages that are, like amoebae, phagocytic cells.

L. pneumophila has the ability to escape the lysosomal degradation pathway of both *A. castellanii* and human alveolar macrophages through the formation of a protective vacuole (the Legionella-containing vacuole or LCV) where it multiplies to high numbers. Once the host cell has been fully exploited and nutrients become limited, *L. pneumophila* exits the host and infects a new cell [14].

To establish the LCV and replicate, *L. pneumophila* secretes over 300 effector proteins into the host cytoplasm via a type four secretion system (T4SS) called Dot/Icm [15], thereby manipulating host pathways and redirecting nutrients to the LCV [16, 17]. In the early stages of infection, many of these proteins target the host secretory pathway, including several small GTPases, to recruit endoplasmic reticulum-derived vesicles to the LCV [18]. During the intracellular cycle, a wider range of processes, including membrane trafficking, cytoskeleton dynamics, and signal transduction pathways, are targeted by these effectors [19, 20]. *L. pneumophila* also directly alters the genome of its host by modifying epigenetic marks of the host genome in human macrophages and *A. castellanii*. It secretes an effector named RomA with histone methyltransferase activity that is targeted to the nucleus. RomA carries out genome-wide trimethylation of K14 of histone H3 [21], leading to transcriptional changes that modulate the host response in favor of bacterial survival [21]. Concomitantly, *L. pneumophila* infection leads to genome-wide changes in gene expression [22]. In many eukaryotes, gene regulation is intertwined within the three-dimensional organization of chromosomes. The functional interplay between gene regulation and higher-order chromatin elements such as loops, self-interacting domains and active/inactive compartments is actively being studied [23, 24]. Therefore, the infection of *A. castellanii* by *L. pneumophila* provides

an amenable model with which to investigate how an intracellular bacterial infection may affect the regulation of chromosome folding, and its consequences, in a eukaryotic host.

The investigation of genome organization and regulation of *A. castellanii* in response to infection requires a highly contiguous genome assembly. The reference genome sequence for *A. castellanii*, NEFF-v1 [25], is based on the Neff strain, isolated from soil in California in 1957 [26]. This assembly is widely used by different laboratories studying *A. castellanii*, but is fragmented into 384 scaffolds comprising 3192 contigs, which makes chromosome-level analyses difficult, if not impossible, and basic features of the *A. castellanii* genome, such as the number of chromosomes and ploidy, remain undetermined. In addition, many teams investigating bacteria-amoeba interactions use the "C3" strain (ATCC 50739), isolated from a drinking water reservoir in Europe in 1994 and identified as a mouse pathogen [27]. However, genomic information is scarce for this strain and little is known about its similarity to the Neff strain. Notably, these two *A. castellanii* strains have been cultivated for several decades and were isolated from different ecological niches, but the extent of conservation between their genomes is unknown. It is difficult to investigate the factors that determine the susceptibility of different *A. castellanii* strains to the pathogen without proper genomic resources. These resources would also be required to apply genome-wide omics approaches.

The goal of this work was to study how the *A. castellanii* C3 strain responds to *L. pneumophila* infection through the lens of the three-dimensional organisation of its genome. This analysis required the generation of a high quality reference genome sequence of the C3 strain, as well as a new and improved assembly of the Neff reference genome. Illumina, Nanopore long read, and Hi-C data were used to generate near chromosome-level assemblies of the genomes of both strains. Surprisingly, the new Neff and C3 assemblies have a (gap-excluded) sequence divergence of 6.7%. We find evidence for strain-specific enrichment of a handful of functions, including ones related to signal transduction in C3, and one relating to viral replication and virion assembly in Neff. Using the C3 assembly, RNA-seq and Hi-C, we were able to analyze the genome folding and expression changes of *A. castellanii* in response to the infection by *L. pneumophila*. We found infection-dependent chromatin loops to be enriched in genes involved in signal transduction and phosphorylation.

Results

The *A. castellanii* Neff and C3 genome assemblies are highly contiguous and complete

We used a combination of Illumina short reads, Oxford Nanopore long reads and Hi-C to assemble each genome to chromosome scale, with 90% of the Neff genome contained within 28 scaffolds. This is in contrast to a previous estimate of approximately 20 chromosomes inferred using pulsed-field gel electrophoresis [28]. For both the Neff and C3 strains, we first generated a raw *de-novo* assembly using Oxford Nanopore long reads. To account for the error prone nature of long reads, we polished the first draft assemblies with paired-end shotgun Illumina sequences using HyPo [29]. The polished assemblies were then scaffolded with long range Hi-C contacts using our probabilistic program instaGRAAL, which exploits a Markov Chain Monte Carlo algorithm to swap DNA segments until the most likely scaffolds are

achieved [30]. Following the post-scaffolding polishing step of the program (see [30]), the final genome assemblies displayed better contiguity (Table 1), completion, and mapping statistics than the previous versions, with the cumulative scaffold lengths quickly reaching a plateau (Fig. 1a). The assemblies of both strains are also slightly longer, with a smaller number of contigs than the original Neff assembly (NEFF-v1) (Fig. 1b). The BUSCO-completeness scores for both assemblies are also improved, with 90.6% (Neff) and 91.8% (C3) complete eukaryotic universal single copy orthologs, compared to 77.6% for NEFF-v1. We also noted an increased proportion of properly paired shotgun reads from 71% for NEFF-v1 to 84% for our new Neff assembly, suggesting a reduced number of short mis-assemblies. Hi-C contact maps present a convenient readout to explore large mis-assemblies in genome sequences [31]. While this allowed us to manually address major unambiguous mis-assemblies, a number of visible mis-assemblies remain in complex regions such as repeated sequences near telomeres and ribosomal DNAs (rDNAs). These mis-assemblies could not be resolved with the data generated herein. In the C3 assembly, there are also a few (at least 5) interchromosomal mis-assemblies which appear to be heterozygous and cannot be resolved without a phased genome. We also found shotgun coverage to be highly heterogeneous between scaffolds, which is suggestive of aneuploidy (Fig. S1).

A. castellanii strains Neff and C3 have partly non-overlapping gene complements

The generation of chromosome-scale genome assemblies for two different *A. castellanii* strains afforded us the first opportunity to compare and contrast their coding capacities. We used both Broccoli [32] and OrthoFinder [33] for inference of orthologous groups. A summary of the inferred orthogroups shared by, and specific to, the Neff and C3 strains of *A. castellanii* is presented in Figure 2, with orthogroup numbers from both orthologous clustering tools included. This figure only compares Neff against C3, irrespective of orthogroup presence or absence in outgroup taxa. In this analysis, each strain-specific gene that was not assigned to an orthogroup by either program was still considered to be a single strain-specific orthogroup in order to account for the presence of genes without any orthologs across the five species. Broccoli predicted more orthogroups overall and more strain-specific genes than OrthoFinder, but predicted fewer shared orthogroups. Despite these differences, the overall trend is similar for the two outputs. The number of orthogroups shared by the two strains is roughly an order of magnitude greater than the number specific to either strain, while the C3 strain has a greater number of strain-specific orthogroups than the Neff strain as predicted by both programs.

To investigate how similar the *A. castellanii* gene complement was to other members of Amoebozoa, *A. castellanii* orthogroups were evaluated for their presence in three outgroup species. Both Broccoli and OrthoFinder outputs were analyzed in this fashion. According to Broccoli, 43.5% of orthogroups shared by the two *A. castellanii* strains were not present in the other three amoebae, while OrthoFinder gave a figure of 48.4%. In the Neff strain, 49.1% of all orthogroups, shared or strain-specific, were not found in the three outgroup amoebae according to Broccoli, compared to 51.0% as predicted by OrthoFinder. In the C3 strain, the Broccoli results indicate that 52.4% of all orthogroups are not present in the outgroup amoebae,

while 52.8% were not found in the outgroup by OrthoFinder. This is in contrast with *A. castellanii* strain C3 sharing an estimated 82.5% (Broccoli) to 89.4% (OrthoFinder) of its orthogroups with the Neff strain, and the Neff strain sharing an estimated 88.9% (Broccoli) to 93.6% (OrthoFinder) of its orthogroups with the C3 strain.

A. castellanii accessory genes show strain-specific functional enrichment

In an attempt to gain insight into the functional significance of strain-specific genes in the C3 and Neff genomes, the top 30 most significantly enriched terms were identified by topGO and plotted in order of decreasing p-value for each strain/ontology combination (Supplementary Figures S8-S13). Notably, among C3-specific genes, only two terms were found to be statistically significantly enriched for each of the three ontologies at a 95% confidence level. Among Neff-specific genes, only one term was significantly enriched in each of the 'cellular component' and 'molecular function' ontologies, while three were significantly enriched in the 'biological process' ontology.

In C3, enriched molecular functions were 'GTP binding' (p = 5e-5) and 'protein serine/threonine phosphatase activity' (p = 0.037), enriched biological processes were 'small GTPase mediated signal transduction' (p = 8.5e-5) and 'ubiquitin-dependent protein catabolic processes' (p = 0.029), and enriched cellular components were 'RNA polymerase II core complex' (p = 0.026) and 'the Golgi membrane' (p = 0.036). In Neff, the enriched molecular function was 'DNA helicase activity' (p = 0.0071), enriched biological processes were 'telomere maintenance' (p = 0.0027), 'protein homooligomerization' (p = 0.0135), and 'DNA replication' (p = 0.0403), and the enriched cellular component was 'virion parts' (p = 0.012). When searched against the nr database with BLASTp [34], the Neff genes found to be responsible for both DNA helicase activity enrichment and telomere maintenance enrichment had their best BLAST hits to PIF1 5'-to-3' DNA helicases, those responsible for protein homooligomerization enrichment had their best BLAST hits to K+ channel tetramerization domains, and the gene annotated as being a virion part had its best BLAST hits to major capsid protein from various nucleocytoplasmic large DNA viruses (NCLDVs).

The Neff strain has a divergent mannose binding protein

One particular gene of interest encodes a mannose binding protein, which is known to be used as a receptor for cell entry by *Legionella* in some *A. castellanii* strains [35]. The MEEI 0184 strain of *A. castellanii*, an isolate from a human corneal infection, was used as a reference sequence, because it is the only strain in which the mannose binding protein is biochemically characterized [36, 37]. The orthologs from C3, Neff, and *Acanthamoeba polyphaga* were retrieved, and all four sequences were aligned (Figure S14). The percent identity of each sequence to the reference was calculated over the sites in the alignment where the *A. polyphaga* sequence was not missing (Table 2). The C3 homolog was found to be 99.5% identical to the MEEI 0184 homolog, whereas the Neff and *A. polyphaga* proteins were more divergent, sharing 91.6% and 97.2% identity to MEEI 0184, respectively. Despite being of the same species as the reference, the Neff strain homolog was found to

be much more divergent than the *A. polyphaga* sequence is from the other two *A. castellanii* strains. Interestingly, we observed that *L. pneumophila* replicates worse in the Neff strain than the C3 strain in culture. This phenotype may result from impaired receptor-mediated entry by *Legionella* into Neff cells due to differences in the receptor encoding gene.

Spatial organisation of the *A. castellanii* genome

To our knowledge, no Hi-C contact maps have been generated from species of Amoebozoa. Therefore, the Hi-C reads we used to generate the chromosome-scale scaffolding of two *A. castellanii* genomes also offer the opportunity to reveal the average genome folding in a species of this clade. Hi-C reads were realigned along the new assemblies of both the C3 and Neff strains to generate genome-wide contact maps. Visualising the Hi-C contact maps of both genomes shows that *A. castellanii* chromosomes are well resolved in our assemblies (Fig. 3). In Neff, the highest intensity contacts are concentrated on the main diagonal, suggesting an absence of large-scale mis-assemblies. On the other hand, the C3 assembly retains a few mis-assembled blocks, mostly in the rDNA region where tandem repeats could not be resolved correctly with the data available to us. However, for both strains the genome-wide contact maps reveal a grid-like pattern, with contact enrichment between chromosome extremities resulting in discrete dots. These contacts can be interpreted as a clustering of the telomeres, or subtelomeres, of the different chromosomes (Fig. 3a). Based on the presence of these inter-telomeric contacts patterns, Hi-C contact maps suggest the presence of at least 35 chromosomes in both strains, ranging from roughly 100 kbp to 2.5 Mbp in length (Fig. S15). Additionally we found 100 copies of 5S rDNA dispersed across most chromosomes for both strains, and 18S/28S rDNA genes show increased contacts with subtelomeres (Fig. 3a).

In addition to large, interchromosomal subtelomeric contacts, we also explored the existence of intrachromosomal chromatin 3D structures in the contact maps using Chromosight, a program that detects patterns reflecting chromatin structures on Hi-C contact maps [38]. For both strains, Chromosight identified arrays of chromatin loops along chromosomes, as well as boundaries separating chromatin domains (Fig. 3b). Most chromatin loops are regularly spaced, with a typical size of 20 kbp (Fig. 3c). The chromatin domains correspond to discrete squares along the diagonal (Fig. S3a). We overlapped all predicted genes in the C3 genome with the domain borders detected from Hi-C data and measured their base expression using RNA-seq we generated from that strain (see Methods). We selected the closest gene to each domain border and found that the genes overlapping domain boundaries are overall more highly expressed than those that do not (Fig. S2c). In addition, the analysis showed that gene expression is negatively correlated with the distance to the closest domain border (Fig. S2d). We performed the same comparison using chromatin loop anchors instead of domain borders. To a lesser extent, genes overlapping chromatin loops are also associated with higher expression (Fig. S2a), although it is not correlated with the distance from the closest loop (Fig. S2b). Altogether, these results suggest that the chromatin structures observed in *cis* are both associated with gene expression, although the association between gene expression and chromatin loop anchors is likely due to their co-localization with domain

borders (Fig. S2e). Some microorganisms (e.g. budding yeasts and euryarchaeotes) organize their chromosomes into micro-domains that correspond to expressed genes [39, 40]. Our findings in *A. castellanii* bear an interesting similarity to this type of organization.

L. pneumophila infection induces chromatin loop changes enriched in infection-related functions

The generation of near-complete assemblies allowed us to tackle the question of whether *L. pneumophila* infection impacts the 3D folding and transcription of the *A. castellanii* C3 strain genome. We harvested cultured *A. castellanii* cells before and 5 hours following infection by *L. pneumophila* strain Paris [41] (Methods). The cells were processed using Hi-C and RNA-seq (Methods), and the resulting reads aligned against the reference genome to assess changes in the genome structure and the host transcription program, respectively. RNA-seq was performed in triplicate, and Hi-C in duplicate (Methods). To measure changes in *trans*-chromosomal contacts, we merged the contact maps from our replicates and applied the serpentine adaptive binning method to improve the signal-to-noise ratio [42]. We then computed average interactions between each pair of chromosomes before and after infection. For each pair of chromosomes, we then used the log ratio of infected over uninfected average contacts. Following infection a global decrease in *trans*-subtelomeric contacts was observed, suggesting a slight de-clustering of chromosome ends (Fig. 4b). In addition, the scaffold bearing 18S and 28S rDNA (scaffold_29), as well as two other small scaffolds (35 and 36) displayed weaker interactions with other scaffolds during infection (Fig. 4a).

We then assessed whether the behavior of *cis* contacts changes during infection. First, we computed the average contact frequencies according to genomic distance p(s) (Methods), which is a convenient way to unveil variations in the compaction state of chromatin [43]. The p(s) curves show a global increase in long range contacts following infection (Fig. S4b). The strengths of chromatin loops and domain borders before and 5h after infection were quantified using Chromosight [38]. However, no significant average increase or decrease in the intensity of these structures (Fig. S4a) was identified when computed over the whole genome. To focus on infection-dependent chromatin structures, we filtered the detected patterns to retain those showing the top 20% strongest change in Chromosight score during infection (either appearing or disappearing). We performed a GO term enrichment analysis for genes associated with infection-dependent chromatin loops (Methods). A significant enrichment for Rho GTPase and phosphorelay signal transduction, protein catabolism and GPI biosynthesis was found (Fig. S6a). The strongest loop changes were associated with genes encoding Rho GTPase, GOLD and SET domains as well as genes for proteins containing leucine-rich repeats and ankyrin repeats (Fig. S7).

We followed the same procedure for domain borders and found that genes associated with infection-dependent domain borders were significantly enriched in 'amino acid transport', 'cyclic nucleotide biosynthetic process', 'protein modification' and 'deubiquitination' (Fig. S6b). Our results suggest that domain borders are generally associated with highly transcribed metabolic genes, consistent with previous findings showing that such borders are associated with high transcription [44].

By analyzing the *A. castellanii* RNA-seq data after infection with *L. pneumophila*, we revealed that the expression of genes was globally impacted at 5h post infection compared to uninfected cells (Fig. S5a). This is consistent with recent results showing that transcription is globally disrupted in *A. castellanii* Neff following infection by *L. pneumophila* [22]. To investigate the relationship between this change in gene expression and chromatin structure, we assigned the closest domain border to each gene and compared their expression and border score changes during infection. For the majority of genes, we found border intensity not to be correlated with gene expression changes (Fig. S5b). Only genes undergoing extreme expression changes during infection corresponded to changes in associated borders (Fig. S5c). This raises the possibility that insulation domains in *A. castellanii* chromosomes do not dictate gene expression programs as they do in mammals.

Recently, Li et al. [22] investigated gene expression changes at 3, 8, 16 and 24h after infection of *A. castellanii* Neff by *L. pneumophila*. To further validate our finding that chromatin domains are not units of regulation in *A. castellanii*, we used these expression results and migrated the gene annotations to our C3 assembly using liftoff [45]. This allowed us to compute co-expression between gene pairs during infection (i.e., expression correlation). We found that gene pairs within the same chromatin domain did not have a higher co-expression than gene pairs from different domains at similar genomic distances (Fig. S3d).

Discussion

Chromosome-level assembly uncovers *A. castellanii* genome organization

Generation, analysis and comparison of the genome sequences of two *A. castellanii* strains revealed heterogeneous coverage across scaffolds, which is consistent with previous findings that *A. castellanii* has a high but variable ploidy of approximately 25n [46]. Previous estimates of the *A. castellanii* Neff karyotype using pulsed-field gel electrophoresis estimated 17 to 20 unique chromosomes ranging from 250 kbp to just over 2 Mbp [28], while our estimate suggests at least 35 unique chromosomes with a similar size range of 100 kbp to 2.5 Mbp. The discrepancy between the number of bands in the electrophoretic karyotype and our estimate may result from chromosomes of similar size co-migrating on the gel, which we were able to resolve using sequence- and contact-based information.

Considering features of the nuclear biology of *A. castellanii*, such as suspected amitosis [47] and probable aneuploidy, our finding that 5S ribosomal DNA is dispersed across all chromosomes may serve to ensure a consistent copy number of 5S rDNA in daughter cells.

It was previously estimated that *A. castellanii* has 24 copies of rDNA genes per haploid genome [48]. Our data show that both strains contain 4 times as many copies as originally thought. The decrease in interchromosomal contacts with rDNA-containing scaffolds during infection may reflect an alteration in the nucleolus structure, probably caused by a global increase in translational activity. This would be consistent with the global transcription shift observed in RNA-seq under infection conditions.

At a first glance, the contact maps show a clustering of subtelomeric regions, but do not display a Rabl conformation, where centromeres cluster to the spindle-pole

body [49]. However, the precise positions of centromeres would be needed to verify that they do not co-localize with subtelomeric regions.

Changes in chromatin structure likely reflect transcriptional changes

Infection of *A. castellanii* with *L. pneumophila* induced significant changes in chromatin loops and borders. Our analyses showed an enrichment in several interesting GO terms at the sites of these infection-induced changes, many of them consistent with known biological processes induced by *L. pneumophila* in amoebae and macrophages. Several enriched terms are related to cell cycle regulation, including mitotic cell cycle, cell cycle processes and cell cycle checkpoints (Fig. S6), which is consistent with recent results showing that *L. pneumophila* prevents proliferation of its natural host *A. castellanii* [50, 22]. *L. pneumophila*-induced alterations of the host cell cycle may serve to avoid cell cycle phases that restrict bacterial replication [51], or to prevent amoebal proliferation, which has been proposed to increase the feeding efficiency of individual amoebae [52].

Several other GO terms that we found to be enriched at infection-dependent loops or borders are related to host cell organelles, such as organelle assembly, microtubule cytoskeleton organization, protein localization to endoplasmic reticulum, mitochondrion organization, electron transport chain, or mitochondrial respiratory chain complexes (Fig. S6). This is interesting given that it is well known that during infection, *L. pneumophila* hijacks host organelles such as the cytoskeleton, the endoplasmic reticulum, and mitochondria in both amoebae and macrophages [53, 54, 55]. Indeed, mitochondrial respiration and electron transport chain complexes were recently shown to be altered in macrophages during *L. pneumophila* infection [54, 56].

Sites of infection-dependent chromatin reorganization also show enrichment in functions related to changes in the general metabolism of the host, such as biosynthetic and catabolic processes, including nucleotide and nucleoside synthesis, lipid metabolism, or transport of amino acids and metal ions. To replicate intracellularly, *L. pneumophila* acquires all its nutrients from the cytoplasm of the host cell. Therefore, it is thought that bacteria-induced modulation of the host metabolism is key to establishing a successful infection [57]. In summary, many of the GO terms associated with changes in chromatin loops and borders during infection align with the known biology of *Legionella* infection, suggesting a link between chromatin organization and many of the observed changes in host cells during infection.

It was previously shown that *L. pneumophila* infection halts host cell division and is associated with a decrease of mRNA of the *A. castellanii* CDC2b gene, a putative regulator of the *A. castellanii* cell cycle [50]. The large scale 3D changes we observed in chromatin compaction (Fig. S4b) and interchromosomal contacts (Fig. 4) are reminiscent of cell cycle changes in yeast and could suggest that the bacterium stops the host's cell cycle at a specific checkpoint.

We identified an array of regularly spaced chromatin loops in *A. castellanii* chromosomes of approximately 20 kbp in size. This is consistent with size range of chromatin loops observed in *S. cerevisiae* during the G2/M stage [58]. This similarity in terms of regularity and size suggests that chromatin loops in *A. castellanii* may serve a similar purpose of chromosome compaction for cell division as in yeast. Our finding that DNA loop anchors and domain borders overlap highly expressed

genes is also concordant with observations made in yeast and other species that domain borders are preferentially located at highly expressed genes [38, 59], and could result from their role in blocking the processing SMC complexes [60], potentially to avoid interferences between cohesin activity and transcription.

Unlike previously shown in Drosophila [61], we did not find an increase in co-expression of genes sharing the same contact domain in *A. castellanii*. This suggests chromatin domains may be caused by highly transcribed genes, and do not act as units of regulation.

A. castellanii accessory genes may permit environmental adaptation

Despite the substantial number of genes predicted to be strain-specific in *A. castellanii*, few functions were found to be significantly enriched in either the Neff or C3 strain set of strain-specific genes. Of these, the most biologically interesting is the enrichment of both 'small GTPase mediated signal transduction' and 'GTP binding' genes in C3. Nearly all of the genes annotated as being involved in 'small GTPase mediated signal transduction' biological processes are also annotated as having 'GTP binding' molecular functions, which is not surprising – GTP binding is an integral part of GTPase functionality. The enrichment of these two GO terms, as well as protein serine/threonine phosphatase activity enrichment, suggests that the C3 strain may have expanded its capacity for environmental sensing and associated cellular responses by expanding gene families involved in signal transduction. Given the extensive gene repertoire in *A. castellanii* dedicated to cell signalling, environmental sensing, and the cellular response [25], which is thought to help the amoeba navigate diverse habitats and identify varied prey, it seems likely that alterations of this gene repertoire in C3 may have permitted further environmental adaptations.

Another enrichment of note is that of 'virion parts' in the Neff strain of *A. castellanii*. This enrichment comprises a single gene with a best BLAST hit to major capsid proteins in various NCLDVs, including a very strong hit to *Mollivirus sibericum*. Many NCLDVs, including *Mollivirus*, are known viruses of *Acanthamoeba* spp. [62]. Although no phylogenetic analyses were performed to investigate the origin of this major capsid protein gene in the Neff genome, it seems plausible that it was acquired by lateral gene transfer during an NCLDV infection, perhaps by *Mollivirus* or some closely related virus.

The remaining enriched functions have no obvious biological significance. They could well be non-adaptive, having been generated through gene duplication, differential loss in the other surveyed amoebae, or lateral gene transfer, without conferring any notable selective advantage. An improved understanding of *Acanthamoeba* cell and molecular biology is needed to make sense of the gene enrichment data presented herein.

Substitutions in the Neff mannose binding protein may inhibit *Legionella* entry

Alignment of the three *A. castellanii* mannose binding proteins (MBPs) and the *A. polyphaga* homolog may help explain the difference in susceptibility to *Legionella* infection between the Neff and C3 strains. The C3 strain mannose binding protein is highly similar to its counterpart in strain MEEI 0184, which was first to

be biochemically characterized. The Neff strain MBP, however, is markedly more divergent than even the *A. polyphaga* MBP, which is not known to participate in *Acanthamoeba-Legionella* interactions [63]. These results are consistent with the hypothesis that the Neff strain of *A. castellanii* is not a very good host for infection by *Legionella* due to an accumulation of amino acid substitutions in its mannose binding protein, substitutions that may prevent *Legionella* from binding to this protein during cell entry. Whether or not *A. castellanii* uses its MBP for feeding or recognition of potential pathogens like *Legionella* is at present unclear, but it is worth noting that the Neff strain has been in axenic culture since 1957, so it may be that relaxed selective pressure on this protein, combined with repeated population bottlenecking during culture maintenance, has allowed for mutations in the Neff strain MBP gene to accumulate. At the present time, without available genome data for strains more closely related to the Neff strain, it cannot be determined whether these mutations arose in nature or in culture. However, given that the divergence of the *A. polyphaga* ortholog to the MEEI 0184 strain is much less than that of the Neff strain, despite all four strains having similar lifestyles in nature, evolution of the Neff strain since being deposited in the culture collection seems likely.

Methods

Strains and growth conditions

A. castellanii strains Neff and C3 were grown on amoeba culture medium (2% Bacto Tryptone, 0.1% sodium citrate, 0.1% yeast extract), supplemented with 0.1 M glucose, 0.1 mM $CaCl_2$, 2.5 mM KH_2PO_4, 4 mM $MgSO_4$, 2.5 mM Na_2HPO_4, 0.05 mM $Fe_4O_{21}P_6$ at 20°C. *L. pneumophila* strain Paris was grown for 3 days on N-(2-acetamido)-2-amino-ethanesulfonic acid (ACES)-buffered charcoal-yeast (BCYE) extract agar, at 37 °C.

Infection timecourse

Infection of *A. castellanii* C3 with *L. pneumophila* was performed using MOI 10 over 5h in infection medium (0.5% sodium citrate supplemented with 0.1 mM $CaCl_2$, 2.5 mM KH_2PO_4, 4 mM $MgSO_4$, 2.5 mM Na_2HPO_4, 0.05 mM $Fe_4O_{21}P_6$ at 20°C. At 5h post-infection, amoebae were collected in a 15 mL tube, pelleted by centrifugation at 300 *g* for 10 minutes and washed twice in PBS, then crosslinked in 3% formaldehyde during 20 minutes at room temperature (RT) with gentle shaking. 2.5 M glycine was then added to reach a final concentration of 0.125 M over 20 minutes, centrifuged, washed, and pellets were stored at -80 °C until DNA extraction.

Library preparations

Hi-C

Cell pellets were suspended in 1.2ml H_2O and transferred to CK14 Precellys tubes. Cells were broken with Precellys (6 cycles: 30 sec ON / 30 sec OFF) at 7500 RPM and transferred into a tube. All Hi-C libraries for *A. castellanii* strains C3 and Neff were prepared using the Arima kit and protocol with only the *DpnII* restriction enzyme. Libraries were sequenced to produce 35 bp paired-end reads on an Illumina NextSeq machine.

Short-read sequencing

Illumina libraries SRX12218478 and SRX12218479 were prepared from A. castellanii strains C3 and Neff genomic DNA, respectively, and sequenced by Novogene at 2x150 bp on an Illumina Novaseq 6000 machine.

For SRX4625411, a PCR-free library was prepared and sequenced by Génome Québec from purified A. castellanii strain Neff genomic DNA. The library was barcoded and run with other samples on an Illumina HiSeq X Ten instrument, producing 150 bp paired-end reads.

RNA-seq

Poly-A selected libraries were prepared from purified *A. castellanii* total RNA. *A. castellanii* strain C3 RNA-seq libraries were prepared using the stranded mRNA Truseq kit from Illumina and sequenced in single-end mode at 150 bp on an Illumina NextSeq machine.

For *A. castellanii* strain Neff (SRX7813524), the library was prepared and sequenced by Génome Québec. The library was barcoded and run with other samples on an Illumina NovaSeq 6000 instrument, producing 300 bp paired-end reads.

Nanopore sequencing

For SRX12218489 and SRX12218490, DNA was extracted from *A. castellanii* strains Neff and C3 using the QIAGEN Blood and Cell Culture DNA Kit (Qiagen) following the specific recommendations detailed by Oxford Nanopore Technologies in the info sheet entitled "High molecular weight gDNA extraction from cell lines (2018)" in order to minimize DNA fragmentation by mechanical constraints. Nanopore libraries were prepared with the ligation sequencing kit LSKQ109, flowcell model MIN106D R9. Basecalling was performed using Guppy v2.3.1-1.

For other libraries, genomic DNA samples were obtained from *A. castellanii* strain Neff using an SDS-based lysis method, followed by digestion with RNase A, then proteinase K, and then a phenol-chloroform-based extraction. DNA samples were cleaned with QIAgen G/20 Genomic Clean-up columns using the manufacturer's protocol, but with double the number of wash steps. Four different libraries were prepared, using the SQK-RAD003 Rapid Sequencing Kit (SRX4620962), the SQK-LSK308 1D2 Ligation Sequencing Kit (SRX4620963), the SQK-RAD004 Rapid Sequencing Kit (SRX4620964), and the SQK-LSK108 Ligation Sequencing Kit (SRX4620965). The SQK-LSK308 and SQK-RAD003 libraries were sequenced on FLO-MIN107 flow cells, and the SQK-LSK108 and SQK-RAD004 libraries were both sequenced on a FLO-MIN106 flow cell. All four libraries were basecalled with Albacore 2.1.7, as they were sequenced prior to the release of Guppy. Adapters were removed from the basecalled reads using Porechop v0.2.3.

Genome assembly

Nanopore reads were filtered using filtlong v0.2.0 with default parameters to keep the best 80% reads according to length and quality. Illumina shotgun libraries were used as reference for the filtering. A *de novo* assembly was generated from the raw (filtered) Nanopore long reads using flye v2.3.6 with three iterations of polishing. The resulting assembly was polished using both Nanopore and Illumina reads with

HyPo v1.0.1. Contigs from the polished assembly bearing more than 60% of their sequence or 51% identity to the mitochondrial sequence from the NEFF_v1 assembly were separated from the rest of the assembly to prevent inclusion of mitochondrial contigs into the nuclear genome during scaffolding. Polished nuclear contigs were scaffolded with Hi-C reads using instagraal v0.1.2 with default parameters. Instagraal-polish was then used to fix potential errors introduced by the scaffolding procedure. Mitotic contigs were then added at the end of the scaffolded assembly and the final assembly was polished with the Illumina shotgun library data using two rounds of pilon polishing. The resulting assembly was edited manually to remove spurious insertion of mitochondrial contigs in the scaffold and other contaminants. The final assembly was polished again using pilon with Rcorrector-corrected reads [64]. Minimap2 v2.17 [65] was used for all long reads alignments, and bowtie2 v2.3.4.1 for short reads alignments.

Genome annotation

The structural genome annotation pipeline employed here was implemented similarly as described in [66]. Briefly, RNA-Seq reads were mapped to the genome assembly using STAR v2.7.3a [67], followed by both *de novo* and genome-guided transcriptome assembly by Trinity v2.12.0 [68]. Both runs of Trinity were performed with jaccard clipping to mitigate artificial transcript fusions. The resulting transcriptome assemblies were combined and aligned to the genome assembly using PASA v2.4.1 [69]. Protein sequences were aligned to the genome using Spaln v2.4.2 [70] to recover the most information from sequence similarity. The *ab initio* predictors employed were Augustus v3.3.2 [71], Snap [72], Genemark v4.33 [73], and CodingQuarry v2.0 [74]. Finally, the PASA assembly, Spaln alignments, as well as Augustus, Snap and Codingquarry gene models, were combined into a single consensus with Evidencemodeler v1.1.1 [75].

Functional annotations were added using funannotate v1.5.3. [76] Repeated sequences were masked using repeatmasker. Predicted proteins were fed to Interproscan v5.22 [77], Phobius v1.7.1 [78] and Eggnog-mapper v2.0.0 [79] were used to generate functional annotations. Ribosomal RNA genes were annotated separately using RNAmmer v1.2 [80] with HMMER 2.3.2.

As described in the Availability of data and materials section, the funannotate-based script "func_annot_from_gene_models.sh" used to add functional annotations to existing gene models is provided in the Zenodo record and on the associated github repository.

Analysis of sequence divergence

To compute the proportion of substituted positions in aligned segments between the C3 and Neff strains, the two genomes were aligned using minimap2 with the map-ont preset and -c flag. The gap-excluded sequence divergence (mismatches / (matches + mismatches) was then computed in each primary alignment and the average of divergences (weighted by segment lengths) was computed. This is implemented in the script "04_compute_seq_divergence.py" available in the genome analysis repository listed in Availability of data and materials

Orthogroup inference

Orthogroups were inferred using the predicted proteomes of both the Neff and C3 strains, with *Dictyostelium discoideum*, *Physarum polycephalum*, and *Vermamoeba vermiformis* as outgroups to improve the accuracy of orthogroup inference. The outgroup predicted proteomes were retrieved from PhyloFisher [81]. Both Broccoli [32] and OrthoFinder [33] were run with default settings for orthogroup inference.

Gene content comparison of Neff and C3 strains

Custom Python scripts were used to retrieve genes unique to each *A. castellanii* strain, as well as orthogroups that were shared between the two strains. Genes were only determined to be strain-specific or shared if both Broccoli and OrthoFinder assigned them as such; genes were excluded from the analysis if both tools did not agree. For both strains, functional assignments for each gene ID were extracted from funannotate output and tabulated. The tabulated assignments and strain-specific gene IDs were fed into the R package topGO [82] to analyze GO term enrichment in the strain-specific genes. Fisher's exact test with the weight algorithm was implemented in topGO for the Neff- and C3-specific genes for each of the three ontologies (biological process, cellular component, and molecular function). When building the GOdata objects for these three ontologies, nodeSize was set to 10 for both the biological process and molecular function ontologies, and 5 for the cellular component ontology due to the lower number of GO terms in this ontology.

Mannose Binding Protein Comparison

Mannose binding protein (MBP) amino acid sequences from three strains of *A. castellanii* (Neff, C3, and MEEI 0184) and one strain of *Acanthamoeba polyphaga* were retrieved, aligned using MAFFT-linsi (v7.475) [83], and visualized in Jalview (v2.11.1.3) [84]. The MEEI 0184 strain sequence was retrieved from NCBI (Accession: AAT37865.1), and the Neff and C3 sequences were retrieved from the predicted proteomes generated in this study with the MEEI 0184 sequence as a BLASTp [34] query. The *A. polyphaga* genome does not have a publicly available predicted proteome, so its MBP protein sequence was manually extracted from several contigs in the genome sequence (NCBI accession: GCA_000826345.1) using tBLASTn with the MEEI 0184 sequence as a query (the sequence encoding the first 8 amino acids of the protein could not be found in the genome due to a truncated contig).

Hi-C analyses

Reads were aligned with bowtie2 v2.4.1, and Hi-C matrices were generated using hicstuff v3.0.1 [85]. For all comparative analyses, matrices were downsampled to the same number of contacts using cooltools (https://www.github.com/mirnylab/cooltools) and balancing normalization was performed using the ICE algorithm [86]. Loops and domain borders were detected using Chromosight v1.6.1 [38] using the merged replicates at a resolution of 2 kbp. We measured the intensity changes in Chromosight scores during infection using pareidolia (v0.6.1) [87] on 3 pseudo replicates generated by sampling the merged contact maps, as described in [88]. This was done to account for contact coverage heterogeneity across replicates. The 20% threshold used to select differential patterns amounts to 1.2% false detections for loops and 2.3% for borders when comparing pseudo-replicates from the same condition.

Acknowledgements
We thank Axel Cournac, Laura Gomez Valero, Christophe Rusniok and Lyam Baudry for their comments on the bioinformatics analysis, Tobias Sahr for RNAseq library construction, Pierrick Moreau for his help with the optimization of the Hi-C protocol, Charlotte Cockram for her help with Nanopore sequencing, as well as all members of the Koszul lab and Buchrieser lab for stimulating discussions.

Funding
C.M.D. is supported by the Pasteur—Paris University (PPU) International PhD Program. This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 to R.K. (ERC grant agreement 771813). The C.B. laboratory is financed by the Institut Pasteur, the Fondation pour la Recherche Médicale (FRM) grant n°EQU201903007847 and the grant n°ANR-10-LABX-62-IBEID. Research in the Archibald Lab was supported by a grant from the Gordon and Betty Moore Foundation (GBMF5782). M.J.C. is supported by graduate student scholarships from NSERC and Dalhousie University. B.F.L. and M.S. were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN- 2017-05411) and by the 'Fonds de Recherche Nature et Technologie', Quebec.

Availability of data and materials
Sequencing datasets have been deposited in SRA under bioprojects PRJNA599339 and PRJNA487265. All processed data, as well as the assemblies and annotations used in this work are available on zenodo record https://zenodo.org/record/5507417. Strains supporting the findings of this study are available from the corresponding authors.
The analyses are packaged into the following snakemake pipelines available on github. Hybrid genome assembly: https://github.com/cmdoret/Acastellanii_hybrid_assembly, functional annotation of *A. castellanii*: https://github.com/cmdoret/Acastellanii_genome_annotation, analyses of genomic features in *A. castellanii*: https://github.com/cmdoret/Acastellanii_genome_analysis, changes during infection by Legionella: https://github.com/cmdoret/Acastellanii_legionella_infection.
The dataset(s) supporting the conclusions of this article are available in the Zenodo repository https://zenodo.org/record/5507417.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
All authors gave their consent for publication of this manuscript.

Authors' contributions
C.M.D and M.J.C performed analyses. P.E. Performed infection experiments and DNA extractions. M.J.C performed DNA extraction. M.J.C and C.M.D did the Nanopore sequencing. A.T. constructed Hi-C and shotgun libraries. M.J.C., C.M.D., B.C., M.S., M.W.G. and B.F.L. contributed to the curation and improvement of the genome assembly. All authors contributed to writing the manuscript.

Author details
[1]Institut Pasteur, Université de Paris, CNRS UMR3525, Unité Régulation Spatiale des Génomes, F-75015, Paris, France. [2]Collège Doctoral, Sorbonne Université, F-75005, Paris, France. [3]Department of Biochemistry and Molecular Biology and Institute for Comparative Genomics, Dalhousie University, Sir Charles Tupper Medical Building, 5850 College Street, B3H 4R2 Halifax, Nova Scotia, Canada. [4]Institut Pasteur, Université de Paris, CNRS UMR3525, Unité Biologie des Bactéries Intracellulaires, F-75015, Paris, France. [5]Robert Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, QC, Canada.

References
1. Puschkarew, B.M.: Über die Verbreitung der Süsswasser-Protozoen durch die Luft. Arch Protistent. (23), 323–362 (1913). Accessed 2021-09-07
2. Volkonsky, M.: Hartmannella castellanii Douglas et classification des Hartmannelles. Archives de zoologie expérimentale et générale de Paris **72**(3), 317–339 (1931)
3. Rodríguez-Zaragoza, S.: Ecology of Free-Living Amoebae. Critical Reviews in Microbiology **20**(3), 225–241 (1994). doi:10.3109/10408419409114556. Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/10408419409114556. Accessed 2021-04-26
4. Siddiqui, R., Khan, N.A.: Biology and pathogenesis of Acanthamoeba. Parasites & Vectors **5**(1), 6 (2012). doi:10.1186/1756-3305-5-6. Accessed 2021-03-25
5. Visvesvara, G.S., Moura, H., Schuster, F.L.: Pathogenic and opportunistic free-living amoebae: Acanthamoeba spp., Balamuthia mandrillaris, Naegleria fowleri, and Sappinia diploidea. FEMS Immunology & Medical Microbiology **50**(1), 1–26 (2007). doi:10.1111/j.1574-695X.2007.00232.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1574-695X.2007.00232.x. Accessed 2021-09-07
6. Castellani, A.: An amoeba found in culture of yeast: preliminary note. J Trop Med Hyg **33**, 160 (1930)
7. Samba-Louaka, A., Delafont, V., Rodier, M.-H., Cateau, E., Héchard, Y.: Free-living amoebae and squatters in the wild: ecological and molecular features. FEMS Microbiology Reviews **43**(4), 415–434 (2019). doi:10.1093/femsre/fuz011. Accessed 2021-04-26
8. Drozanski, W.: Fatal bacterial infection in soil amoebae. Acta Microbiologica Polonica (1952) **5**(3-4), 315–317 (1956)
9. Rowbotham, T.J.: Preliminary report on the pathogenicity of Legionella pneumophila for freshwater and soil amoebae. Journal of Clinical Pathology **33**(12), 1179–1183 (1980). doi:10.1136/jcp.33.12.1179. Accessed 2021-03-25

10. Kilvington, S., Price, J.: Survival of *Legionella pneumophila* within cysts of *Acanthamoeba polyphaga* following chlorine exposure. Journal of Applied Bacteriology **68**(5), 519–525 (1990). doi:10.1111/j.1365-2672.1990.tb02904.x. Accessed 2021-03-25
11. Pagnier, I., Raoult, D., La Scola, B.: Isolation and identification of amoeba-resisting bacteria from water in human environment by using an Acanthamoeba polyphaga co-culture procedure. Environmental Microbiology **10**(5), 1135–1144 (2008). doi:10.1111/j.1462-2920.2007.01530.x. Accessed 2021-03-25
12. Lasheras, A., Boulestreau, H., Rogues, A.-M., Ohayon-Courtes, C., Labadie, J.-C., Gachie, J.-P.: Influence of amoebae and physical and chemical characteristics of water on presence and proliferation of Legionella species in hospital water systems. American Journal of Infection Control **34**(8), 520–525 (2006). doi:10.1016/j.ajic.2006.03.007. Accessed 2021-03-25
13. Ikedo, M., Yabuuchi, E.: Ecological Studies of *Legionella* Species: I. Viable Counts of *Legionella pneumophila* in Cooling Tower Water. Microbiology and Immunology **30**(5), 413–423 (1986). doi:10.1111/j.1348-0421.1986.tb02967.x. Accessed 2021-03-25
14. Mondino, S., Schmidt, S., Rolando, M., Escoll, P., Gomez-Valero, L., Buchrieser, C.: Legionnaires' Disease: State of the Art Knowledge of Pathogenesis Mechanisms of Legionella. Annual Review of Pathology: Mechanisms of Disease **15**(1), 439–466 (2020). doi:10.1146/annurev-pathmechdis-012419-032742. Publisher: Annual Reviews. Accessed 2021-04-26
15. Kubori, T., Nagai, H.: The Type IVB secretion system: an enigmatic chimera. Current Opinion in Microbiology **29**, 22–29 (2016). doi:10.1016/j.mib.2015.10.001
16. Isberg, R.R., O'Connor, T.J., Heidtman, M.: The Legionella pneumophila replication vacuole: making a cosy niche inside host cells. Nature Reviews Microbiology **7**(1), 13–24 (2009). doi:10.1038/nrmicro1967. Accessed 2021-03-25
17. Ensminger, A.W.: Legionella pneumophila, armed to the hilt: justifying the largest arsenal of effectors in the bacterial world. Current Opinion in Microbiology **29**, 74–80 (2016). doi:10.1016/j.mib.2015.11.002
18. Swart, A.L., Gomez-Valero, L., Buchrieser, C., Hilbi, H.: Evolution and function of bacterial RCC1 repeat effectors. Cellular Microbiology **22**(10), 13246 (2020). doi:10.1111/cmi.13246
19. Hubber, A., Roy, C.R.: Modulation of host cell function by Legionella pneumophila type IV effectors. Annual Review of Cell and Developmental Biology **26**, 261–283 (2010). doi:10.1146/annurev-cellbio-100109-104034
20. Qiu, J., Luo, Z.-Q.: Legionella and Coxiella effectors: strength in diversity and activity. Nature Reviews. Microbiology **15**(10), 591–605 (2017). doi:10.1038/nrmicro.2017.67
21. Rolando, M., Sanulli, S., Rusniok, C., Gomez-Valero, L., Bertholet, C., Sahr, T., Margueron, R., Buchrieser, C.: Legionella pneumophila Effector RomA Uniquely Modifies Host Chromatin to Repress Gene Expression and Promote Intracellular Bacterial Replication. Cell Host & Microbe **13**(4), 395–405 (2013). doi:10.1016/j.chom.2013.03.004. Accessed 2021-03-25
22. Li, P., Vassiliadis, D., Ong, S.Y., Bennett-Wood, V., Sugimoto, C., Yamagishi, J., Hartland, E.L., Pasricha, S.: Legionella pneumophila Infection Rewires the Acanthamoeba castellanii Transcriptome, Highlighting a Class of Sirtuin Genes. Frontiers in Cellular and Infection Microbiology **10**, 428 (2020). doi:10.3389/fcimb.2020.00428. Accessed 2021-03-25
23. Rennie, S., Dalby, M., van Duin, L., Andersson, R.: Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. Nature Communications **9**(1), 487 (2018). doi:10.1038/s41467-017-02798-1. Accessed 2021-03-25
24. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., Heard, E.: Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature **485**(7398), 381–385 (2012). doi:10.1038/nature11049. Accessed 2021-03-25
25. Clarke, M., Lohan, A.J., Liu, B., Lagkouvardos, I., Roy, S., Zafar, N., Bertelli, C., Schilde, C., Kianianmomeni, A., Bürglin, T.R., Frech, C., Turcotte, B., Kopec, K.O., Synnott, J.M., Choo, C., Paponov, I., Finkler, A., Heng Tan, C., Hutchins, A.P., Weinmeier, T., Rattei, T., Chu, J.S., Gimenez, G., Irimia, M., Rigden, D.J., Fitzpatrick, D.A., Lorenzo-Morales, J., Bateman, A., Chiu, C.-H., Tang, P., Hegemann, P., Fromm, H., Raoult, D., Greub, G., Miranda-Saavedra, D., Chen, N., Nash, P., Ginger, M.L., Horn, M., Schaap, P., Caler, L., Loftus, B.J.: Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. Genome Biology **14**(2), 11 (2013). doi:10.1186/gb-2013-14-2-r11. Accessed 2021-03-25
26. Neff, R.J.: Purification, Axenic Cultivation, and Description of a Soil Amoeba, Acanthamoeba sp. The Journal of Protozoology **4**(3), 176–182 (1957). doi:10.1111/j.1550-7408.1957.tb02505.x. Accessed 2021-03-25
27. Michel, R., Hauröder, B.: Isolation of an Acanthamoeba Strain with Intracellular Burkholderia pickettii Infection. Zentralblatt für Bakteriologie **285**(4), 541–557 (1997). doi:10.1016/S0934-8840(97)80116-8. Accessed 2021-03-25
28. Rimm, D.L., Pollard, T.D., Hieter, P.: Resolution of Acanthamoeba castellanii chromosomes by pulsed field gel electrophoresis and construction of the initial linkage map. Chromosoma **97**(3), 219–223 (1988). doi:10.1007/BF00292964. Accessed 2021-03-25
29. Kundu, R., Casey, J., Sung, W.-K.: HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies. preprint, Bioinformatics (December 2019). doi:10.1101/2019.12.19.882506. http://biorxiv.org/lookup/doi/10.1101/2019.12.19.882506 Accessed 2021-03-25
30. Baudry, L., Guiglielmoni, N., Marie-Nelly, H., Cormier, A., Marbouty, M., Avia, K., Mie, Y.L., Godfroy, O., Sterck, L., Cock, J.M., Zimmer, C., Coelho, S.M., Koszul, R.: instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder. Genome Biology **21**(1), 148 (2020). doi:10.1186/s13059-020-02041-z. Accessed 2021-04-27
31. Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D.P., Syan, S., Guillén, N., Margeot, A., Zimmer, C., Koszul, R.: High-quality genome (re)assembly using chromosomal contact data. Nature Communications **5**(1), 5695 (2014). doi:10.1038/ncomms6695. Accessed 2021-03-25
32. Derelle, R., Philippe, H., Colbourne, J.K.: Broccoli: Combining Phylogenetic and Network Analyses for

Orthology Assignment. Molecular Biology and Evolution **37**(11), 3389–3396 (2020). doi:10.1093/molbev/msaa159. Accessed 2021-10-06
33. Emms, D.M., Kelly, S.: OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology **20**(1), 1–14 (2019). doi:10.1186/s13059-019-1832-y. Number: 1 Publisher: BioMed Central. Accessed 2021-10-06
34. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2. Accessed 2021-10-06
35. Declerck, P., Behets, J., De Keersmaecker, B., Ollevier, F.: Receptor-mediated uptake of Legionella pneumophila by Acanthamoeba castellanii and Naegleria lovaniensis. Journal of Applied Microbiology **103**(6), 2697–2703 (2007). doi:10.1111/j.1365-2672.2007.03530.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2672.2007.03530.x. Accessed 2021-10-04
36. Garate, M., Cao, Z., Bateman, E., Panjwani, N.: Cloning and Characterization of a Novel Mannose-binding Protein of Acanthamoeba*. Journal of Biological Chemistry **279**(28), 29849–29856 (2004). doi:10.1074/jbc.M402334200. Publisher: Elsevier. Accessed 2021-10-04
37. Garate, M., Cubillos, I., Marchant, J., Panjwani, N.: Biochemical Characterization and Functional Studies of Acanthamoeba Mannose-Binding Protein. Infection and Immunity **73**(9), 5775–5781 (2005). doi:10.1128/IAI.73.9.5775-5781.2005. Publisher: American Society for Microbiology. Accessed 2021-10-04
38. Matthey-Doret, C., Baudry, L., Breuer, A., Montagne, R., Guiglielmoni, N., Scolari, V., Jean, E., Campeas, A., Chanut, P.H., Oriol, E., Méot, A., Politis, L., Vigouroux, A., Moreau, P., Koszul, R., Cournac, A.: Computer vision for pattern detection in chromosome contact maps. Nature Communications **11**(1), 5795 (2020). doi:10.1038/s41467-020-19562-7. Number: 1 Publisher: Nature Publishing Group. Accessed 2021-05-07
39. Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., Rando, O.J.: Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell **162**(1), 108–119 (2015). doi:10.1016/j.cell.2015.05.048. Accessed 2021-08-23
40. Cockram, C., Thierry, A., Gorlas, A., Lestini, R., Koszul, R.: Euryarchaeal genomes are folded into SMC-dependent loops and domains, but lack transcription-mediated compartmentalization. Molecular Cell **81**(3), 459–47210 (2021). doi:10.1016/j.molcel.2020.12.013. Accessed 2021-08-23
41. Cazalet, C., Rusniok, C., Brüggemann, H., Zidane, N., Magnier, A., Ma, L., Tichit, M., Jarraud, S., Bouchier, C., Vandenesch, F., Kunst, F., Etienne, J., Glaser, P., Buchrieser, C.: Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity. Nature Genetics **36**(11), 1165–1173 (2004). doi:10.1038/ng1447. Accessed 2021-03-25
42. Baudry, L., Millot, G.A., Thierry, A., Koszul, R., Scolari, V.F.: Serpentine: a flexible 2D binning method for differential Hi-C analysis. Bioinformatics **36**(12), 3645–3651 (2020). doi:10.1093/bioinformatics/btaa249. Accessed 2021-08-24
43. Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A., Nicodemi, M.: Complexity of chromatin folding is captured by the strings and binders switch model. Proceedings of the National Academy of Sciences **109**(40), 16173–16178 (2012). doi:10.1073/pnas.1204799109. Publisher: National Academy of Sciences Section: Biological Sciences. Accessed 2021-08-24
44. 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**(6159), 731–734 (2013). doi:10.1126/science.1242059. Publisher: American Association for the Advancement of Science Section: Report. Accessed 2021-08-10
45. Shumate, A., Salzberg, S.L.: Liftoff: accurate mapping of gene annotations. Bioinformatics **37**(12), 1639–1643 (2021). doi:10.1093/bioinformatics/btaa1016. Accessed 2021-08-24
46. Byers, T.J.: Molecular Biology of DNA inAcanthamoeba, Amoeba, Entamoeba, and Naegleria. In: International Review of Cytology vol. 99, pp. 311–341. Elsevier, ??? (1986). doi:10.1016/S0074-7696(08)61430-8. https://linkinghub.elsevier.com/retrieve/pii/S0074769608614308 Accessed 2021-03-25
47. Gicquaud, C., Tremblay, N.: Observations with Hoechst Staining of Amitosis in *Acanthamoeba castellanii*. The Journal of Protozoology **38**(3), 221–224 (1991). doi:10.1111/j.1550-7408.1991.tb04432.x. Accessed 2021-03-25
48. Yang, Q., Zwick, M.G., Paule, M.R.: Sequence organization of the *Acanthamoeba* rRNA intergenic spacer: identification of transcriptional enhancers. Nucleic Acids Research **22**(22), 4798–4805 (1994). doi:10.1093/nar/22.22.4798. Accessed 2021-03-25
49. Rabl, C.: Uber zelltheilung. Morphol. Jahrbuch (10), 214–330 (1885)
50. Mengue, L., Régnacq, M., Aucher, W., Portier, E., Héchard, Y., Samba-Louaka, A.: Legionella pneumophila prevents proliferation of its natural host Acanthamoeba castellanii. Scientific Reports **6**(1), 36448 (2016). doi:10.1038/srep36448. Accessed 2021-03-25
51. de Jesús-Díaz, D.A., Murphy, C., Sol, A., Dorer, M., Isberg, R.R.: Host Cell S Phase Restricts Legionella pneumophila Intracellular Replication by Destabilizing the Membrane-Bound Replication Compartment. mBio **8**(4), 02345–16 (2017). doi:10.1128/mBio.02345-16. Publisher: American Society for Microbiology. Accessed 2021-09-07
52. Quinet, T., Samba-Louaka, A., Héchard, Y., Van Doninck, K., Van der Henst, C.: Delayed cytokinesis generates multinuclearity and potential advantages in the amoeba Acanthamoeba castellanii Neff strain. Scientific Reports **10**(1), 12109 (2020). doi:10.1038/s41598-020-68694-9. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Evolutionary ecology;Microbial ecology;Parasitology Subject_term_id: evolutionary-ecology;microbial-ecology;parasitology. Accessed 2021-09-07
53. Rothmeier, E., Pfaffinger, G., Hoffmann, C., Harrison, C.F., Grabmayr, H., Repnik, U., Hannemann, M., Wölke, S., Bausch, A., Griffiths, G., Müller-Taubenberger, A., Itzen, A., Hilbi, H.: Activation of Ran GTPase by a Legionella Effector Promotes Microtubule Polymerization, Pathogen Vacuole Motility and Infection. PLOS Pathogens **9**(9), 1003598 (2013). doi:10.1371/journal.ppat.1003598. Publisher: Public Library of Science. Accessed 2021-09-07
54. Escoll, P., Song, O.-R., Viana, F., Steiner, B., Lagache, T., Olivo-Marin, J.-C., Impens, F., Brodin, P., Hilbi,

H., Buchrieser, C.: Legionella pneumophila Modulates Mitochondrial Dynamics to Trigger Metabolic Repurposing of Infected Macrophages. Cell Host & Microbe **22**(3), 302–3167 (2017). doi:10.1016/j.chom.2017.07.020. Accessed 2021-03-25

55. Escoll, P., Mondino, S., Rolando, M., Buchrieser, C.: Targeting of host organelles by pathogenic bacteria: a sophisticated subversion strategy. Nature Reviews Microbiology **14**(1), 5–19 (2016). doi:10.1038/nrmicro.2015.1. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Bacterial host response;Bacterial pathogenesis;Bacterial secretion;Pathogens Subject_term_id: bacterial-host-response;bacterial-pathogenesis;bacterial-secretion;pathogens. Accessed 2021-09-07
56. Escoll, P., Platon, L., Dramé, M., Sahr, T., Schmidt, S., Rusniok, C., Buchrieser, C.: Reverting the mode of action of the mitochondrial FOF1-ATPase by Legionella pneumophila preserves its replication niche. Technical report (May 2021). doi:10.1101/2021.05.12.443790. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. https://www.biorxiv.org/content/10.1101/2021.05.12.443790v1 Accessed 2021-09-07
57. Escoll, P., Buchrieser, C.: Metabolic reprogramming: an innate cellular defence mechanism against intracellular bacteria? Current Opinion in Immunology **60**, 117–123 (2019). doi:10.1016/j.coi.2019.05.009. Accessed 2021-09-07
58. Garcia-Luis, J., Lazar-Stefanita, L., Gutierrez-Escribano, P., Thierry, A., Cournac, A., García, A., González, S., Sánchez, M., Jarmuz, A., Montoya, A., Dore, M., Kramer, H., Karimi, M.M., Antequera, F., Koszul, R., Aragon, L.: FACT mediates cohesin function on chromatin. Nature structural & molecular biology **26**(10), 970–979 (2019). doi:10.1038/s41594-019-0307-x. Accessed 2021-08-23
59. Marbouty, M., Le Gall, A., Cattoni, D., Cournac, A., Koh, A., Fiche, J.-B., Mozziconacci, J., Murray, H., Koszul, R., Nollmann, M.: Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. Molecular Cell **59**(4), 588–602 (2015). doi:10.1016/j.molcel.2015.07.020. Accessed 2021-03-25
60. Anchimiuk, A., Lioy, V.S., Bock, F.P., Minnen, A., Boccard, F., Gruber, S.: A low Smc flux avoids collisions and facilitates chromosome organization in Bacillus subtilis. eLife **10**, 65467 (2021). doi:10.7554/eLife.65467. Publisher: eLife Sciences Publications, Ltd. Accessed 2021-08-23
61. Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., Manke, T.: High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nature Communications **9**(1), 189 (2018). doi:10.1038/s41467-017-02525-w. Number: 1 Publisher: Nature Publishing Group. Accessed 2021-04-28
62. Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M., Alempic, J.-M., Ramus, C., Bruley, C., Labadie, K., Shmakova, L., Rivkina, E., Couté, Y., Abergel, C., Claverie, J.-M.: In-depth study of *Mollivirus sibericum* , a new 30,000-y-old giant virus infecting *Acanthamoeba*. Proceedings of the National Academy of Sciences **112**(38), 5327–5335 (2015). doi:10.1073/pnas.1510795112. Accessed 2021-03-25
63. Harb, O.S., Venkataraman, C., Haack, B.J., Gao, L.-Y., Kwaik, Y.A.: Heterogeneity in the Attachment and Uptake Mechanisms of the Legionnaires' Disease Bacterium, Legionella pneumophila, by Protozoan Hosts. Applied and Environmental Microbiology **64**(1), 126–132 (1998). doi:10.1128/AEM.64.1.126-132.1998. Publisher: American Society for Microbiology. Accessed 2021-10-07
64. Song, L., Florea, L.: Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience **4**(1), 1–8 (2015). doi:10.1186/s13742-015-0089-y. Number: 1 Publisher: BioMed Central. Accessed 2021-08-05
65. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics (Oxford, England) **34**(18), 3094–3100 (2018). doi:10.1093/bioinformatics/bty191
66. Gray, M.W., Burger, G., Derelle, R., Klimeš, V., Leger, M.M., Sarrasin, M., Vlcek, C., Roger, A.J., Elias, M., Lang, B.F.: The draft nuclear genome sequence and predicted mitochondrial proteome of Andalucia godoyi, a protist with the most gene-rich and bacteria-like mitochondrial genome. BMC biology **18**(1), 22 (2020). doi:10.1186/s12915-020-0741-6
67. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) **29**(1), 15–21 (2013). doi:10.1093/bioinformatics/bts635
68. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology **29**(7), 644–652 (2011). doi:10.1038/nbt.1883. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Software;Transcriptomics Subject_term_id: software;transcriptomics. Accessed 2021-10-06
69. Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr, R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., White, O.: Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research **31**(19), 5654–5666 (2003). doi:10.1093/nar/gkg770. Accessed 2021-10-06
70. Gotoh, O.: A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Research **36**(8), 2630–2638 (2008). doi:10.1093/nar/gkn105. Accessed 2021-10-06
71. Stanke, M., Schöffmann, O., Morgenstern, B., Waack, S.: Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics **7**(1), 1–11 (2006). doi:10.1186/1471-2105-7-62. Number: 1 Publisher: BioMed Central. Accessed 2021-10-06
72. Korf, I.: Gene finding in novel genomes. BMC Bioinformatics **5**(1), 1–9 (2004). doi:10.1186/1471-2105-5-59. Number: 1 Publisher: BioMed Central. Accessed 2021-10-06
73. Lomsadze, A., Burns, P.D., Borodovsky, M.: Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Research **42**(15), 119 (2014). doi:10.1093/nar/gku557. Accessed 2021-10-06

74. Testa, A.C., Hane, J.K., Ellwood, S.R., Oliver, R.P.: CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics **16**(1), 1–12 (2015). doi:10.1186/s12864-015-1344-4. Number: 1 Publisher: BioMed Central. Accessed 2021-10-06
75. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., Wortman, J.R.: Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology **9**(1), 1–22 (2008). doi:10.1186/gb-2008-9-1-r7. Number: 1 Publisher: BioMed Central. Accessed 2021-10-06
76. Palmer, J., Stajich, J.: nextgenusfs/funannotate: funannotate v1.5.3. Zenodo (2019). doi:10.5281/zenodo.2604804. https://zenodo.org/record/2604804 Accessed 2021-10-06
77. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R.: InterProScan: protein domains identifier. Nucleic Acids Research **33**(Web Server issue), 116–120 (2005). doi:10.1093/nar/gki442
78. Käll, L., Krogh, A., Sonnhammer, E.L.L.: Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. Nucleic Acids Research **35**(Web Server issue), 429–432 (2007). doi:10.1093/nar/gkm256
79. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., Huerta-Cepas, J.: eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution, 293 (2021). doi:10.1093/molbev/msab293
80. Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T., Ussery, D.W.: RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research **35**(9), 3100–3108 (2007). doi:10.1093/nar/gkm160
81. Tice, A.K., Žihala, D., Pánek, T., Jones, R.E., Salomaki, E.D., Nenarokov, S., Burki, F., Eliáš, M., Eme, L., Roger, A.J., Rokas, A., Shen, X.-X., Strassert, J.F.H., Kolísko, M., Brown, M.W.: PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. PLOS Biology **19**(8), 3001365 (2021). doi:10.1371/journal.pbio.3001365. Publisher: Public Library of Science. Accessed 2021-10-06
82. Alexa, A., Rahnenfuhrer, J.: topGO: Enrichment Analysis for Gene Ontology. Bioconductor version: Release (3.13) (2021). doi:10.18129/B9.bioc.topGO. https://bioconductor.org/packages/topGO/ Accessed 2021-10-06
83. Katoh, K., Standley, D.M.: MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution **30**(4), 772–780 (2013). doi:10.1093/molbev/mst010. Accessed 2021-10-06
84. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J.: Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**(9), 1189–1191 (2009). doi:10.1093/bioinformatics/btp033. Accessed 2021-10-06
85. Matthey-Doret, C., baudrly, axelcournac, Amaury, Remi-Montagne, Guiglielmoni, N., Foutel-Rodier, T., Scolari, V.F.: koszullab/hicstuff: Standardized help messages. Zenodo (2021). doi:10.5281/zenodo.4722873. https://zenodo.org/record/4722873 Accessed 2021-10-06
86. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., Mirny, L.A.: Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods **9**(10), 999–1003 (2012). doi:10.1038/nmeth.2148. Number: 10 Publisher: Nature Publishing Group. Accessed 2021-04-27
87. Matthey-Doret, C.: koszullab/pareidolia: v0.6.1 (2021). doi:10.5281/zenodo.5062485. https://zenodo.org/record/5062485 Accessed 2021-07-02
88. Yang, T., Zhang, F., Yardımcı, G.G., Song, F., Hardison, R.C., Stafford, W., Yue, F., Li, Q.: HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient, 37

Figures

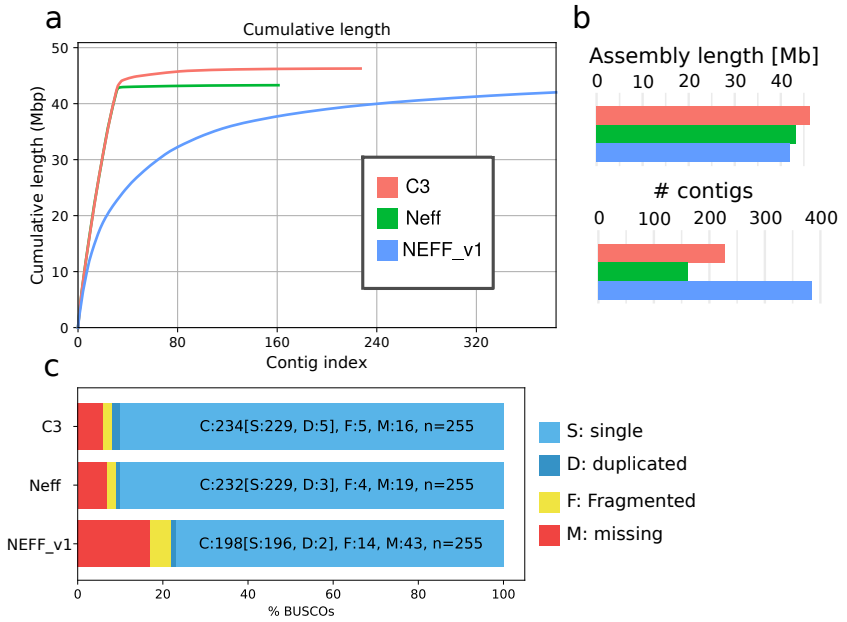


Figure 1 Assembly statistics for A. castellanii genomes. Comparison of genome assemblies for strains C3 and Neff, versus the previous NEFF-v1 genome assembly [25]. **a,** Cumulative length plot showing the relationship between number of contigs (largest to smallest) and length of the assembly. **b,** General continuity metrics. **c,** BUSCO statistics showing the status of universal single copy orthologs in eukaryotes for each assembly.

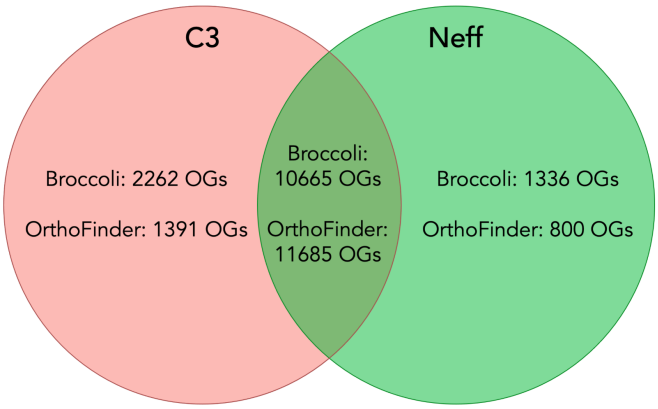


Figure 2 Numbers of strain-specific and shared orthologous groups in the genomes of A. castellanii strains C3 and Neff. Orthology inference was conducted with both Broccoli and OrthoFinder. Dictyostelium discoideum, Physarum polycephalum, and Vermamoeba vermiformis were used as outgroups to improve accuracy of orthogroup inference.

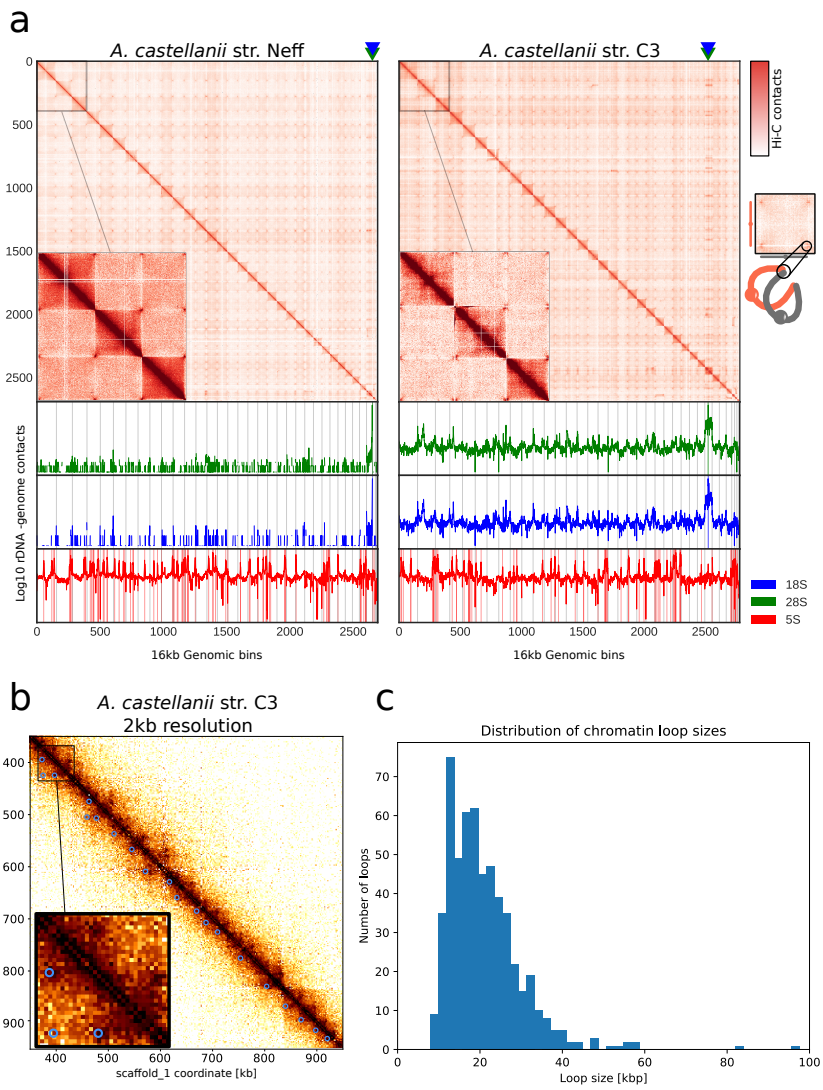


Figure 3 Spatial organisation of the A. castellanii genome. a, Top: Whole genome Hi-C contact maps of the Neff (left) and C3 (right) genomes, with a magnification of the 3 largest scaffolds. The genomes are divided into 16 kbp bins, and each pixel represents the contact intensity between a pair of bins. Each scaffold is visible as a red square on the diagonal. In both strains, there is an enrichment of inter-scaffold contacts towards telomeres, suggesting a spatial clustering of telomeres, as shown on the model in the right margin. Bottom: 4C-like representation of spatial contacts between rDNA and the rest of the genome. Scaffolds are delimited by grey vertical lines. Contacts of all rDNAs are enriched towards telomere. The genomic position of 18S and 28S genes are highlighted with triangles on the top panel and the occurences of 8S rDNA sequences are shown with vertical red lines on the bottom panel. **b,** High resolution the contact map for a region of the C3 genome showing chromatin loops detected by Chromosight as blue circles. **c,** Size distribution of chromatin loops detected in the C3 strain.

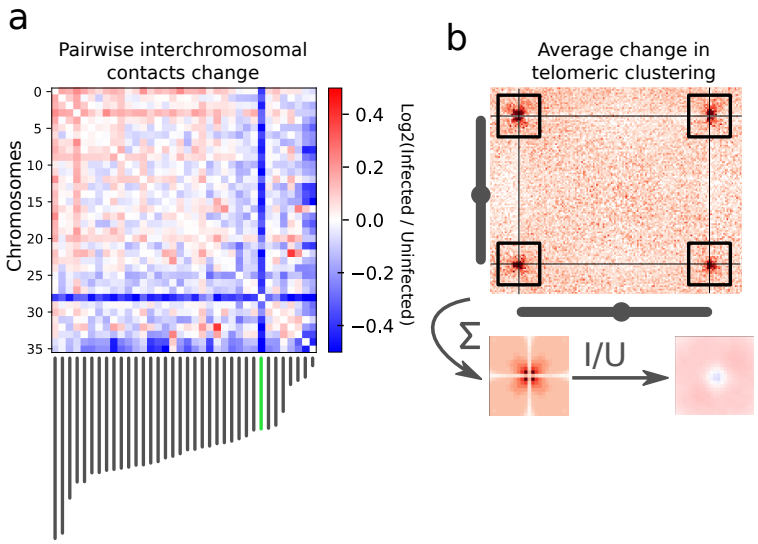


Figure 4 Changes in trans-chromosomal contacts between A. castellanii chromosomes following L. pneumophila infection. a, Average contact change during infection between each pair of chromosomes. Chromosome lengths are shown below the interaction matrix, with the chromosome bearing 18S and 28S rDNA highlighted in green **b,** Representative inter-telomeric contacts between a pair of chromosomes (Neff scaffolds 3 and 11). The average inter-telomeric contact profile generated from all pairs of chromosomes is shown as a pileup. The Log ratio between the infected (I) and uninfected (U) profiles is shown as a ratio (right).

Tables

Assembly	Neff-v1	Neff	C3
Genome size (Mbp)	42.0	43.8	46.1
# scaffolds	384	111	174
# of Ns (Mbp)	2.6 (6.1%)	0 (0%)	0 (0%)
N50 (Mbp)	0.3	1.3	1.4
Largest scaffold (Mbp)	2.0	2.5	2.4
GC%	57.90	58.44	58.64
# protein coding genes	14,974	15,497	16,837

Table 1 Genome statistics for the finished assemblies of Neff, C3 (this study) and the reference Neff-v1 genome.

Strain	Identity	Gaps
Neff	757/826 (91.6%)	1/826 (0.12%)
C3	821/825 (99.5%)	0
A. polyphaga	802/825 (97.5%)	0

Table 2 Identity of mannose binding proteins from *A. polyphaga* and *A. castellanii* strains Neff and C3 to their homolog in *A. castellanii* strain MEEI 0184 across 788 sites of a 834-site amino acid alignment. The first 46 sites of the alignment were excluded from the calculation because the 5' end of the gene in *A. polyphaga* was missing due to a truncated contig.

Supplementary figures

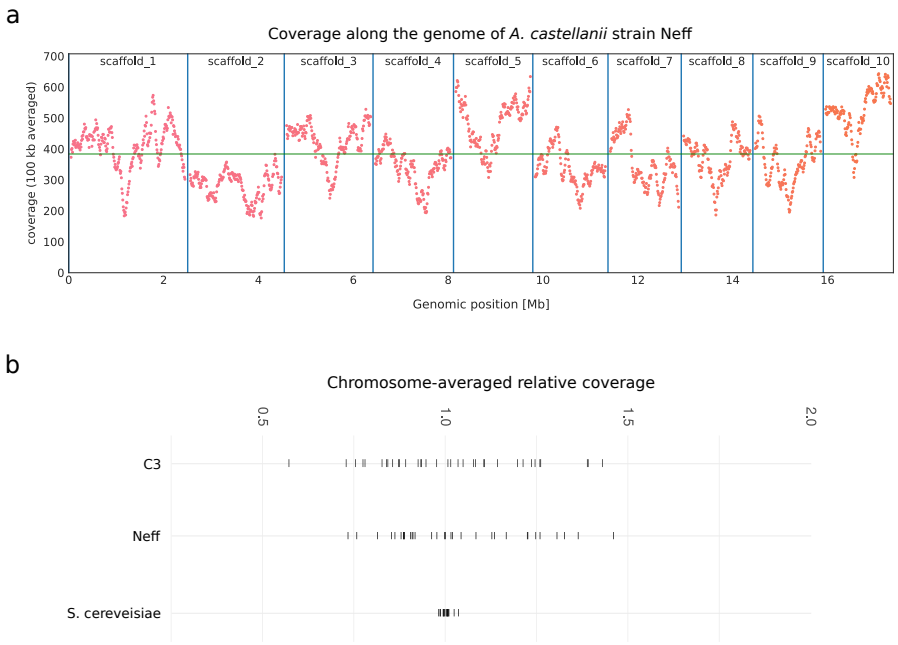


Figure S1 Coverage across scaffolds of A. castellanii compared to a known haploid. a, Illumina short-reads coverage along the 10 largest scaffolds of *A. castellanii* Neff in a 100 kbp sliding window, with the horizontal green line showing genome median coverage. **b,** Variability of median coverage per chromosome (relative to genome median) for *A. castellanii* strains C3 and Neff, and asynchronous *Saccharomyces cerevisiae* strain BY4741, a known haploid. For *S. cerevisiae*, library SRR1569870 was used.

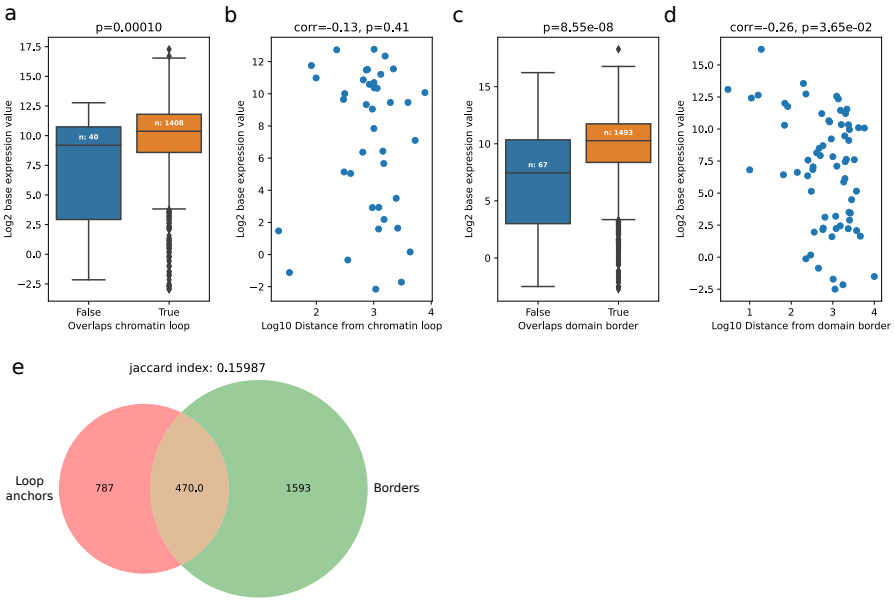


Figure S2 Gene expression according to position relative to chromatin structures. Expression of the closest gene to each loop anchors versus **a,** overlap status with chromatin loops and **b,** distance to closest loop. Expression of the closest gene to domain borders versus **c,** overlap status with domain borders and **d,** distance to closest border. P-values reported for overlap comparisons are obtained using Mann-Whitney U test, correlation coefficients and associated p-values are computed using Spearman's correlation test. **e,** Overlap between chromatin loop anchors and domain borders represented as a Venn diagram.

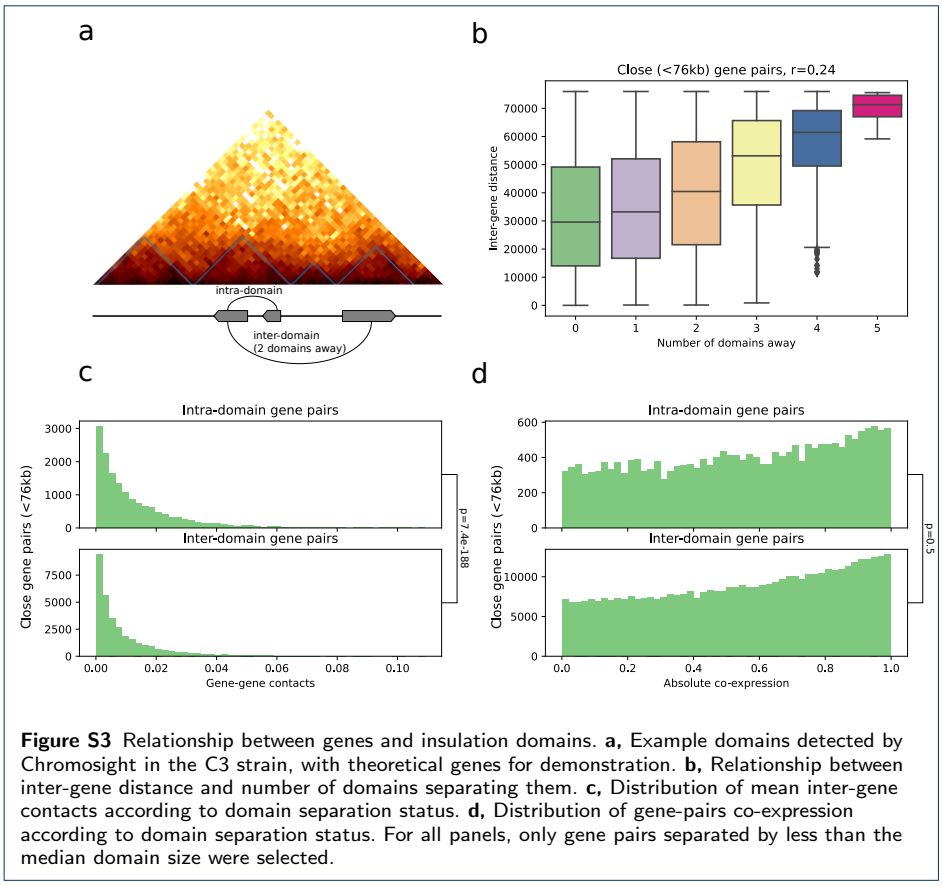


Figure S3 Relationship between genes and insulation domains. **a,** Example domains detected by Chromosight in the C3 strain, with theoretical genes for demonstration. **b,** Relationship between inter-gene distance and number of domains separating them. **c,** Distribution of mean inter-gene contacts according to domain separation status. **d,** Distribution of gene-pairs co-expression according to domain separation status. For all panels, only gene pairs separated by less than the median domain size were selected.

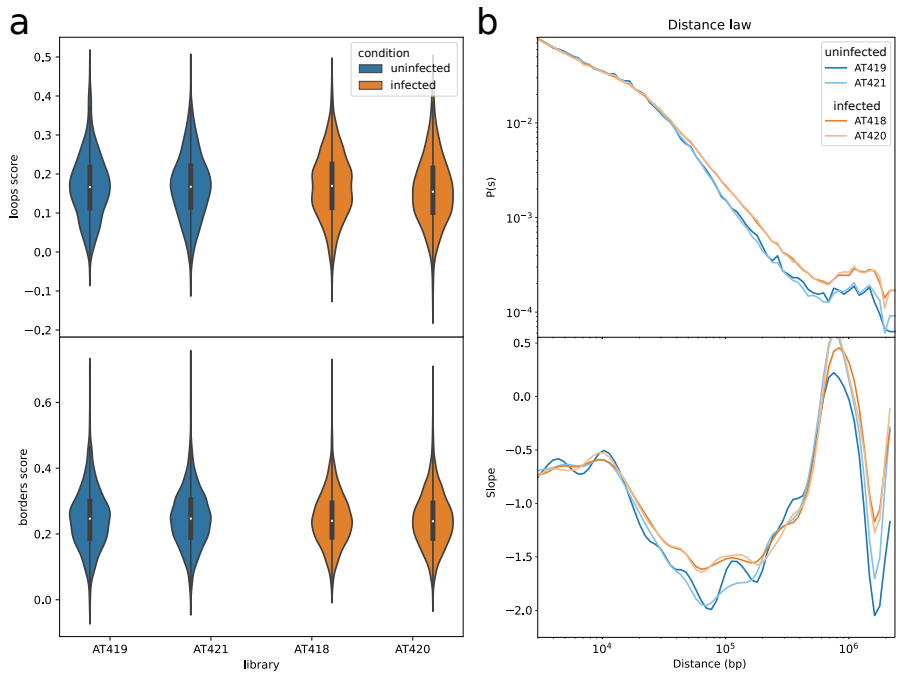


Figure S4 Global comparisons of infection Hi-C results between replicates. **a,** Distribution of Chromosight loops and borders scores for all 4 samples. **b,** Distance-contact decay function (denoted P(s)) and its slope.

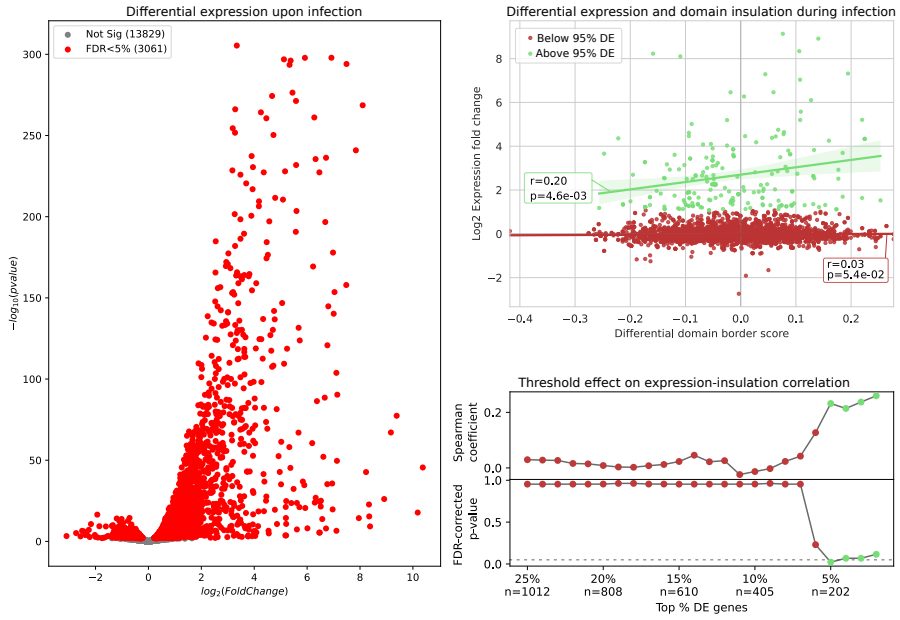


Figure S5 Relationship between differential expression and domain insulation during infection. **a,** Volcano plot showing differential gene expression (DE) of infected (5h p.i.) versus uninfected amoeba. Genes with significant corrected p-values (FDR¡5%) are shown in red. **b,** Changes in gene expression and insulation strength of closest domain border during infection. Linear regression lines, Spearman correlation coefficients and associated p-values are shown separately for genes with extreme fold change values (95% quantile) and the rest. **c,** Spearman correlation coefficient between expression fold change and domain insulation change, and associated FDR-corrected p-values (FDR¡5%) for different subsets of genes according to the threshold of extreme fold change. Values are colored according to the 95% threshold selected in b.

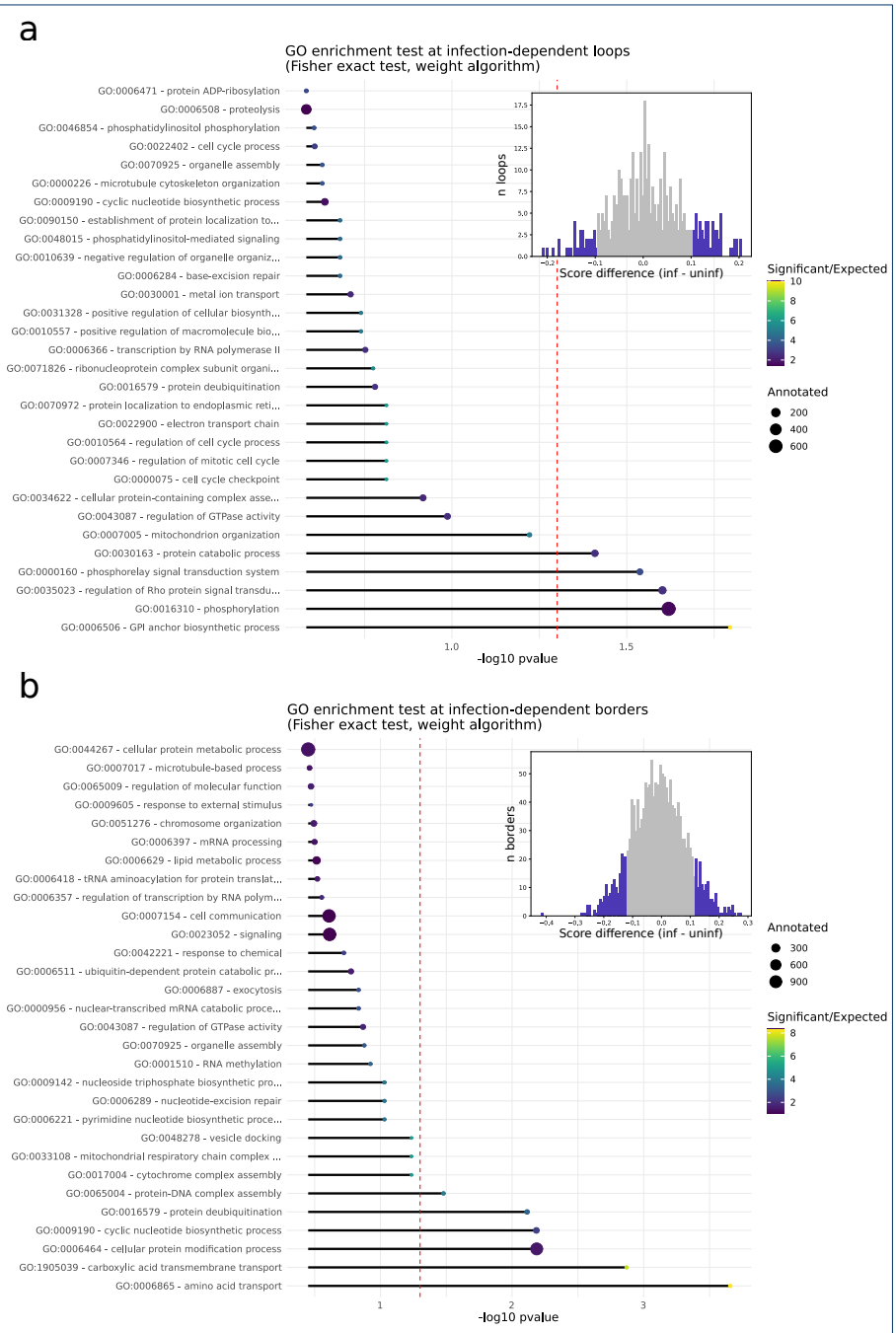


Figure S6 GO term enrichment test results for genes overlapping infection-dependent **a,** chromatin loops and **b,** domain borders. Histograms show the distribution of loop and border score changes during infection, with highlighted portions showing the 80% percentile threshold used to include genes in the GO enrichment test.

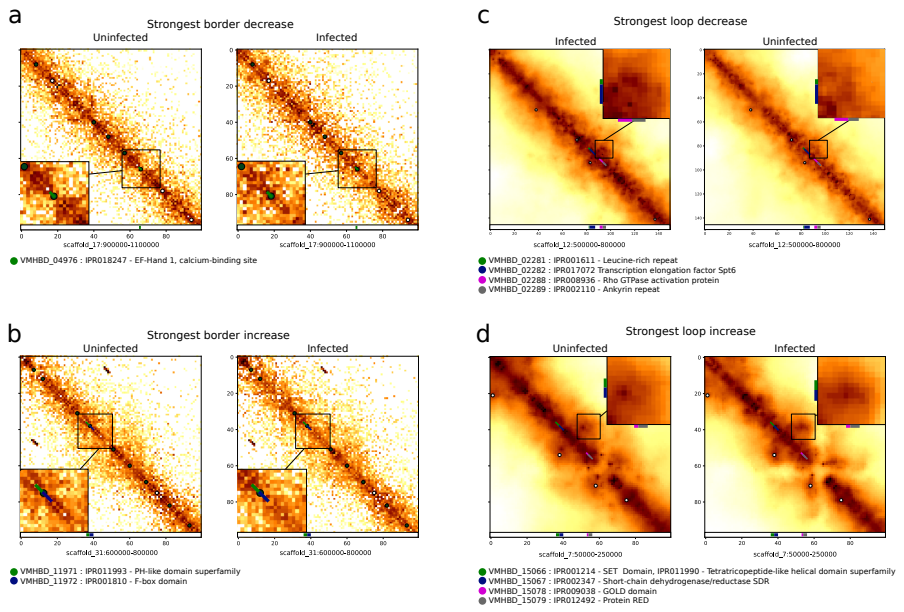


Figure S7 Hi-C zooms on strongest pattern changes during infection. Description of the closest genes are shown below each zoom. Balanced contact map zooms showing **a,** strongest border decrease and **b,** decrease. Serpentine-binned contact maps showing **c,** strongest loop decrease and **d,** increase.

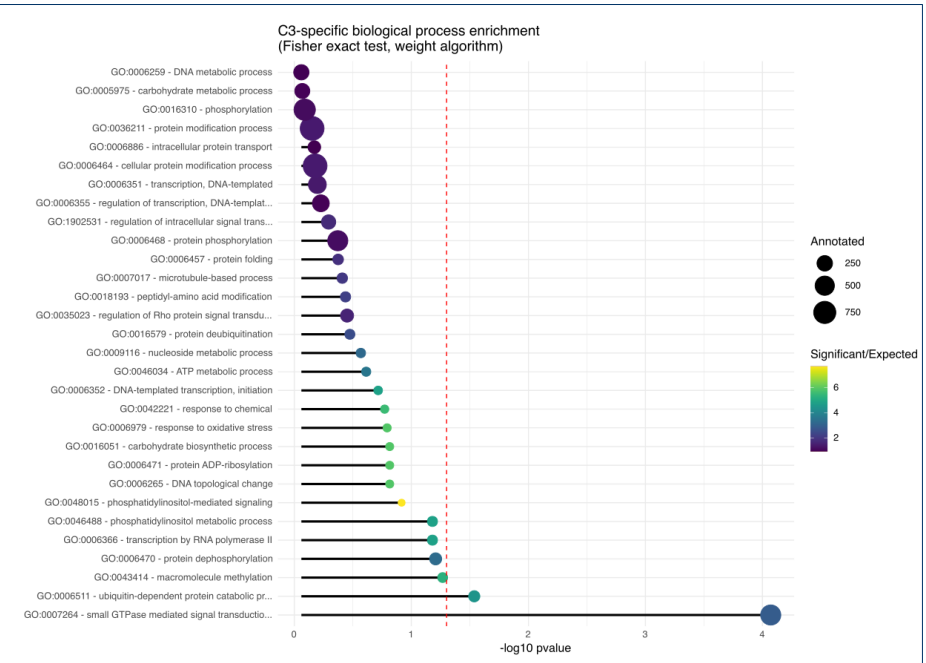


Figure S8 Most significant biological process GO term enrichments in genes specific to Acanthamoeba castellanii strain C3. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

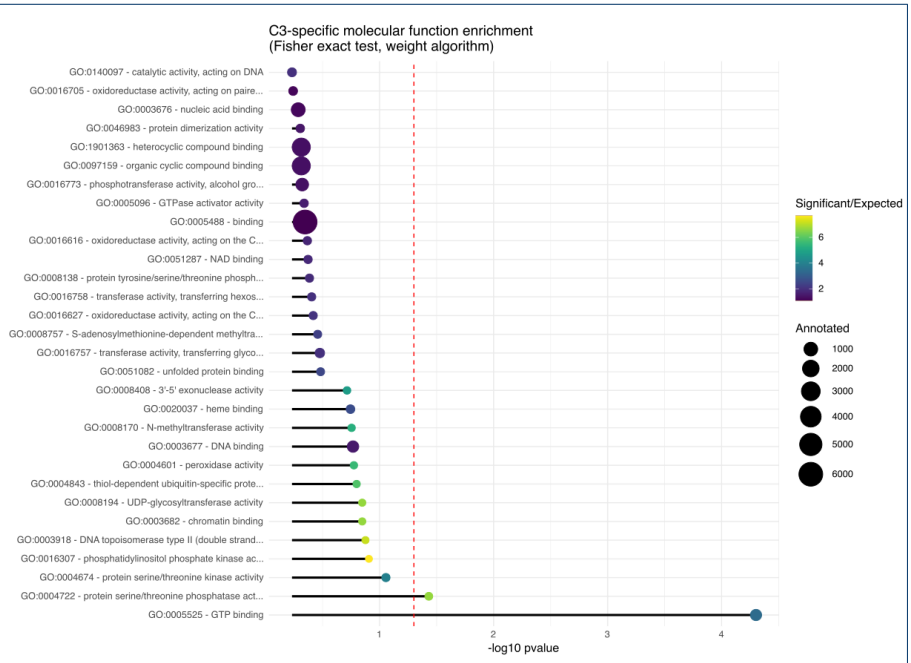


Figure S9 Most significant molecular function GO term enrichments in genes specific to Acanthamoeba castellanii strain C3. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

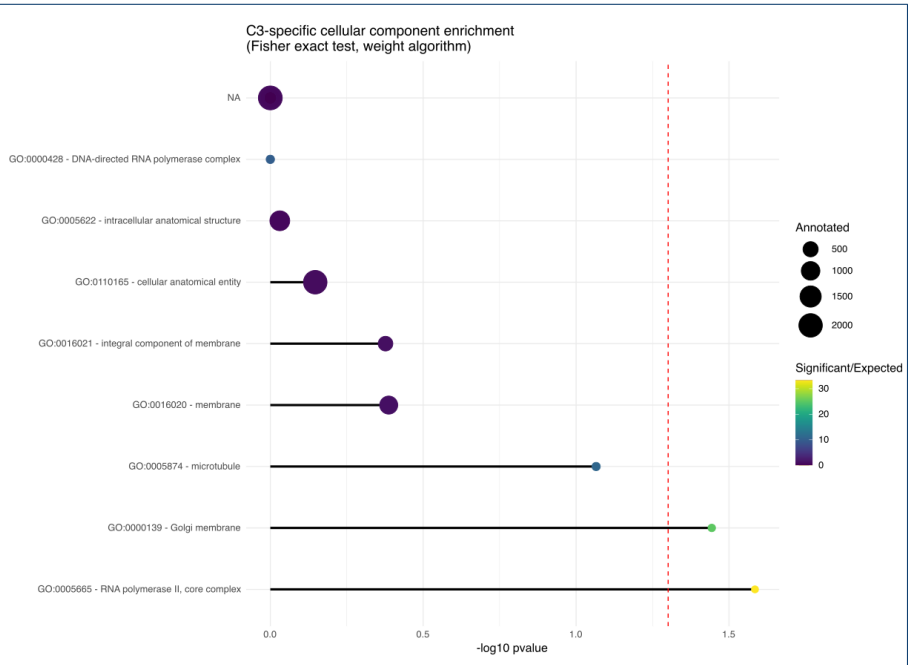


Figure S10 Most significant cellular component GO term enrichments in genes specific to Acanthamoeba castellanii strain C3. Enrichment was determined using topGO, with nodeSize set to 5 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

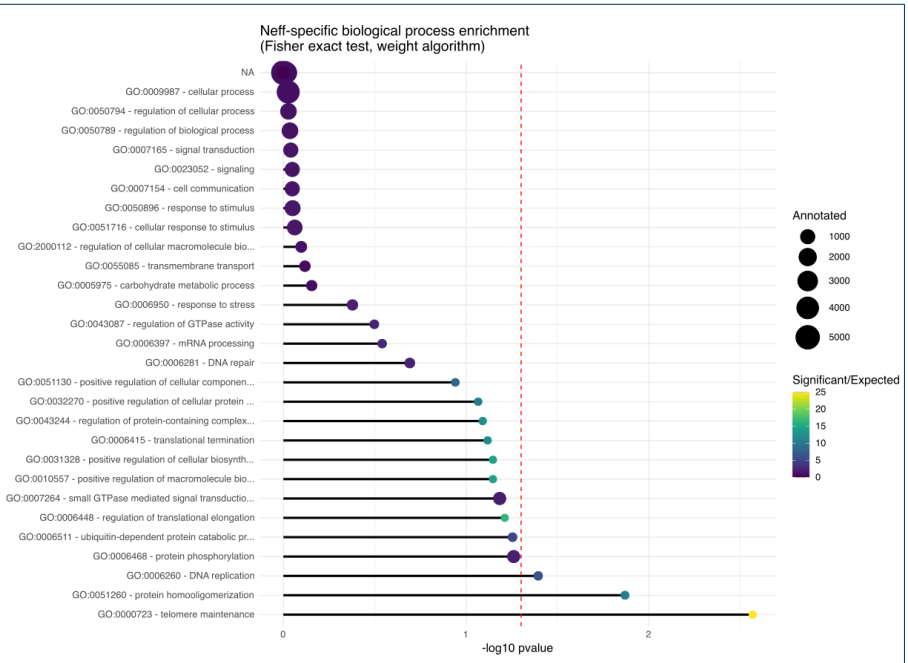


Figure S11 Most significant biological process GO term enrichments in genes specific to *Acanthamoeba castellanii* strain Neff. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

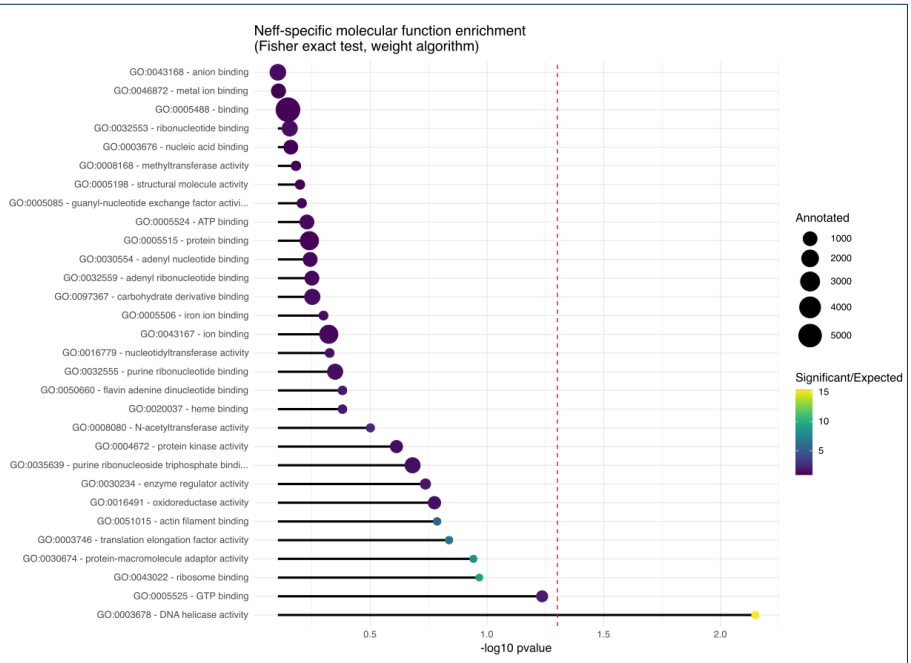


Figure S12 Most significant molecular function GO term enrichments in genes specific to *Acanthamoeba castellanii* strain Neff. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

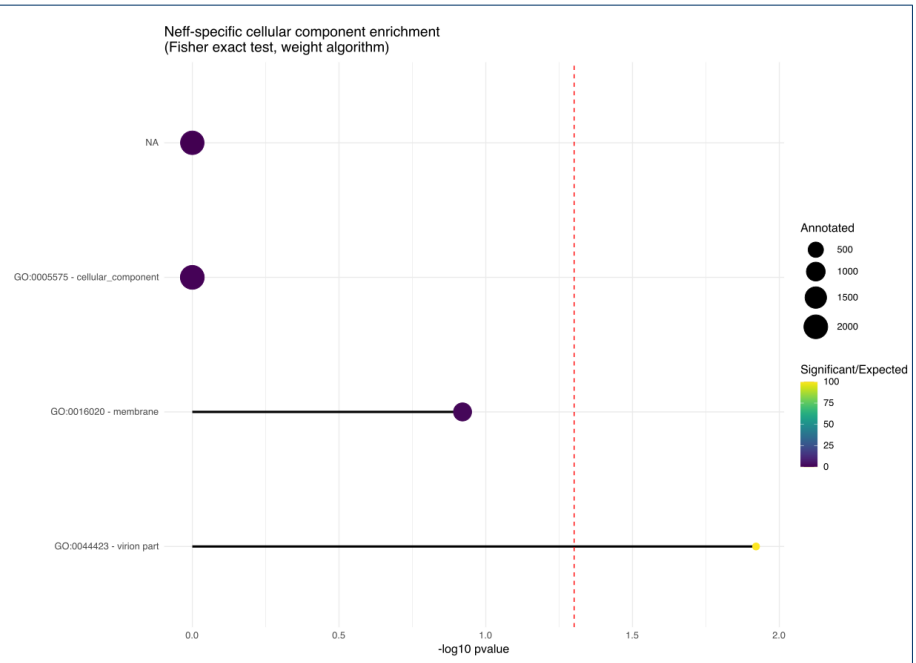


Figure S13 Most significant cellular component GO term enrichments in genes specific to Acanthamoeba castellanii strain Neff. Enrichment was determined using topGO, with nodeSize set to 5 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

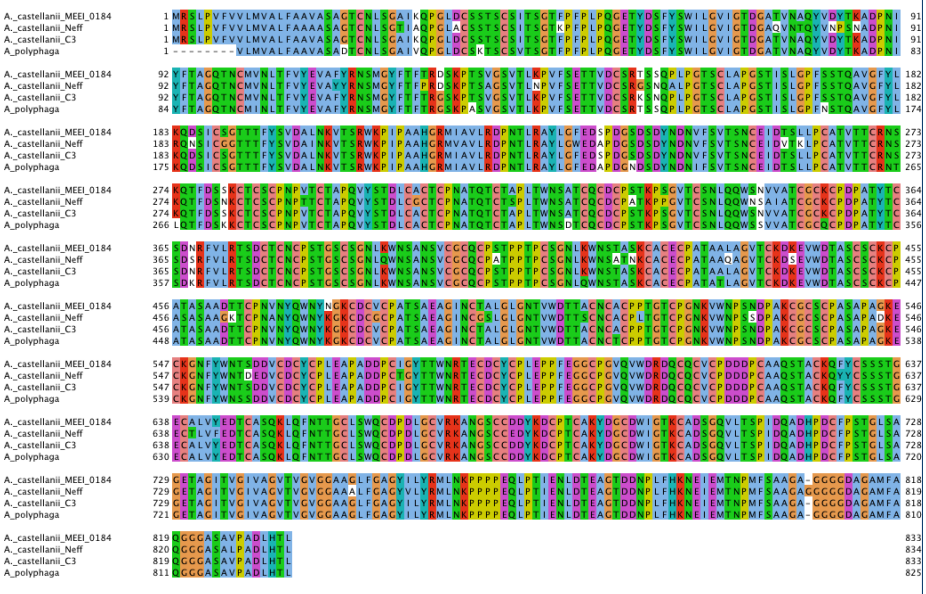


Figure S14 Multiple sequence alignment of mannose binding protein orthologs across three strains of Acanthamoeba castellanii and one strain of Acanthamoeba polyphaga. Sites are coloured according to the Clustalx colour scheme and residues differing from the consensus at any given site are not coloured. The alignment was generated with MAFFT-linsi, and was viewed and coloured in Jalview.

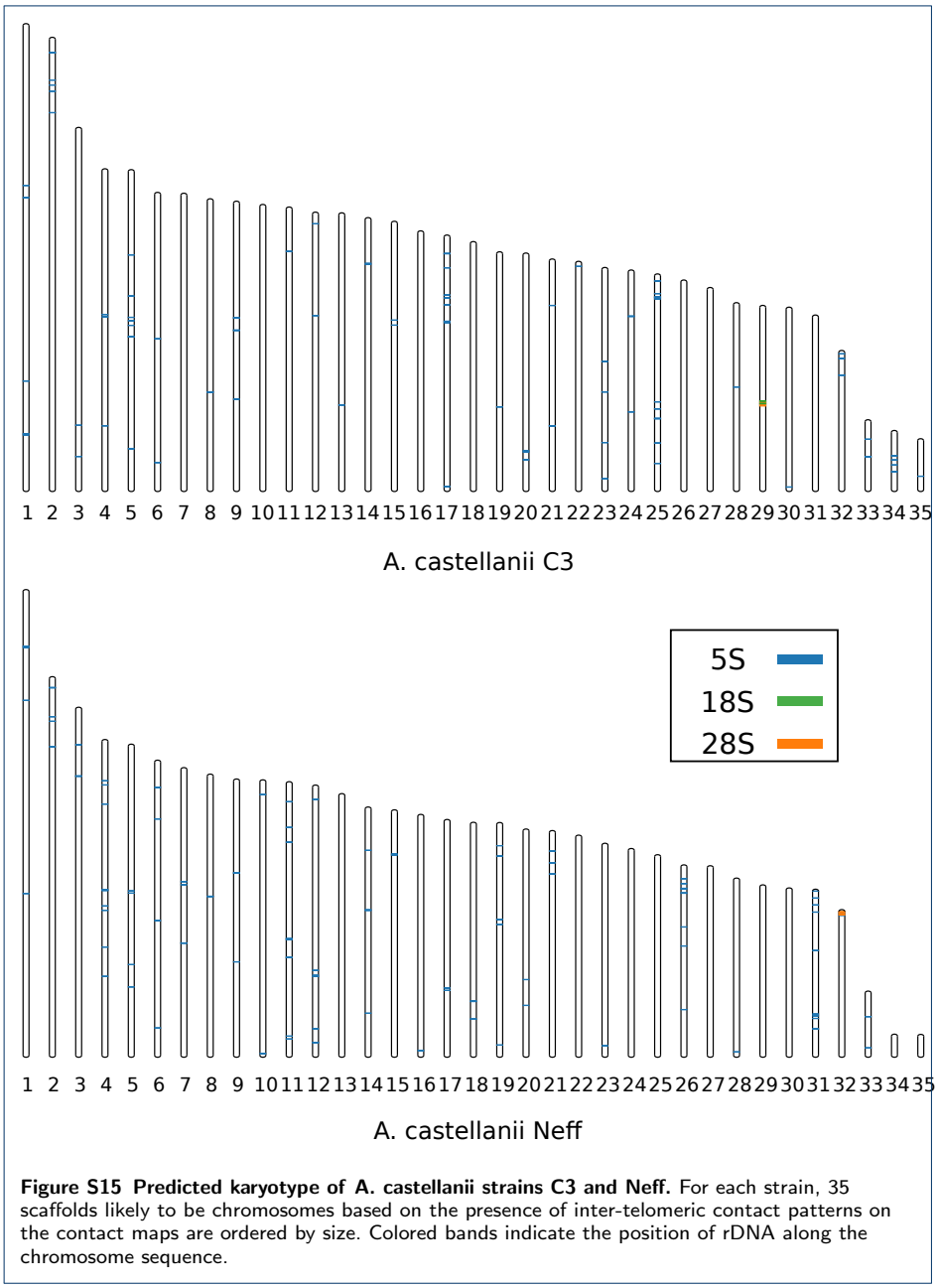


Figure S15 Predicted karyotype of A. castellanii strains C3 and Neff. For each strain, 35 scaffolds likely to be chromosomes based on the presence of inter-telomeric contact patterns on the contact maps are ordered by size. Colored bands indicate the position of rDNA along the chromosome sequence.

2.2 Inter-strain sequence divergence

To put into perspective the divergence between *A. castellanii* strains Neff and C3, here we compare them to 11 other amoeba species whose genomes are available. As described in the previous section, we extracted all predicted coding sequences from each species and used orthofinder (v2.3.3) to constitute groups of orthologous genes among these proteomes.

We constructed a phylogenetic tree using orthofinder's built-in implementation of the STAG (Species Tree inference from All Genes) procedure [174]. Briefly, Orthofinder builds a tree for each gene based on multiple sequence alignment, and then uses all gene trees where all species are available to build a species tree. The species tree is then rooted by Orthofinder using the STRIDE (Species Tree Root Inference from gene Duplication Events) procedure [175]. In the resulting tree, C3 and Neff are highly divergent from the closest available relative, *Planoprotostelium fungivorum* (Fig. II.Ia). Both strains exhibit a high (protein) sequence divergence, with about 0.084 amino-acid substitution per site. This is comparable to inter-species divergence in some groups, for example 0.061 between *Entamoeba histolytica* and *Entamoeba dispar*. It is important to note, however, that branch length and topology may be inaccurate for distantly related groups. This could be due to numerous factors, including low quality of assemblies used to infer proteomes, contaminations and and horizontal gene transfers [176]. When computing the DNA sequence divergence across aligned genomic segments between the two strains, we observe an average of 6.65% nucleotide substitutions (Fig. II.J). In comparison, computing this same metric between Neff and *Dictyostelium discoideum* yields 11.8% substitutions, however this could be an underestimate, since focusing on aligned segments will likely select for regions with high conservation.

As mentioned earlier, over 10% of *A. castellanii* orthogroups are strain-specific (Fig. II.Ib), further emphasizing their strong differences. There also seems the be a few major genomic rearrangements between the two strains, however it is hard to assess whether these are genuine inter-strain differences. These segments could also be the product of mis-assemblies induced by the collapse of heterozygous structural variants in either strains.

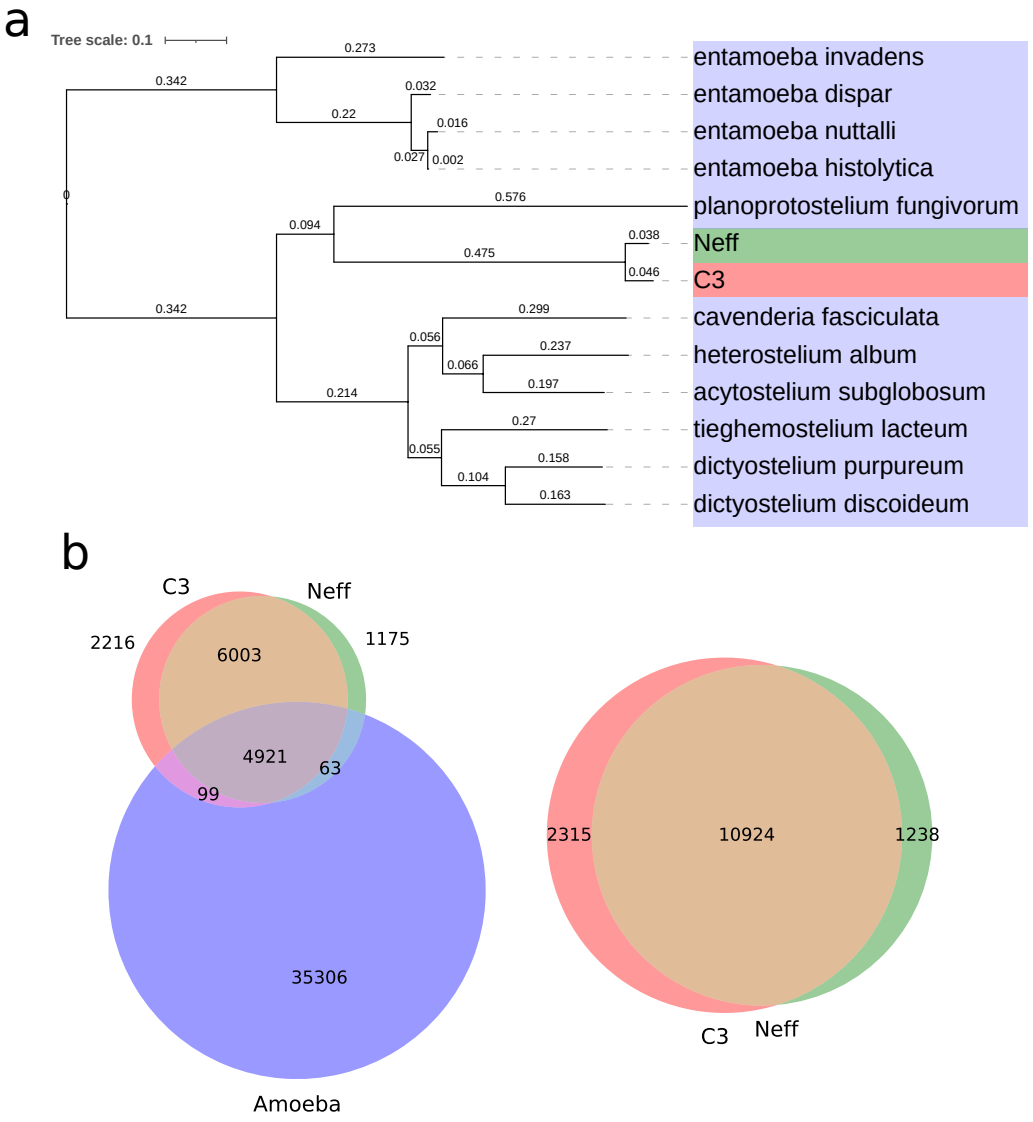


Fig. II.I: Comparison of *A. castellanii* strains C3 and Neff with other species. **a:** Phylogenetic tree of *A. castellanii* strains C3 and Neff and 11 other amoeba species built based on all coding sequences. Distances represent substitution per position. **b:** Comparison of orthogroups content between *A. castellanii* and the 11 other amoeba species used in the tree.

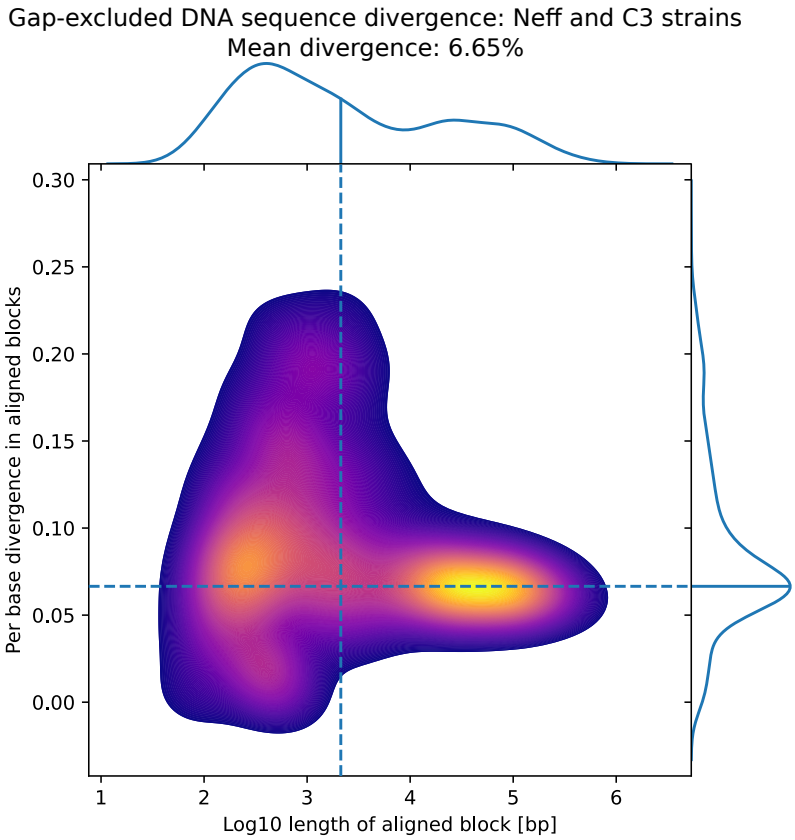


Fig. II.J: Density plot showing the distribution of sequence (gap-excluded) divergence across aligned blocks between *A. castellanii* strains C3 and Neff.

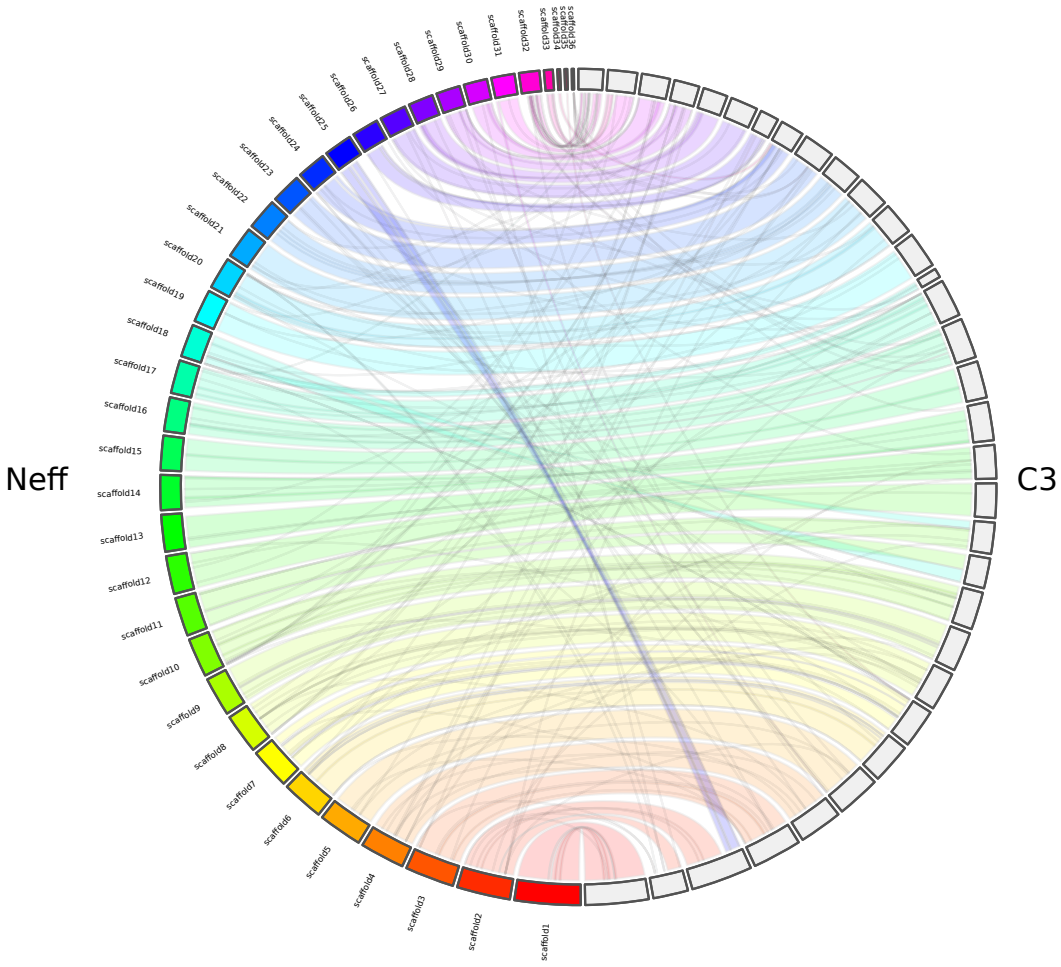


Fig. II.K: Circos plot showing homologous blocks for all scaffolds of *A. castellanii* strains C3 and Neff longer than 50 kb.

3 Infection of murine macrophages by *Salmonella*

In the former chapter, we presented the investigation of the infection process of a natural host by a bacterial pathogen. In this chapter, we investigate the opportunistic infection of mammalian macrophages by *Salmonella enterica* through the prism of genome folding. Notably whether, and how *S. enterica* infection affects the chromatin of its eukaryotic host. Much like *Legionella*, *Salmonella* manipulates its host cell's defense and signalling to promote its own survival in their cytoplasm [41]. Using a mouse Bone Marrow Macrophage (BMM) model, we measured changes in chromatin architecture, accessibility and gene expression in different infection conditions and timepoints to explore the potential epigenetic deregulations happening during this process.

In the case of *A. castellanii*, infection involved an unicellular host. Here, infection takes place in mammalian cells, whose much larger genomes have an intertwined, complex tridimensional organization. Notably, they are segmented into active and inactive compartments, and they contain long-range regulatory elements organized into chromatin domains and exhibiting cell type specificity [177].

Here we focus on bone marrow-derived macrophages. These cells originate from hematopoietic stem cells and go through a complex differentiation process (Fig. II.La). After differentiation, they retain a strong plasticity and can be activated by cues in their environment to become "polarized" into one of two main activation states (Fig. II.Lb). M1 macrophages secrete high amounts of cytokines and promote inflammation, while M2 macrophages suppress immune response and focus on tissue repair [178]. Together with neutrophils, M1 Macrophages are first responders to infection and act as key modulators and effectors cells during the immune response. During a bacterial infection, they will indeed phagocytose bacterial cells and initiate adaptive immunity by activating T cells through antigen presentation via the Major Histocompatibility Complex II (MHC-II). Additionally, they release cytokines and chemokines, which promote inflammation and further recruitment of other immune cells, and secrete anti-microbial molecules to destroy infectious cells. However, in the case of a prolonged infection, this strong immune reaction can have detrimental outcomes to the host and result in organ damages, or even lethal shock [179]. This overstimulation is avoided by a process known as endotoxin- or LPS-tolerance. This state of reduced immune response is triggered by continuous exposure to lipopolysaccharides exposed on the bacterial cell surface and skews the macrophage

population towards M2 polarization [180]. LPS-tolerance must also be tightly balanced, as suppressed immunity can lead to secondary infection or even sepsis. Such regulation is known to involve a combination of signalling and gene-specific chromatin changes through histone modifications [181].

Throughout the next sections, we describe an ongoing investigation, done in collaboration with the laboratory of Sophie Helaine at Harvard Medical School and her postdoc Peter W. Hill currently at Imperial College, of changes in chromosome conformation in macrophages following infection by *Salmonella*. As we will focus on changes happening during late infection, LPS-tolerance is especially relevant to the understanding of this chapter.

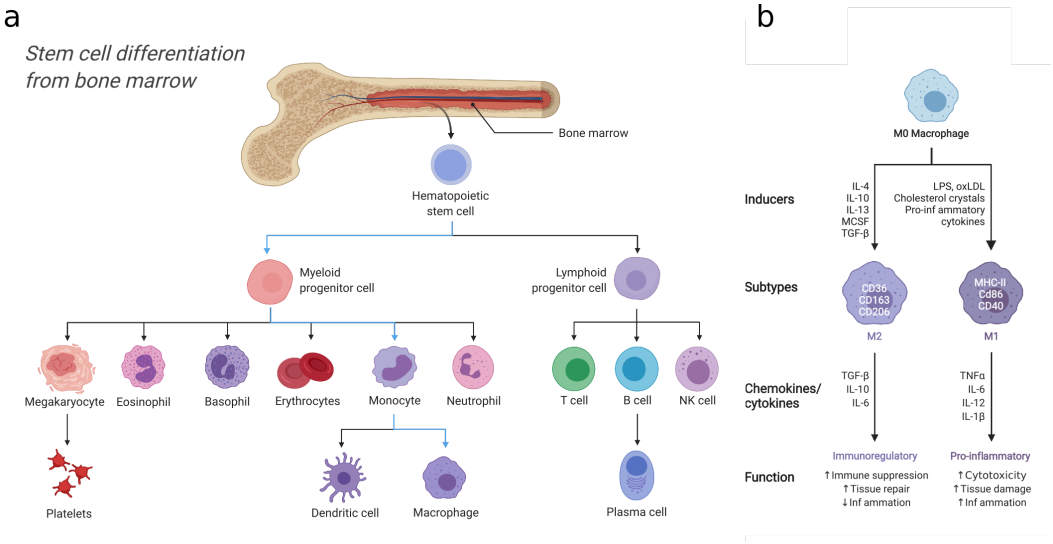


Fig. II.L: Macrophage differentiation from hematopoietic stem cells. **a:** Cellular differentiation pathway leading from bone marrow stem cells to macrophages **b:** Macrophage polarization from M0 to M1 (pro-inflammatory) or M2 (anti-inflammatory) macrophages. For either forms, molecules associated with induction of polarization, surface exposure and secretion are listed, as well as its functions. Adapted from "Stem cell differentiation from bone marrow" and "Macrophage subtypes in atherosclerosis", by BioRender.com (2020). Retrieved from https://app.biorender.com/biorender-templates

ALTERATION OF CHROMATIN STRUCTURE DURING LPS TOLERANCE ACQUISITION IN *Salmonella enterica*-INFECTED MACROPHAGES

A PREPRINT

Cyril Matthey-Doret[1,2], Peter W. Hill[3], Agnès Thierry[1], Sophie Helaine[4,*], Romain Koszul[1,*]

October 28, 2021

1 Institut Pasteur, Spatial Regulation of Genomes unit, CNRS, UMR 3525, C3BI USR 3756, Paris, France
2 Sorbonne Université, Collège Doctoral, F-75005 Paris, France
3 Department of Medicine, MRC CMBI, Imperial College London, London, UK
4 Department of Microbiology, Harvard Medical School, 77 Ave Pasteur, Boston, MA 02115, USAHMS Department of Microbiology, Harvard, US

ABSTRACT

In vertebrates, the immune response to bacterial infection involves a complex balance to clear infectious agents without damaging the tissues. During prolonged infections, LPS tolerance is key to this balance, causing a temporary reduction in the inflammatory response to regenerate and preserve tissues. Many regulatory layers are important to coordinate infectious response, including phosphorylation cascades, histone modifications, and micro-RNAs. Gene expression and epigenetic changes are often intertwined with spatial organization of the genome. In this work, we study changes in chromatin structure during infection of murine bone macrophages by *Salmonella enterica*. We find global changes during late infection, when LPS tolerance presumably take place, and identify several pathways associated with chromatin changes.

***K*eywords** Genomics · Hi-C · Host-parasite · Infection

Introduction

Salmonella is an intracellular bacterium and human pathogen causing an enteric disease known as salmonellosis in many animals. It is usually contracted by ingestion of contaminated food or water, and infiltrates the intestinal epithelium. Salmonella destabilizes the tight junctions inbetween epithelial cells, favoring the migration of neutrophils through the epithelial layer by stimulating the mitogen-activated protein kinase (MAPK) and NF-κB pathways [1, 2]. The sensing of bacterial lipopolysaccharides (LPS) present on the bacterial cell surface by immune cells through Toll-like receptors (TLR) elicits the production of interleukins and activation of caspase genes [3], which further increases intestinal inflammation.

Salmonella cellular infection is mediated by two type 3 secretion systems (T3SS), encoded by distinct pathogenicity islands, *Salmonella pathogenicity island* (SPI) 1 and 2. Both T3SS mediate the transfer of bacterial proteins (or effectors) into the host-cell cytoplasm, but they are active at different time during infection. The SPI1 T3SS is used mainly for invasion of non-phagocytic cells and induction of inflammatory response, whereas the SPI2 T3SS is important for bacterial survival in macrophages and establishment of systemic disease [4].

Salmonella can infect many cell types in the epithelium by T3SS SPI1-mediated endocytosis, and enter macrophages through phagocytosis [5]. During systemic infection, macrophages phagocytose *Salmonella* at mesentric lymph nodes and transport the bacteria to other sites such as the spleen, liver and bone marrow [6]. The bacterium can survive for

long periods in these macrophages and form ganulomas. Upon cell entry, Salmonella secretes effector proteins through its T3SS to manipulate the host defenses and metabolism, and replicate inside the cell. A combination of replicating and non-replicating bacteria can co-exist within the host cell [7]. Non-replicating, dormant cells - also called persisters - display an increased antibiotic recalcitrance and are of particular concern for the relapse of *Salmonella* infection following stoppage of antibiotic treatment [8].

In response to bacterial infection, host macrophages secrete chemokines and cytokines to recruit other immune cells to the infection site. They also produce reactive oxygen species (ROS) and microbicidal molecules to kill surrounding bacterial cells [9, 10]. This response is stimulated by LPS present on the cell surface of gram-negative bacteria such as *Salmonella*. However, in cases of intense and prolonged exposure, this can lead to over-stimulation of the immune system, sometimes resulting in an endotoxic shock which poses the threat of tissue damage, organ failure and death. To avoid such outcome, the body can enter a transient state of hyporesponsiveness to infection known as endotoxin-tolerance, or LPS-tolerance. During this period, macrophages are reprogrammed to cease production of inflammatory molecules and instead focus on tasks such as tissue repair and phagocytosis of cellular debris [11].

A complex interplay takes place between host inflammatory factors and bacterial effectors. Upon invasion of the intestinal lumen by *Salmonella*, the release of ROS by macrophages leads to a growth advantage for the pathogen over resident bacteria from the microbiome [12]. Conversely, after cellular entry, *Salmonella* effectors dampen inflammation to favor intracellular survival, reducing IL-8 secretion and MAPK-mediated inflammation using its effector proteins [13]. This suggests that *Salmonella* can increase or decrease inflammatory response depending on the stage of infection. Understanding how bacteria manipulate the host immune response is an important step towards treating and mitigating risks associated with *Salmonella* infection. Many levels or regulation are affected by the bacterium, including signal transduction pathways [14], mitochondrial metabolism [15], RNA splicing [16] and histone marks [17].

In mammals, gene regulation is intertwined with genome compaction and folding. At the broadest level, chromatin is segregated into active and inactive compartments, which can change according to the needs of the cell [18]. The genome is also partitioned into Topologically Associating Domains (TADs) which insulate genes and regulatory elements from each others [19, 20]. Within TADs, chromatin loops mediated by the structural maintenance complex cohesin (SMC, [21]) can modulate the folding of chromatin. Cohesins mediate the expansion of small loops through an active process called loop extrusion [22]. As a consequence, it has been proposed that these loops could mediate regulatory interactions by bringing physically closer together promoters and regulatory elements such as enhancers. In mammals, the boundaries of these chromatin structures are mainly formed by the transcription CCTC-binding factor (CTCF) [23]. CTCF-bound positions have been shown to delimit the anchors of loop basis. CTCF-mediated loops have been proposed to play roles in immunity, for instance by increasing the expression of genes in the major histocompatibility complex (MHC) locus [24, 25, 26], or coordinating the expression of interleukins [27, 28]. The LPS-tolerance phenomenon is thought to be regulated by epigenetic mechanisms such as histone modifications [29, 30, 31], but thus far there has been little investigation of the implication of genome conformation in that process.

Here we investigate the consequences of *Salmonella* infection on the genome structure of mammalian host cells. Using Hi-C in Mouse bone marrow macrophages (BMM), we describe the spatial genomic features at early and late Salmonella infection, and how they relate to gene deregulation. We find genome-wide changes in chromatin compartments and overall organization during late infection (around 20h post infection). This coincides with the time at which LPS tolerance is acquired [32]. We find large compartment switches associated with the MHC complex and chemokine genes. We also identify strong changes in chromatin loops, compartment and expression associated with chemokine genes, known to regulate cell migration and chemotaxis. Finally, we observe changes in expression and long range interactions during infection for several markers of LPS tolerance, as well as the anti-inflammatory cytokine Interleukin-10.

Results

Chromosome folding is altered in late infection

We used Hi-C to capture the chromosome conformation of both murine BMM cells infected by *Salmonella* as well as bystander cells exposed to *Salmonella* (Methods). Hi-C was performed in uninfected cells, and at two time points representing early (2h) and late (20h) infection. BMM cells infected by a *Salmonella* ΔSsaV mutant strain deficient for the T3SS SPI2 and unable to inject effector proteins into the host cytoplasm, were also processed by Hi-C. We used 3 different Hi-C derived features to measure chromosome structural changes. First, the stratum-adjusted correlation coefficient [33], which measures the overall contact similarity between Hi-C matrices of sample pairs (Fig. 1a). Second, the slope of the distance-contacts decay function (Fig. 1b), which reflects chromatin compaction averaged over the genome. Finally, the A/B compartment eigenvectors (Fig. 1c), which encode the segmentation of the genome into active and inactive chromatin. The strongest changes took place during late (20h) infection, regardless of *Salmonella*

genotype or bystander versus infected status. Infection time point (20h vs 2h) was the main determinant with respect to all 3 aforementioned features.

The time at which we observed the strongest conformational changes coincides with the time range for the acquisition of LPS tolerance (16 - 48h) [34]. To focus on such changes, we re-sequenced Hi-C libraries from samples infected by WT Salmonella at those time points (each time point in duplicates). This allowed us to inspect changes in fine grained chromatin structures, such as chromatin loops and TAD borders.

We used Hi-C to measure compartment changes at two time points in infected cells. Genome-wide A/B compartmentalization was more pronounced at late infection (Fig. 1d) compared to early infection or uninfected cells. These large scale changes could be attributed to physiological changes in late infection.

Using RNA-seq we generated from the same samples at 2h and 20h post infection (Methods), we investigated the expression of known LPS-response marker genes [32]. We found that the negative regulators of LPS response were upregulated in late infection (Fig. S1a), with some of these changes occurring concomitantly to changes in chromatin loops patterns in the Hi-C data (Fig. S1b). The groups of previously reported positive and negative regulators of LPS-tolerance were largely consistent with our differential expression results (Fig. S1c)

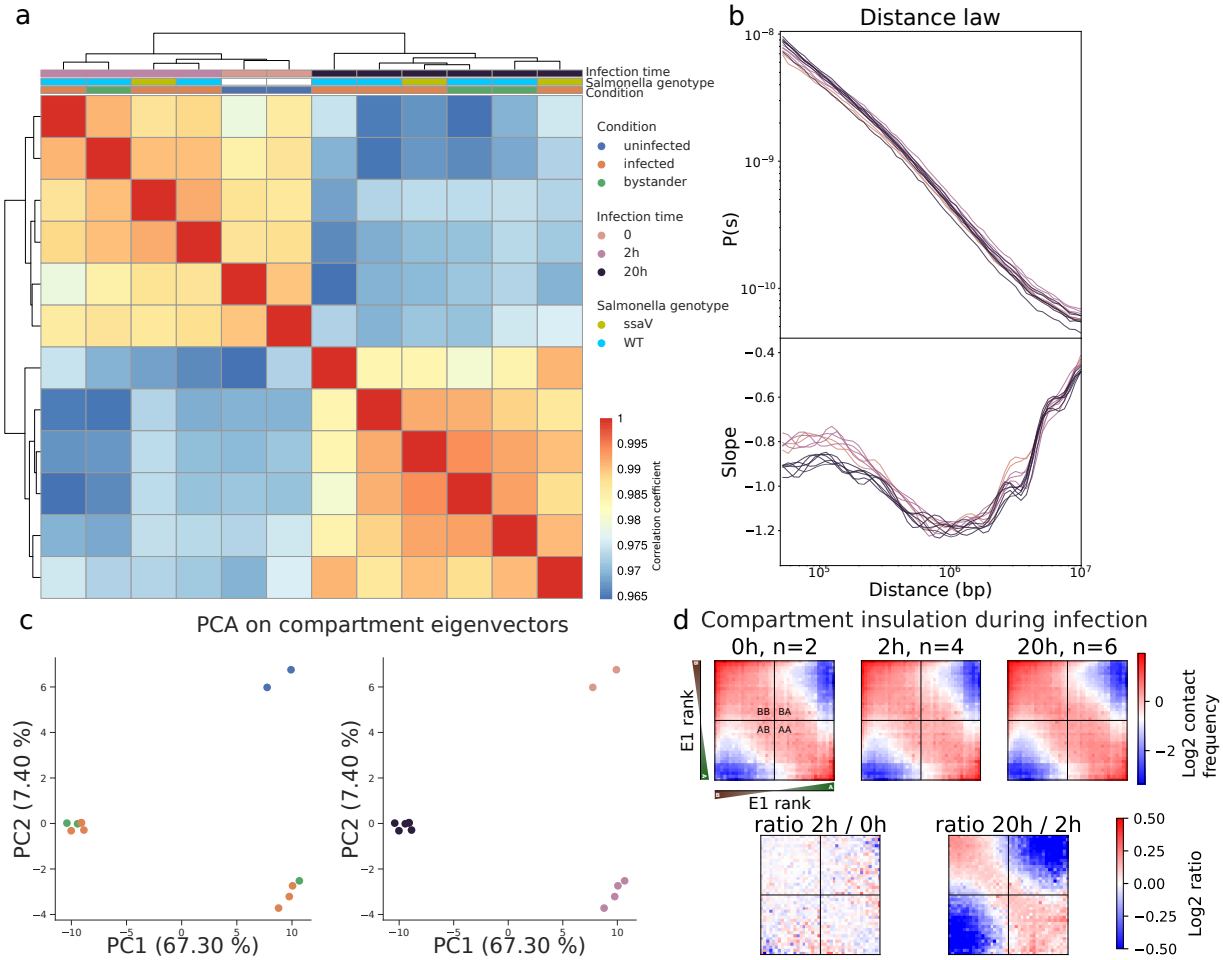


Figure 1: Global changes in genome conformation happen during late Salmonella infection. **a,** Heatmap of HiCRep stratum adjusted correlation coefficients between all pairs of samples. **b,** Distance-dependent contact decay (top) and its slope (bottom) according to infection time. **c,** PCA of chromatin compartment vectors with samples colored by condition (top) and infection time (bottom). **d,** Saddle plots showing compartment insulation intensity during infection. Hi-C interactions are binned according to their rank on the compartment eigenvector (E1) and discretized into quantiles. Saddle plots show the average intensity of interactions between each pair of eigenvector quantile (top) and their change during infection (bottom) using the Log ratio of saddles between different time points.

Global structural changes in the MHC region

We used CHESS [35] to detect structural changes occurring between early and late infection. CHESS extracted 350 features corresponding to structural changes, encompassing a total of 1,110 genes. We performed a functional enrichment analysis using Gprofiler [36] to identify gene ontology terms presenting enrichment in structural changes. Among all 155 significant terms (full list in table S1), the most strongly enriched terms are related to antigen presentation and the Major Histocompatibility Complex (MHC) (Fig. 2a). A visible compartment change is observed in the magnified contact map of the corresponding region, (Fig. 2b) as well as an insulation change, at a finer resolution. Out of the 7 MHC genes located in the region identified by CHESS (H2-Q1,2,4,6,7,10 and H2-D1), 3 show significantly higher expression during late infection (H2-Q4,6,7, log fold changes 0.6, 1.9 and 1.9).

Chromatin alterations are enriched in cell migration pathways

We ran a gene set enrichment analysis (GSEA) independently for four features to identify changes occurring during late infection (20h vs 2h): A/B compartment, chromatin loops, domain borders and gene expression. After multiple testing correction (Benjamini Hochberg, FDR rate=0.1), all four features became enriched at 20h in gene sets related to leukocyte chemotaxis and migration. To visually explore the relationships between structural features and expression, and the gene set overlap between GO terms, we generated a graph from all gene sets with (non-corrected) p-values below 0.05 using Enrichment Map [37] (methods). In this graph, each node represents the gene set of a GO term, and nodes are connected if they share at least 37.5% of their genes. Nodes are colored according to the features in which they were found to have a significant p-values. The largest connected component of this graph contains GO terms related to chemotaxis and migration, as well as other pathways (Fig. 3a). Chromatin features are limited to certain modules of that graph, while gene expression is deregulated in most nodes.

The genes associated with structural and expression changes in sets pertaining to chemotaxis mostly belong to a cluster of chemokine ligand (CXCL) genes (Fig. 3b). These genes produce small cytokines controlling the migration and adhesion of monocytes and have previously been associated with increased expression in LPS tolerance [30]. In addition, CXCL5 and CXCL9 expression is thought to be maintained through histone acetylation and methylation [30]. Our results suggest that these histone modifications are accompanied by changes in chromatin loops and borders, as well as a global switch to the A compartment (Fig. 3b).

Increase in chromatin looping at the IL-10 locus

In order to refine the position of chromatin loops anchors, we generated ATAC-seq data at 2h and 20h post infection (Methods). We intersected chromatin loop anchors with ATAC-seq accessibility peaks, and classified them based on their location (TSS, TTS, inter-gene, intronic, exonic). This classification was further expanded using publicly available ChIP-seq datasets of histone marks in BMM to include enhancer, promoter and repressed (Methods, S2). Among loops overlapping differential ATAC-seq peaks (20h vs 2h p.i), we found 36 loops anchored at the promoters of differentially expressed genes, including genes related to cell adhesion and cytoskeleton (ACTN1, ICAM5, P2RX4, TGFB1).

We also found that a chromatin loop appears next to the Il-10 gene during late infection, bridging it with the Fcmr and Il-24 genes, both of which harbour repressive marks and did not have detectable expression in our RNA-seq data. The anti-inflammatory cytokine interleukin 10 (IL-10) downregulates the inflammatory response to prevent damage to the host. Its expression is regulated by CTCF [27], and it is thought that chromatin looping coordinates the gene expression in that locus [27]. Our results forther support the role of CTCF looping in interleukin regulation.

Il-10 is activated by the TLR4 pathway which directly depends on LPS stimulation. While its expression upon LPS stimulation is known to be stronger in tolerized macrophages compared to naive macrophages, we found a lower expression of IL-10 in late infection compared to early infection (log2 fold change: -4.55, q-value: 4e-147). This is likely due to the absence of LPS-restimulation in late infection.

Discussion

In this work, we studied changes in chromatin organization of muring BMM following infection by *Salmonella enterica*. We found that most changes in global chromatin structure happened at late infection (20h p.i.), whereas it is mostly unchanged in early (2h p.i.) infection. This time point corresponds with the onset of LPS tolerance, which was shown to be dependant on histone modifications [30]. While response to acute infection is accompanied by extensive changes in local chromatin accessibility within 1h of infection [38], we found no concomitant substantial changes in 3D chromatin reorganisation. These results are largely in agreement with the proposed role of regulators of 3D chromatin structure (i.e.

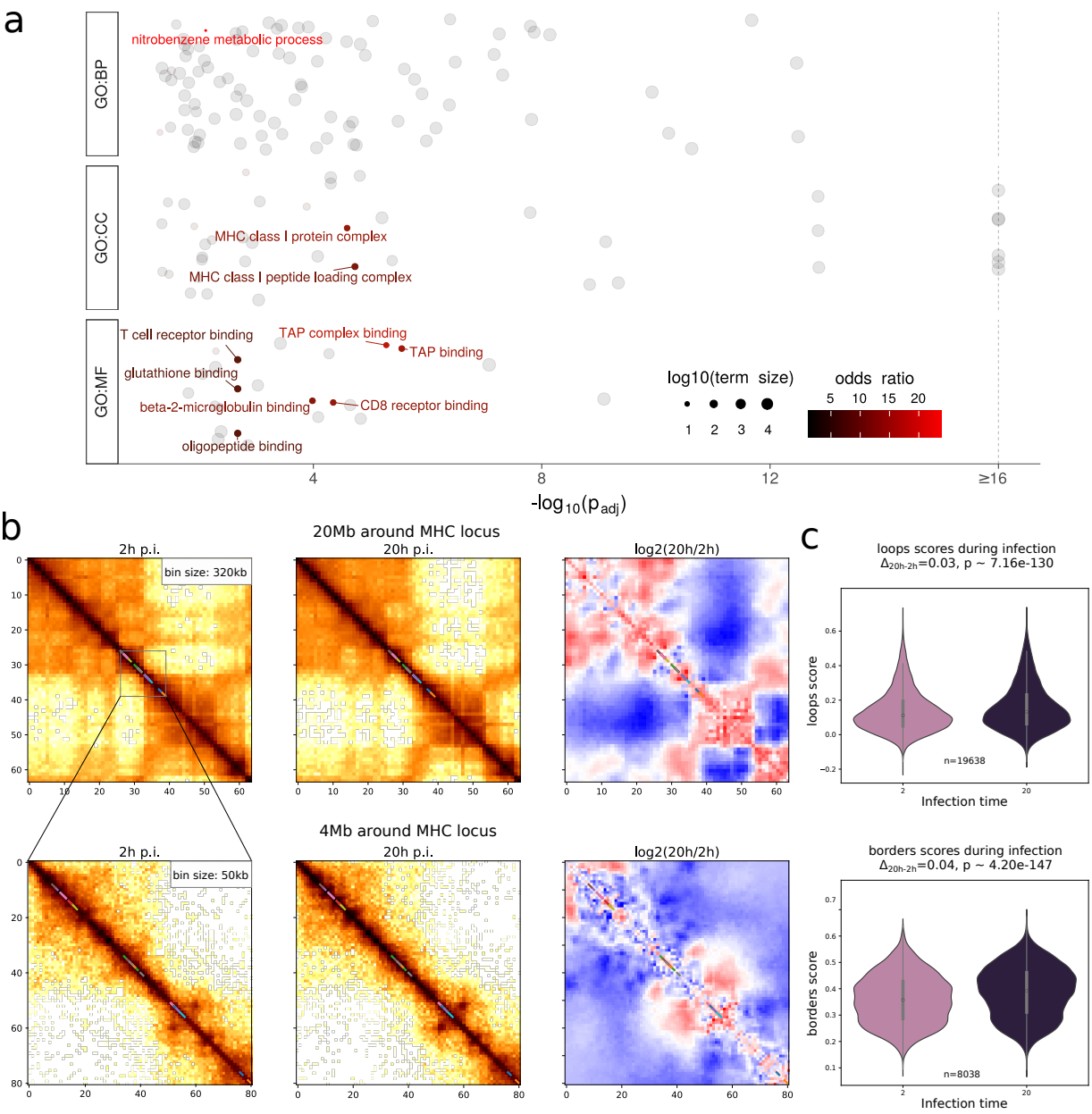


Figure 2: Hi-C changes during late infection. **a,** Overrepresentation analysis analysis of gene ontology terms in CHESS-positive regions showing structural changes during infection. The top 10 terms with highest enrichment odds ratio are highlighted. Point size represent the number of genes constituting a GO term, and the horizontal position represents the p-value from Fisher exact test adjusted for multiple testing using g:profiler's SCS algorithm [36]. Terms are split into 3 based on their GO category: CC (cellular component), BP (biological process) or MF (molecular function). **b,** Hi-C contacts around the MHC locus at low (top) and medium (bottom) resolutions. Contacts are shown during early (left) and late (middle) infection. The serpentine-binned ratio showing contact changes during infection is shown on the right. Colored lines on the main diagonal represent MHC genes identified by CHESS as part of a structural change. **c,** Distribution of loops (top) and borders (bottom) intensity throughout the whole genome at early and late infection. Pileup plots show the average 2D profile of patterns in each condition. P-values are computed using Wilcoxon's signed-rank test.

cohesin, CTCF) keeping the macrophage genome organised in a way that facilitates rapid response to TLR signalling [27, 39].

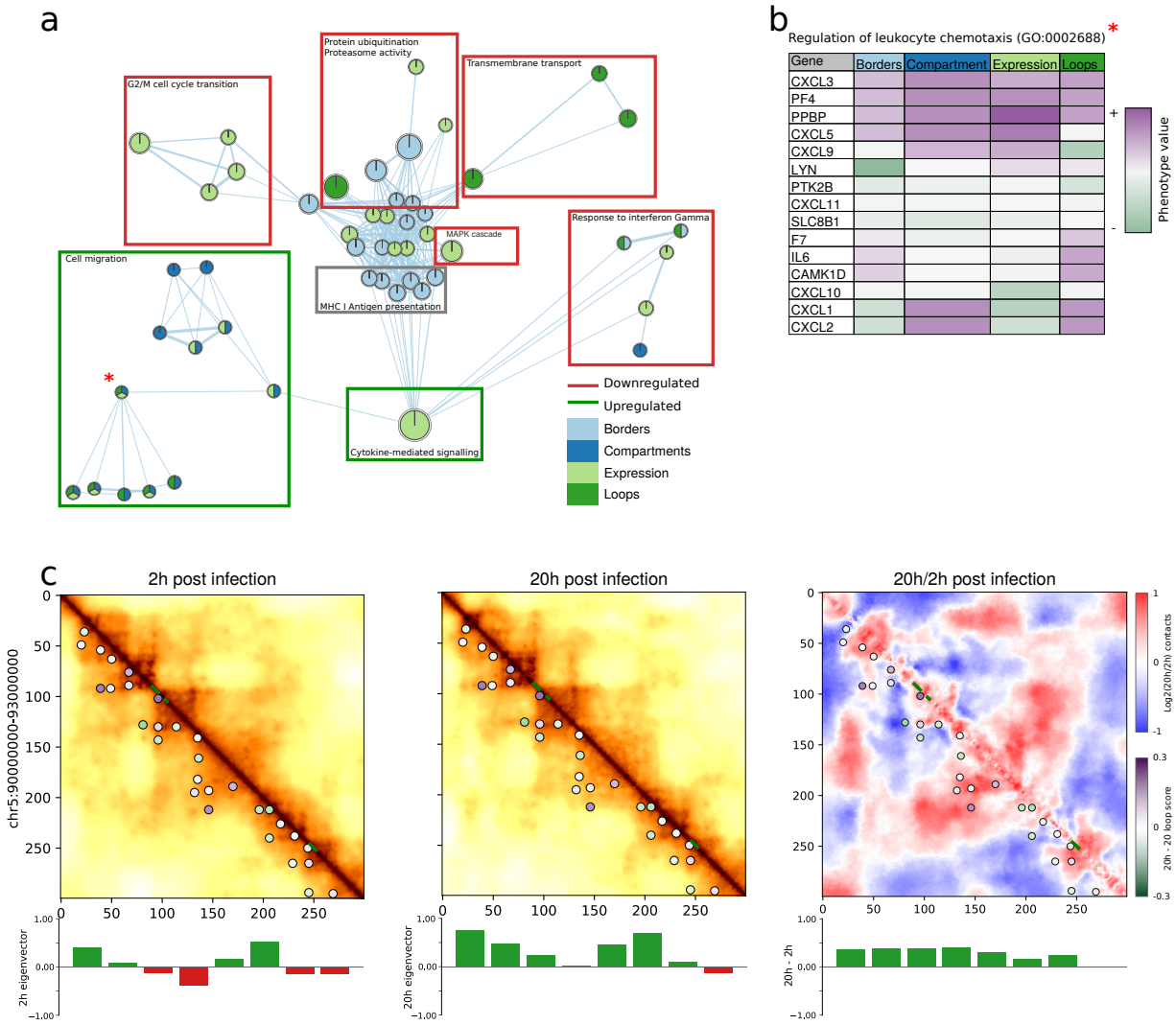


Figure 3: Gene set enrichment analysis of chromatin features. **a,** Largest connected component of the GSEA graph from chromatin features and gene expression change in . Each node is a GO:BP term, edges represent the proportion of gene overlap between terms (minimum cutoff 37.5%). Nodes are colored according to the feature (expression, compartment, border or loop change) in which they are significantly enriched during late infection (20h vs 2h). Functional subregions of the graph have been manually annotated, and the frame is colored based on its gene expression change (red: significantly downregulated, green: significantly upregulated, grey: neutral). **b,** Feature enrichment for genes involved in the GO term "Regulation of leukocyte chemotaxis", denoted by a red star on the graph. **c,** Hi-C contacts in the region containing chemokine genes CXCL3 and CXCL5. Chemokine genes are highlighted in green along the main diagonal. All matrices were binned at 10kb resolution and smoothed using Serpentine adaptive binning.

During late infection (20h), we observed an enrichment of chromatin structural changes, and a general up-regulation of genes involved in chemotaxis and cell migration. Interestingly, it was shown that M2 macrophages, which share other key characteristics with the macrophage phenotype associated with LPS tolerance, are more motile [40].

The large compartment switch we observe at the MHC locus, along with increased expression of several MHC genes of the H2-Q region are consistent with previous findings based on microscopy observations that transcriptional changes at the MHC locus are associated with chromatin reorganization [41], however the role of H2-Q family genes is still poorly understood [42].

Similarly, the chromatin looping observed at Il-10 confirms previous observations [27]. It is especially interesting that the inhibition of Il-10 expression is associated with specific interactions at chromosomal regions harboring repressive

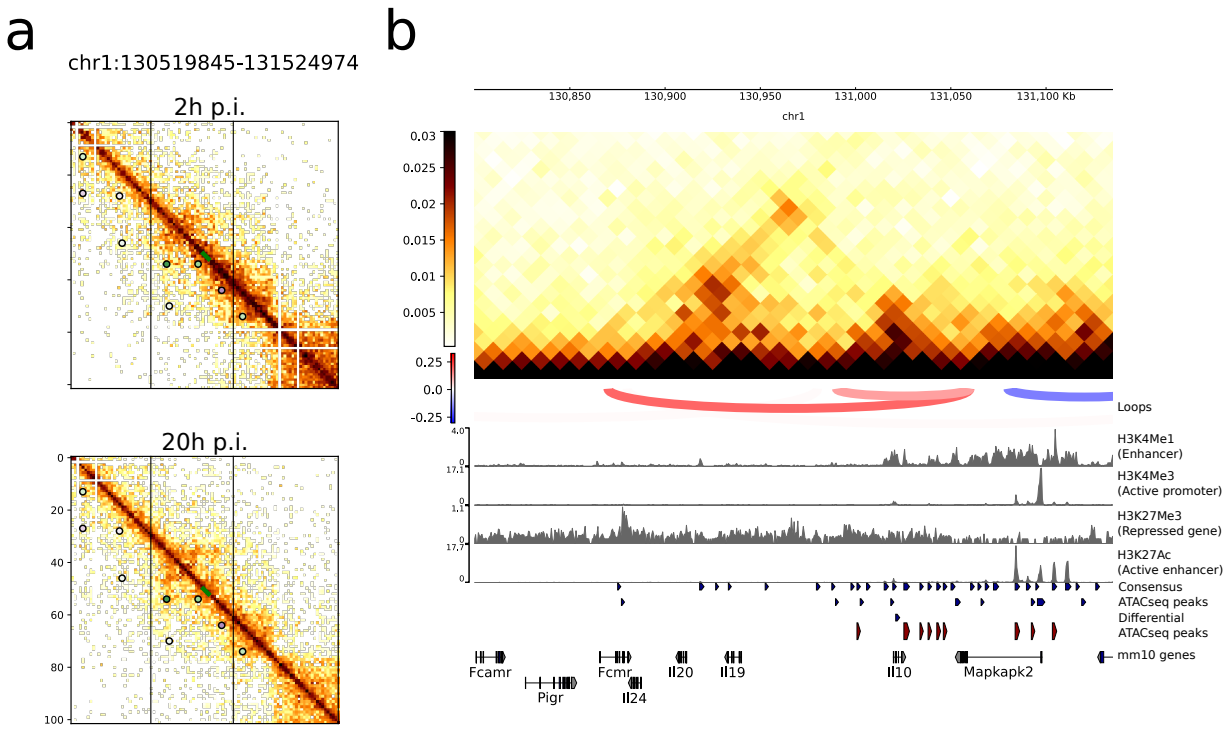


Figure 4: **a,** Contact map region in a 1Mbp region around the Il10 gene in early (2h) and late (20h) infection. Vertical black liines indicate the region shown in b. **b,** Epigenetic landscape around IL-10. A chromatin loop anchored next to the Il10 gene appears at 20h p.i. (top). The right anchor falls into a region with enhancer epigenetic marks. The left anchor falls close to Il24 and is rich in repressive marks.

marks. Further investigation will require histone ChIP-seq from infected macrophages which was not performed in this study, as histone marks used here are derived from resting BMM and may be markedly different from LPS-stimulated BMM.

Generally, the absence of chromatin changes during early infection could suggest that large structural changes operate on a slower time scale and correspond to the establishment of long term tolerance.

Methods

Libraries preparation

Hi-C library preparation

Hi-C libraries were prepared according to the Arima protocol using only the DpnII and HinfI enzymes. Libraries were sequenced using paired end sequencing at 35bp on an Illumina NextSeq 500 machine.

ATAC-seq library preparation

ATAC-seq library preparation was carried out according to the published Omni-ATAC protocol [43]. Libraries were sequenced on an Illumina NextSeq 500 using paired end sequencing.

RNA-seq library preparation

RNA was isolated using the Quick DNA-RNA Miniprep Kit following the manufacturer's protocol. Following RNA isolation, macrophage rRNA was depleted using the NEBNext rRNA Depletion Kit following the manufacturer's protocol. RNA-Seq libraries were generated using the NEBNext Ultra II Directional RNA Library Kit for Illumina and

the NEBNext Multiplex Oligos for Illumina following the manufacturer's protocol. Libraries were sequenced on an Illumina NextSeq 500 using single read sequencing.

RNA-Seq/ATAC-Seq infections and FACS sorting

Wild type or ΔssaV *Salmonella enterica* serovar Typhimurium (strain SL1344) expressing the pFCcGssaG plasmid (i.e. ssaG promoter expressed GFP, constitutive mCherry) [44] were grown overnight in MgMES pH 5.0 medium (170 mM 2-(N-morpholino)ethanesulfonic acid (MES) at pH 5.0, 5 mM KCl, 7.5 mM $(NH_4)_2SO_4$, 0.5 mM K_2SO_4, 1 mM KH_2PO_4, 8 mM $MgCl_2$, 38 mM glycerol, and 0.1% casamino acids). Stationary phase bacteria were then opsonized with 8% mouse serum for 20 minutes and added to the BMDM at a MOI of 10. At 30 min post-infection, macrophages were washed 3x with PBS, and fresh BMDM medium containing gentamicin (50 μg/ml) was added. At 2.5 h post-infection, macrophages were washed 1x with PBS, and fresh BMDM medium containing gentamicin (10 μg/ml) was added. Prior to isolation by FACS, uninfected macrophages and macrophages at 2h and 18h post-infection were washed with PBS three times and detached from the surface with cold PBS and scraping. Macrophages were then centrifuged at 4 °C at 300g, the supernatant discarded, and macrophages resuspended in cold sterile PBS. For infected macrophages, 5E4 macrophages containing either wild type or ΔssaV Salmonella (mCherry+ population) were isolated by FACS. For uninfected macrophages, 5E4 macrophages were isolated by FACS. For FACS isolation, apoptotic macrophages and doublets were excluded by gating, and the samples were sorted under continuous cooling to 4 °C by a BD Aria III into cold sterile PBS. Isolated macrophages in PBS were then centrifuged at 500g for 5 min at 4 °C, the supernatant was removed, and macrophage pellets were either used immediately for ATAC-seq library preparation or snap frozen in liquid nitrogen and stored at -80 °C until RNA was isolated for RNA-Seq library preparation.

Analyses

All analyses were done using the mm10 reference genome assembly.

Differential accessibility peaks

ATACseq data was processed using the nf-core/atacseq pipeline (v1.2.1). Within nfcore/atacseq, consensus peaks were obtained using MACS2 (v2.2.7.1) and differential peaks from DESeq2 with FDR<0.05 were selected.

Histone marks ChIPseq

Publicly available histone mark ChIPseq datasets were retrieved from ENCODE and processed using the nf-core/chipseq pipeline (v1.2.1) in single-end mode. The following marks and respective accession numbers were used: H3K4Me1 (ENCSR000CFE), H3K4Me2 (SRR930721, SRR930722), H3K4Me3 (ENCSR000CFF), H3K27Me3 (SRR930746), H3K27Ac (ENCSR000CFD).

Differential expression

Libraries were aligned using Hisat2 (1.24.0.123) and transcripts were quantified into TPM using salmon (v0.14.1). Differential expression was measured between 2h and 20h p.i. using DESeq2.

Hi-C analyses

Hi-C matrices were generated using hicstuff (v3.0.1) [45]. Matrix balancing (normalization) was performed using the Cooler implementation of the ICE algorithmm [46] Compartments were extracted using the cooltools API [47].

Reproducibility between replicates was assessed using the hicreppy implementation (`https://github.com/cmdoret/hicreppy`)of the HiCrep algorithm [33]. Briefly, a correlation coefficient is computed between pairs of sample for each diagonal separately, and a weighted average of correlation coefficients is then returned, with the weights being inversely proportional to the genomic distance corresponding to each diagonal. The operation is performed separately on each chromosome, and the average of all chromosomes, weighted by their length is used.

Chromatin loops and domain borders were detected using chromosight [48] (v1.5.1) and pattern intensity changes between conditions were computed using pareidolia (v.0.6.0) [49]. Compartment segmentation was performed using cooltools (v0.3.2) [47] using the correlation with gene density to orient eigenvectors. Hi-C matrices were binned at 320kb for compartment detection and Hi-C rep and 10kb for all other analyses.

Each gene was assigned the closest loop anchor and domain border within 200kb (if any). CHESS was used to identify genes located in regions undergoing major structural changes during infection.

Gene set enrichment analysis

GSEA was performed using the python package gseapy [50]. The analysis was run 4 times independently on different phenotypes representing changes between 2h and 20h p.i. The phenotypes used were: differential pattern scores (loop and borders) from Pareidolia, compartment eigenvector differences from cooltools, and gene expression log2 fold change from DeSeq2. For structural phenotypes, values were assigned to genes using bedtools genome arithmetic operations [51]: Borders and loops were assigned to the closest gene within 200kb, and compartment values were assigned to genes using bedtools intersect.

For visualizing the network graph, the union of all terms with p-values below 0.05 (without multiple testing correction) for all phenotypes was used. The graph was generated using cytoscape with the enrichmentMap plugin [37]. Nodes were colored by dataset (i.e. where the phenotypes' p-values are below 0.05) for visualization.

Integration of epigenomic data

Chromatin loops in figure S2 were intersected with differntial ATAC peaks to refine retain only loops with both anchors within 10kb (a margin of 1 pixel on the Hi-C contact map) of differentially accessible ATAC peaks (FDR <0.05). Average normalized histone mark intensity scores were assigned to each peak and K-means clustering (k=3) was used on those intensities to classify anchors into 3 groups: promoter (highest H3K4Me3), enhancer (highest H3K4Me1) or low activity (other peaks). Peaks were histone marks were not available were labelled "unknown". Promoter anchors were further refined to include only those located in promoter regions (-1kb to +100bp from TSS).

Code availability

All codes to reproduce analyses is available on a Github repository at `https://github.com/cmdoret/mouse_salmonella_infection.git` where data processing is packaged into a Snakemake pipeline, and downstream analyses are provided as jupyter notebooks.

References

[1] B A McCormick, S I Miller, D Carnes, and J L Madara. Transepithelial signaling to neutrophils by salmonellae: A novel virulence mechanism for gastroenteritis. *Infection and Immunity*, 63(6):2302–2309, June 1995.

[2] B A McCormick, P M Hofman, J Kim, D K Carnes, S I Miller, and J L Madara. Surface attachment of Salmonella typhimurium to intestinal epithelia imprints the subepithelial matrix with gradients chemotactic for neutrophils. *Journal of Cell Biology*, 131(6):1599–1608, December 1995.

[3] Nicholas Arpaia, Jernej Godec, Laura Lau, Kelsey E. Sivick, Laura M. McLaughlin, Marcus B. Jones, Tatiana Dracheva, Scott N. Peterson, Denise M. Monack, and Gregory M. Barton. TLR Signaling Is Required for Salmonella typhimurium Virulence. *Cell*, 144(5):675–688, March 2011.

[4] Andrea Haraga, Maikke B. Ohlson, and Samuel I. Miller. Salmonellae interplay with host cells. *Nature Reviews Microbiology*, 6(1):53–66, January 2008.

[5] M Martínez-Moya, M. A de Pedro, H Schwarz, and F García-del Portillo. Inhibition of Salmonella intracellular proliferation by non-phagocytic eucaryotic cells. *Research in Microbiology*, 149(5):309–318, May 1998.

[6] Hanna K. de Jong, Chris M. Parry, Tom van der Poll, and W. Joost Wiersinga. Host–Pathogen Interaction in Invasive Salmonellosis. *PLOS Pathogens*, 8(10):e1002933, October 2012.

[7] K. Z. Abshire and F. C. Neidhardt. Growth rate paradox of Salmonella typhimurium within host macrophages. *Journal of Bacteriology*, 175(12):3744–3748, June 1993.

[8] Daphne A. C. Stapels, Peter W. S. Hill, Alexander J. Westermann, Robert A. Fisher, Teresa L. Thurston, Antoine-Emmanuel Saliba, Isabelle Blommestein, Jörg Vogel, and Sophie Helaine. Salmonella persisters undermine host immune defenses during antibiotic treatment. *Science*, 362(6419):1156–1160, December 2018.

[9] John J. O'Shea and Peter J. Murray. Cytokine signaling modules in inflammatory responses. *Immunity*, 28(4):477–487, April 2008.

[10] David C. Dale, Laurence Boxer, and W. Conrad Liles. The phagocytes: Neutrophils and monocytes. *Blood*, 112(4):935–945, August 2008.

[11] Subhra K. Biswas and Eduardo Lopez-Collazo. Endotoxin tolerance: New mechanisms, molecules and clinical significance. *Trends in Immunology*, 30(10):475–487, October 2009.

[12] Sebastian E. Winter, Parameth Thiennimitr, Maria G. Winter, Brian P. Butler, Douglas L. Huseby, Robert W. Crawford, Joseph M. Russell, Charles L. Bevins, L. Garry Adams, Renée M. Tsolis, John R. Roth, and Andreas J. Bäumler. Gut inflammation provides a respiratory electron acceptor for Salmonella. *Nature*, 467(7314):426–429, September 2010.

[13] Sumati Murli, Robert O. Watson, and Jorge E. Galán. Role of tyrosine kinases and the tyrosine phosphatase SptP in the interaction of Salmonella with host cells. *Cellular Microbiology*, 3(12):795–810, 2001.

[14] Doris L. LaRock, Anu Chaudhary, and Samuel I. Miller. Salmonellae interactions with host processes. *Nature Reviews Microbiology*, 13(4):191–205, April 2015.

[15] Haihua Ruan, Zhen Zhang, Li Tian, Suying Wang, Shuangyan Hu, and Jian-Jun Qiao. The Salmonella effector SopB prevents ROS-induced apoptosis of epithelial cells by retarding TRAF6 recruitment to mitochondria. *Biochemical and Biophysical Research Communications*, 478(2):618–623, September 2016.

[16] Athma A. Pai, Golshid Baharian, Ariane Pagé Sabourin, Jessica F. Brinkworth, Yohann Nédélec, Joseph W. Foley, Jean-Christophe Grenier, Katherine J. Siddle, Anne Dumaine, Vania Yotova, Zachary P. Johnson, Robert E. Lanford, Christopher B. Burge, and Luis B. Barreiro. Widespread Shortening of 3' Untranslated Regions and Increased Exon Inclusion Are Evolutionarily Conserved Features of Innate Immune Responses to Infection. *PLOS Genetics*, 12(9):e1006338, September 2016.

[17] Marcelo B. Sztein, Andrea C. Bafford, and Rosângela Salerno-Goncalves. Salmonella enterica serovar Typhi exposure elicits ex vivo cell-type-specific epigenetic changes in human gut cells. *Scientific Reports*, 10(1):13581, August 2020.

[18] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, October 2009.

[19] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, May 2012.

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

[21] J. Dekker. Capturing Chromosome Conformation. *Science*, 295(5558):1306–1311, February 2002.

[22] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, 15(9):2038–2049, May 2016.

[23] Diego Ottaviani, Elliott Lever, Shihong Mao, Rossitza Christova, Babatunji W. Ogunkolade, Tania A. Jones, Jaroslaw Szary, Johan Aarum, Muhammad A. Mumin, Christopher A. Pieri, Stephen A. Krawetz, and Denise Sheer. CTCF binds to sites in the major histocompatibility complex that are rapidly reconfigured in response to interferon-gamma. *Nucleic Acids Research*, 40(12):5262–5270, July 2012.

[24] Rossitza Christova, Tania Jones, Pei-Jun Wu, Andreas Bolzer, Ana P. Costa-Pereira, Diane Watling, Ian M. Kerr, and Denise Sheer. P-STAT1 mediates higher-order chromatin remodelling of the human MHC in response to IFNγ. *Journal of Cell Science*, 120(18):3262–3270, September 2007.

[25] Parimal Majumder, Jorge A. Gomez, Brian P. Chadwick, and Jeremy M. Boss. The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. *The Journal of Experimental Medicine*, 205(4):785–798, April 2008.

[26] Parimal Majumder and Jeremy M. Boss. Cohesin regulates major histocompatibility complex class II genes through interactions with MHC-II insulators. *Journal of immunology (Baltimore, Md. : 1950)*, 187(8):4236–4244, October 2011.

[27] Tatjana Nikolic, Dowty Movita, Margaretha EH Lambers, Claudia Ribeiro de Almeida, Paula Biesta, Kim Kreefft, Marjolein JW de Bruijn, Ingrid Bergen, Niels Galjart, Andre Boonstra, and Rudi Hendriks. The DNA-binding factor Ctcf critically controls gene expression in macrophages. *Cellular & Molecular Immunology*, 11(1):58–70, January 2014.

[28] Shutao Cai, Charles C. Lee, and Terumi Kohwi-Shigematsu. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nature Genetics*, 38(11):1278–1288, November 2006.

[29] Mohamed El Gazzar, Barbara K. Yoza, Jean Y. Q. Hu, Sue L. Cousart, and Charles E. McCall. Epigenetic Silencing of Tumor Necrosis Factor α during Endotoxin Tolerance*. *Journal of Biological Chemistry*, 282(37):26857–26864, September 2007.

[30] Simmie L. Foster, Diana C. Hargreaves, and Ruslan Medzhitov. Gene-specific control of inflammation by TLR-induced chromatin modifications. *Nature*, 447(7147):972–978, June 2007.

[31] Hnin Thanda Aung, Kate Schroder, Stewart R. Himes, Kristian Brion, Wendy Van Zuylen, Angela Trieu, Harukazu Suzuki, Yoshihide Hayashizaki, David A. Hume, Matthew J. Sweet, and Timothy Ravasi. LPS regulates proinflammatory gene expression in macrophages by altering histone deacetylase expression. *The FASEB Journal*, 20(9):1315–1327, 2006.

[32] Jörg Mages, Harald Dietrich, and Roland Lang. A genome-wide analysis of LPS tolerance in macrophages. *Immunobiology*, 212(9-10):723–737, January 2008.

[33] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C Hardison, William Stafford, Feng Yue, and Qunhua Li. HiCRep: Assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. page 37.

[34] John J. Seeley and Sankar Ghosh. Molecular mechanisms of innate memory and tolerance to LPS. *Journal of Leukocyte Biology*, 101(1):107–119, January 2017.

[35] Silvia Galan, Nick Machnik, Kai Kruse, Noelia Díaz, Marc A. Marti-Renom, and Juan M. Vaquerizas. CHESS enables quantitative comparison of chromatin contact data and automatic feature extraction. *Nature Genetics*, 52(11):1247–1255, November 2020.

[36] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. G:profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, July 2019.

[37] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D. Bader. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLOS ONE*, 5(11):e13984, November 2010.

[38] Sergi Cuartero, Felix D. Weiss, Gopuraja Dharmalingam, Ya Guo, Elizabeth Ing-Simmons, Silvia Masella, Irene Robles-Rebollo, Xiaolin Xiao, Yi-Fang Wang, Iros Barozzi, Dounia Djeghloul, Mariane T. Amano, Henri Niskanen, Enrico Petretto, Robin D. Dowell, Kikuë Tachibana, Minna U. Kaikkonen, Kim A. Nasmyth, Boris Lenhard, Gioacchino Natoli, Amanda G. Fisher, and Matthias Merkenschlager. Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation. *Nature Immunology*, 19(9):932–941, September 2018.

[39] Grégoire Stik, Enrique Vidal, Mercedes Barrero, Sergi Cuartero, Maria Vila-Casadesús, Julen Mendieta-Esteban, Tian V. Tian, Jinmi Choi, Clara Berenguer, Amaya Abad, Beatrice Borsari, François le Dily, Patrick Cramer, Marc A. Marti-Renom, Ralph Stadhouders, and Thomas Graf. CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. *Nature Genetics*, 52(7):655–661, July 2020.

[40] Laurel E. Hind, Emily B. Lurier, Micah Dembo, Kara L. Spiller, and Daniel A. Hammer. Effect of M1–M2 Polarization on the Motility and Traction Stresses of Primary Human Macrophages. *Cellular and Molecular Bioengineering*, 9(3):455–465, September 2016.

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

[42] Katharine J. Goodall, Angela Nguyen, Lucy C. Sullivan, and Daniel M. Andrews. The expanding role of murine class Ib MHC in the development and activation of Natural Killer cells. *Molecular Immunology*, 115:31–38, November 2019.

[43] M. Ryan Corces, Alexandro E. Trevino, Emily G. Hamilton, Peyton G. Greenside, Nicholas A. Sinnott-Armstrong, Sam Vesuna, Ansuman T. Satpathy, Adam J. Rubin, Kathleen S. Montine, Beijing Wu, Arwa Kathiria, Seung Woo Cho, Maxwell R. Mumbach, Ava C. Carter, Maya Kasowski, Lisa A. Orloff, Viviana I. Risca, Anshul Kundaje, Paul A. Khavari, Thomas J. Montine, William J. Greenleaf, and Howard Y. Chang. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*, 14(10):959–962, October 2017.

[44] Rita Figueira, Kathryn G. Watson, David W. Holden, and Sophie Helaine. Identification of Salmonella Pathogenicity Island-2 Type III Secretion System Effectors Involved in Intramacrophage Replication of S. enterica Serovar Typhimurium: Implications for Rational Vaccine Design. *mBio*, 4(2):e00065–13.

[45] Cyril Matthey-Doret, Lyam Baudry, Amaury Bignaud, Axel Cournac, Remi Montagne, Nadège Guiglielmoni, Foutel-Rodier Théo, and Scolari Vittore F. Simple library/pipeline to generate and handle Hi-C data. https://github.com/koszullab/hicstuff, March 2021.

[46] Nezar Abdennur and Leonid A Mirny. Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, 36(1):311–316, January 2020.

[47] Open2c/cooltools. Open Chromosome Collective, April 2021.

[48] Cyril Matthey-Doret, Lyam Baudry, Axel Breuer, Rémi Montagne, Nadège Guiglielmoni, Vittore Scolari, Etienne Jean, Arnaud Campeas, Philippe Henri Chanut, Edgar Oriol, Adrien Méot, Laurent Politis, Antoine Vigouroux, Pierrick Moreau, Romain Koszul, and Axel Cournac. Computer vision for pattern detection in chromosome contact maps. *Nature Communications*, 11(1):5795, November 2020.

[49] Cyril Matthey-Doret. Koszullab/pareidolia: V0.6.1. https://zenodo.org/record/5062485, July 2021.

[50] Zhuoqing Fang. Gseapy: Gene Set Enrichment Analysis in Python.

[51] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

Supplementary figures

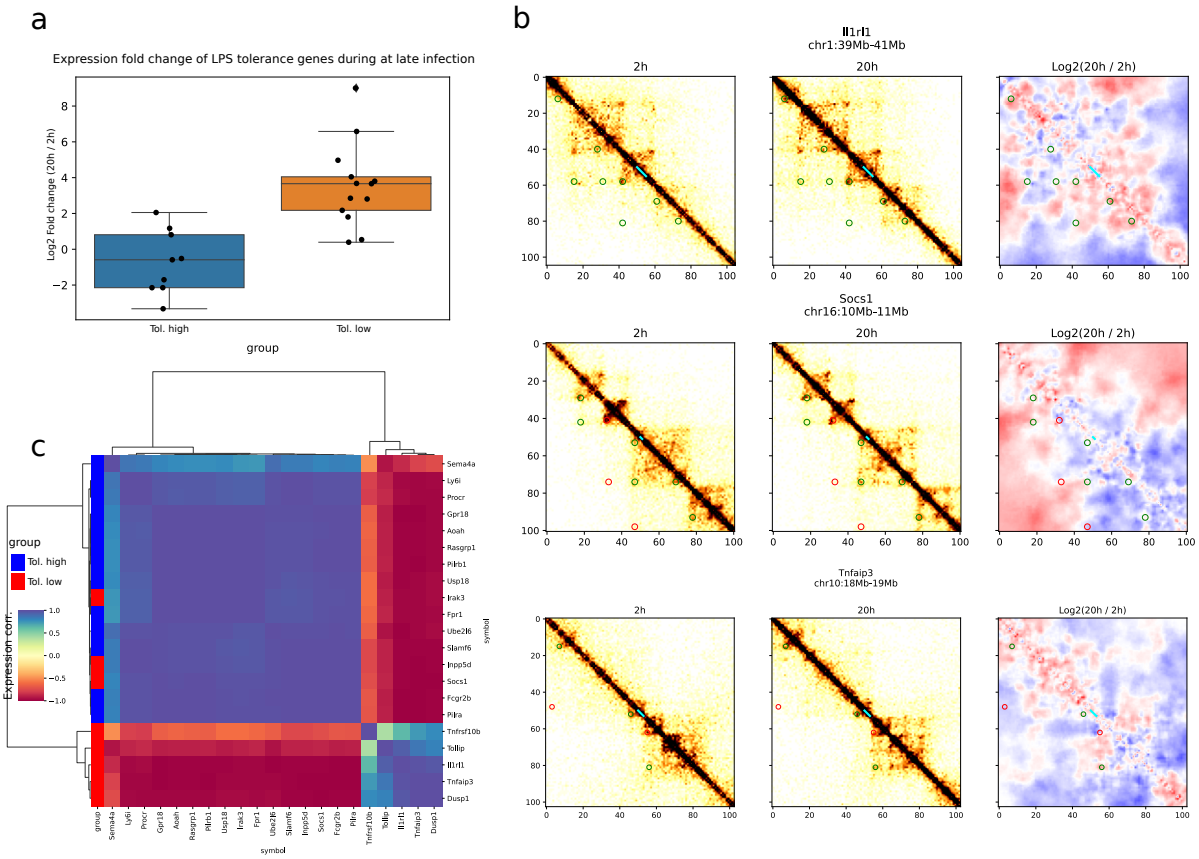


Figure S1: Analysis of select LPS response marker genes from [32]. **a,** Expression of 22 genes known to be positive (Tol. up) or negative (Tol. low) regulators of the LPS response, in our RNAseq results. **b,** Example Hi-C regions from genes with strong loop changes. Contacts at early (2h, left) late (20h, middle) and change during late infection (20h/2h, right) are shown. All matrices were binned at 10kb and ratios are smoothed using Serpentine adaptive binning. **c,** Gene expression correlation between LPS marker genes. Pearson correlation across all 4 samples is shown (duplicates at 2h and 20h).

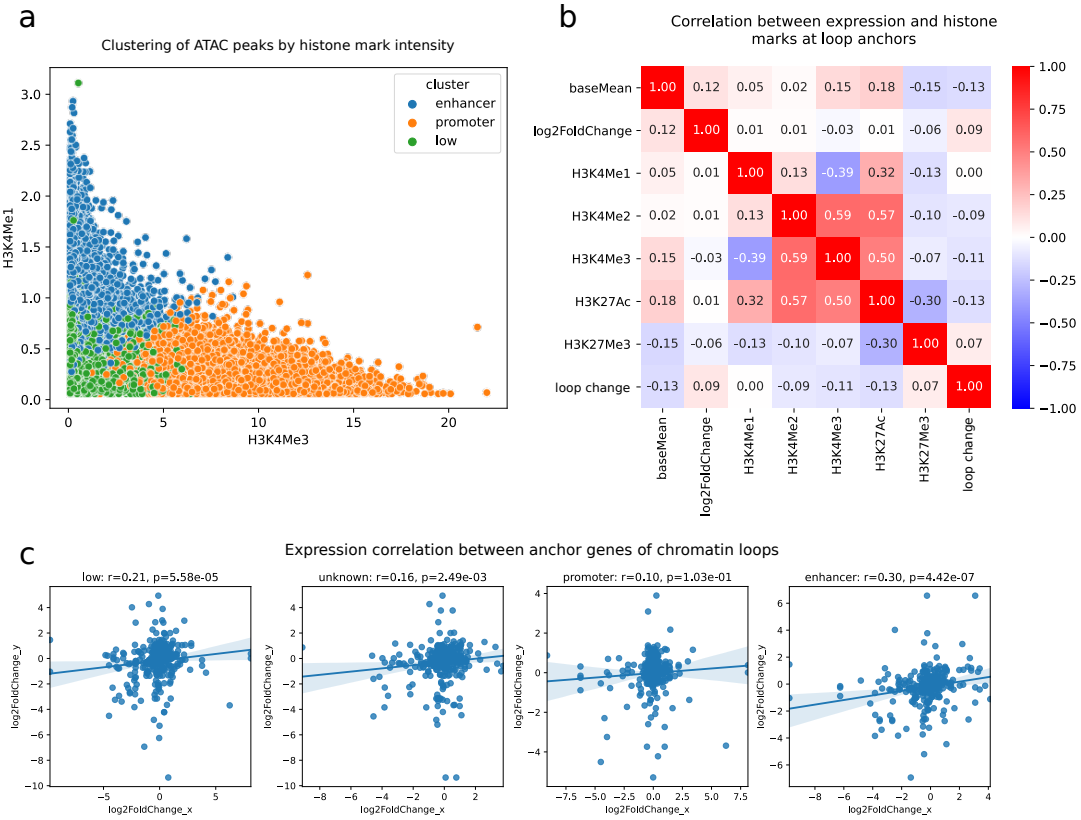


Figure S2: Analysis of epigenetic marks at loop anchors. **a,** Distribution of ATACseq peaks based on histone mark intensities H3K4Me3 and H3K4Me1. Colors represent cluster value assigned on the basis of 5 histone marks (H3K4Me1, H3K4Me2, H3K4Me3, H3K27Ac, H3K27Me3). **b,** Correlation between base gene expression (baseMean), gene expression fold change during infection (log2FoldChange), histone marks and loop intensity change during infection (loop change) at loop anchors. **c,** Expression correlation between gene pairs at loop anchors based on loop categories. Loop categories are defined as either anchor containing a histone mark-derived annotation.

Table S1: Table of significantly enriched GO terms for genes in CHESS-detected regions with structural changes.

source	name	id	logqval	size	intersect
GO:MF	protein binding	GO:0005515	9.089	10189	534
GO:MF	binding	GO:0005488	7.078	14895	712
GO:MF	TAP binding	GO:0046977	5.547	12	8
GO:MF	TAP complex binding	GO:0062061	5.276	9	7
GO:MF	enzyme binding	GO:0019899	4.827	2164	144
GO:MF	identical protein binding	GO:0042802	4.644	2140	142
GO:MF	CD8 receptor binding	GO:0042610	4.346	11	7
GO:MF	transcription factor binding	GO:0008134	4.272	571	53
GO:MF	protein-containing complex binding	GO:0044877	4.083	1464	104
GO:MF	beta-2-microglobulin binding	GO:0030881	3.982	12	7
GO:MF	ion binding	GO:0043167	3.421	5862	314
GO:MF	anion binding	GO:0043168	3.026	2420	149
GO:MF	transferase activity	GO:0016740	2.852	2277	141
GO:MF	T cell receptor binding	GO:0042608	2.673	17	7
GO:MF	oligopeptide binding	GO:1900750	2.673	17	7
GO:MF	glutathione binding	GO:0043295	2.673	17	7
GO:MF	organic cyclic compound binding	GO:0097159	2.382	5768	303

source	name	id	logqval	size	intersect
GO:MF	glutathione transferase activity	GO:0004364	2.366	33	9
GO:MF	catalytic activity	GO:0003824	2.332	5665	298
GO:MF	peptide antigen binding	GO:0042605	2.292	19	7
GO:MF	heterocyclic compound binding	GO:1901363	2.286	5672	298
GO:MF	transcription coregulator activity	GO:0003712	2.273	448	40
GO:MF	protein homodimerization activity	GO:0042803	2.063	676	53
GO:MF	protein dimerization activity	GO:0046983	1.809	1052	73
GO:BP	metabolic process	GO:0008152	12.491	11429	598
GO:BP	cellular metabolic process	GO:0044237	12.464	10350	554
GO:BP	organic substance metabolic process	GO:0071704	11.676	10989	577
GO:BP	primary metabolic process	GO:0044238	10.626	10243	542
GO:BP	cellular nitrogen compound metabolic process	GO:0034641	10.222	6201	364
GO:BP	nitrogen compound metabolic process	GO:0006807	9.931	9669	515
GO:BP	macromolecule metabolic process	GO:0043170	8.141	9484	499
GO:BP	biosynthetic process	GO:0009058	7.867	5882	339
GO:BP	cellular macromolecule metabolic process	GO:0044260	7.821	7962	432
GO:BP	cellular biosynthetic process	GO:0044249	7.801	5715	331
GO:BP	organic cyclic compound metabolic process	GO:1901360	7.314	5829	334
GO:BP	cellular aromatic compound metabolic process	GO:0006725	7.256	5618	324
GO:BP	organic substance biosynthetic process	GO:1901576	7.165	5779	331
GO:BP	nucleobase-containing compound metabolic process	GO:0006139	6.484	5416	311
GO:BP	cellular response to stress	GO:0033554	6.394	1805	132
GO:BP	heterocycle metabolic process	GO:0046483	6.381	5535	316
GO:BP	cellular macromolecule biosynthetic process	GO:0034645	6.147	4744	278
GO:BP	regulation of cellular biosynthetic process	GO:0031326	5.978	4088	246
GO:BP	cellular nitrogen compound biosynthetic process	GO:0044271	5.903	4705	275
GO:BP	macromolecule biosynthetic process	GO:0009059	5.773	4782	278
GO:BP	regulation of biosynthetic process	GO:0009889	5.593	4167	248
GO:BP	nucleic acid metabolic process	GO:0090304	5.483	4962	285
GO:BP	protein localization	GO:0008104	5.040	2452	161
GO:BP	regulation of cellular macromolecule biosynthetic process	GO:2000112	4.908	3880	231
GO:BP	organonitrogen compound metabolic process	GO:1901564	4.822	6326	344
GO:BP	aromatic compound biosynthetic process	GO:0019438	4.752	4042	238
GO:BP	nucleobase-containing compound biosynthetic process	GO:0034654	4.706	3963	234
GO:BP	heterocycle biosynthetic process	GO:0018130	4.684	4028	237
GO:BP	organic cyclic compound biosynthetic process	GO:1901362	4.628	4181	244
GO:BP	regulation of macromolecule biosynthetic process	GO:0010556	4.581	3913	231
GO:BP	RNA metabolic process	GO:0016070	4.305	4470	256
GO:BP	transcription DNA-templated	GO:0006351	4.236	887	211
GO:BP	nucleic acid-templated transcription	GO:0097659	4.217	3534	211
GO:BP	RNA biosynthetic process	GO:0032774	4.068	3549	211
GO:BP	gene expression	GO:0010467	4.055	5933	322
GO:BP	cellular protein metabolic process	GO:0044267	3.910	4815	270
GO:BP	regulation of transcription DNA-templated	GO:0006355	3.796	887	203
GO:BP	regulation of nucleic acid-templated transcription	GO:1903506	3.777	3413	203
GO:BP	regulation of RNA biosynthetic process	GO:2001141	3.738	3417	203
GO:BP	protein metabolic process	GO:0019538	3.677	5474	299
GO:BP	regulation of nitrogen compound metabolic process	GO:0051171	3.582	5706	309
GO:BP	cellular macromolecule localization	GO:0070727	3.453	1726	117
GO:BP	localization	GO:0051179	3.440	6168	329
GO:BP	macromolecule localization	GO:0033036	3.393	2897	176
GO:BP	cellular protein localization	GO:0034613	3.360	1714	116
GO:BP	regulation of cellular metabolic process	GO:0031323	3.318	6073	324
GO:BP	regulation of primary metabolic process	GO:0080090	3.221	5886	315
GO:BP	cellular localization	GO:0051641	3.164	2777	169
GO:BP	nitrogen compound transport	GO:0071705	3.094	1986	129
GO:BP	regulation of nucleobase-containing compound metabolic process	GO:0019219	3.059	3931	224
GO:BP	regulation of RNA metabolic process	GO:0051252	2.941	3691	212

source	name	id	logqval	size	intersect
GO:BP	regulation of macromolecule metabolic process	GO:0060255	2.835	6363	334
GO:BP	positive regulation of cellular biosynthetic process	GO:0031328	2.755	1956	126
GO:BP	establishment of localization	GO:0051234	2.720	4658	256
GO:BP	positive regulation of nitrogen compound metabolic process	GO:0051173	2.714	3069	181
GO:BP	positive regulation of nucleic acid-templated transcription	GO:1903508	2.650	1616	108
GO:BP	regulation of metabolic process	GO:0019222	2.646	6841	354
GO:BP	positive regulation of RNA biosynthetic process	GO:1902680	2.637	1617	108
GO:BP	establishment of protein localization	GO:0045184	2.502	1647	109
GO:BP	macromolecule modification	GO:0043412	2.358	3823	215
GO:BP	positive regulation of biosynthetic process	GO:0009891	2.268	1999	126
GO:BP	positive regulation of macromolecule metabolic process	GO:0010604	2.149	3467	197
GO:BP	nitrobenzene metabolic process	GO:0018916	2.111	4	4
GO:BP	cellular component organization or biogenesis	GO:0071840	2.090	6270	325
GO:BP	positive regulation of metabolic process	GO:0009893	2.007	3760	210
GO:BP	localization within membrane	GO:0051668	2.003	654	53
GO:BP	protein transport	GO:0015031	1.984	1535	101
GO:BP	regulation of gene expression	GO:0010468	1.978	4888	262
GO:BP	protein localization to membrane	GO:0072657	1.948	589	49
GO:BP	transport	GO:0006810	1.929	4502	244
GO:BP	positive regulation of nucleobase-containing compound metabolic process	GO:0045935	1.920	1912	120
GO:BP	organic substance transport	GO:0071702	1.903	2416	145
GO:BP	positive regulation of cellular metabolic process	GO:0031325	1.817	3275	186
GO:BP	positive regulation of RNA metabolic process	GO:0051254	1.804	1745	111
GO:BP	positive regulation of macromolecule biosynthetic process	GO:0010557	1.799	1844	116
GO:BP	transcription by RNA polymerase II	GO:0006366	1.762	2492	148
GO:BP	cellular protein modification process	GO:0006464	1.745	3644	203
GO:BP	protein modification process	GO:0036211	1.745	3644	203
GO:BP	cellular response to organic substance	GO:0071310	1.744	2453	146
GO:BP	cellular response to chemical stimulus	GO:0070887	1.725	3097	177
GO:BP	cellular component organization	GO:0016043	1.655	6090	314
GO:BP	positive regulation of protein-containing complex assembly	GO:0031334	1.598	240	26
GO:BP	antigen processing and presentation of peptide antigen	GO:0048002	1.513	64	12
GO:BP	establishment of localization in cell	GO:0051649	1.512	2031	124
GO:BP	cellular response to organic cyclic compound	GO:0071407	1.497	595	48
GO:BP	response to stress	GO:0006950	1.404	3817	209
GO:BP	negative regulation of cellular process	GO:0048523	1.348	5068	266
GO:BP	xenobiotic catabolic process	GO:0042178	1.312	15	6
GO:CC	intracellular anatomical structure	GO:0005622	29.657	14111	750
GO:CC	membrane-bounded organelle	GO:0043227	24.868	11812	654
GO:CC	organelle	GO:0043226	23.998	12827	690
GO:CC	intracellular membrane-bounded organelle	GO:0043231	22.652	11363	630
GO:CC	intracellular organelle	GO:0043229	22.543	12510	674
GO:CC	cytoplasm	GO:0005737	19.392	11093	609
GO:CC	intracellular organelle lumen	GO:0070013	12.853	4581	296
GO:CC	organelle lumen	GO:0043233	12.840	4582	296
GO:CC	membrane-enclosed lumen	GO:0031974	12.840	4582	296
GO:CC	nucleus	GO:0005634	9.341	7208	404
GO:CC	nuclear lumen	GO:0031981	9.119	4062	256
GO:CC	nucleoplasm	GO:0005654	8.838	3560	230
GO:CC	cytosol	GO:0005829	7.797	3849	240
GO:CC	intracellular protein-containing complex	GO:0140535	5.380	712	64
GO:CC	protein-containing complex	GO:0032991	5.201	5312	297
GO:CC	bounding membrane of organelle	GO:0098588	4.812	1676	117
GO:CC	MHC class I peptide loading complex	GO:0042824	4.727	15	8
GO:CC	MHC class I protein complex	GO:0042612	4.590	11	7
GO:CC	endomembrane system	GO:0012505	4.303	4006	231
GO:CC	mitochondrion	GO:0005739	4.303	1827	123
GO:CC	organelle membrane	GO:0031090	4.074	3076	185

source	name	id	logqval	size	intersect
GO:CC	endoplasmic reticulum exit site	GO:0070971	3.882	31	10
GO:CC	nuclear protein-containing complex	GO:0140513	3.060	1168	83
GO:CC	intracellular non-membrane-bounded organelle	GO:0043232	3.041	4500	247
GO:CC	non-membrane-bounded organelle	GO:0043228	2.928	4515	247
GO:CC	Golgi medial cisterna	GO:0005797	2.817	24	8
GO:CC	Golgi apparatus	GO:0005794	2.804	1449	97
GO:CC	endoplasmic reticulum	GO:0005783	2.761	1761	113
GO:CC	endoplasmic reticulum protein-containing complex	GO:0140534	2.220	135	18
GO:CC	endoplasmic reticulum membrane	GO:0005789	2.138	993	70
GO:CC	catalytic complex	GO:1902494	2.103	1334	88
GO:CC	endoplasmic reticulum subcompartment	GO:0098827	2.054	999	70
GO:CC	nuclear outer membrane-endoplasmic reticulum membrane network	GO:0042175	2.047	1018	71
GO:CC	COPII-coated ER to Golgi transport vesicle	GO:0030134	1.990	58	11
GO:CC	MHC protein complex	GO:0042611	1.912	23	7
GO:CC	perinuclear region of cytoplasm	GO:0048471	1.846	776	57
GO:CC	organelle subcompartment	GO:0031984	1.737	1580	99
GO:CC	intercellular bridge	GO:0045171	1.470	77	12
GO:CC	intrinsic component of endoplasmic reticulum membrane	GO:0031227	1.451	154	18
GO:CC	mitochondrial outer membrane	GO:0005741	1.354	185	20
GO:CC	cytoplasmic vesicle membrane	GO:0030659	1.346	489	39

III

Discussion and conclusion

In this last part, we discuss the findings from this work and place them in the context of the field in its current state. We also discuss the limitations of our approaches and how to improve on it. Finally we briefly reflect upon future developments and perspectives for chromosome conformation analyses in infection biology.

1 Biological and technical discussions

1.1 Representation of protozoan genomes

The representation of species with fully sequenced genomes is traditionally skewed towards mammals and vertebrate animals [182]. Many groups which are much more abundant in nature or ecologically important are underrepresented in reference genome databases [183]. *Acanthamoeba* provides a fitting example, as they are ubiquitous in aquatic environment and form important interactions with numerous other microorganisms, yet have no chromosome quality reference available. This, despite being used in numerous studies as the host of viruses or bacteria [31, 184, 185]. The reference genomes of the two *A. castellanii* strains generated here provide the first high quality reference in the *Acanthamoeba* group and thus represent a valuable resource for the comparative study of amoebae.

1.2 Host plasticity of intracellular bacteria

Throughout the previous part, we have developed new approaches to detect chromatin features and quantify their changes during infection (chapter 1). We then applied these methods in two different infection settings: infection of the amoeba *A. castellanii* by *L. pneumophila* (chapter 2) and of murine bone marrow-derived macrophages by *S. enterica* (chapter 3). Although both are intracellular bacteria with similar infection strategies, they can infect very different eukaryotic hosts with highly divergent evolutionary histories. The size of the mouse haploid genome outclasses that of *A. castellanii* by two orders of magnitude (4.3 Gbp vs 45 Mbp) and its spatial organization appears much more complex, with A/B compartmentation and intricate nested loops bridging very long distances. Despite all their differences, both unicellular and human hosts are susceptible to *L. pneumophila* infection. This is most impressive knowing that human is an evolutionary dead-end for *L. pneumophila* due to the absence of human-to-human transmission. The bacterial genome is therefore shaped exclusively by selective pressure in its natural unicellular hosts. The ability to infect multicellular hosts is probably associated to the high conservation of the targeted pathways and has been attributed to the wide range of protozoan hosts infected by the bacterium [186].

This conservation was also visible in our results, as several processes deregulated during infection are common between *Legionella* and *Salmonella*, such as cell cycle regulation, cytoskeleton organization, protein ubiquitination and transmembrane transport.

1.3 Combination of effects

The analyses presented in this work focus on the description of changes happening in global chromatin structure during infection. One issue with this type of experiments is that we observe the combined effect of the pathogen activity and the host immune response. There are means to dampen one of these effects, such as the use of mutant pathogens which are unable to secrete effector proteins as control to trigger host response (as used in Chapter chapter 3). Although these controls do not completely emulate the pathogen activity, as it will not replicate [187] and therefore will not elicit the same immune response, they are still useful to separate the effect of the infection and pathogen exposure. In the case of *L. pneumophila* (chapter 2), reproducing the infections with mutants for the dotA secretion system and romA methyl-transferase would allow further isolation of the chromatin changes caused by the infection and romA activity.

At large scale, decoupling and deciphering the individual factors at play during infection ultimately requires the use of mutagenesis screens, such as Transposon insertion or CRISPR. Such approaches assess the effect on host survival and not chromatin changes, and to our knowledge there is no method to screen for chromatin structure modifiers, the closest being MAP-C, a screen to assess the effect of mutations on contacts between a pair of predetermined loci [188]. Regardless, more descriptive approaches such as the ones used in this work are still important to understand the extent of changes happening during infection. Specifically, they can still inform us on the type of structural changes that the genome undergoes and global importance of genome organization during infection. They can also outline discrete regions of change in the contact pattern, and point at regulatory pathways perturbations induced by the pathogens.

1.4 Power limitations

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

It was proposed that RNA-seq experiments for differential expression analysis should comprise at least 6 replicates and ideally 12 [189]. 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. Although Hi-C contacts may be less susceptible to technical variataion (especially at shorter range) and exhibit spatial dependency, there can still be substantial variability across biological replicates [106].

The core issue with low replicate numbers is the lack of power to distinguish between biological variability across replicates and differences due to the condition of interest. As a consequence, when fewer replicates are used, lower effect sizes (fold changes in the case of gene expression) become undetectable. This is especially problematic when studying gene regulation, where small changes in expression could be relevant.

Unlike RNA-seq, where the standard for analyses is well established and most softwares can account for replicates and experimental design, most methods available for Hi-C analysis do not leverage replicate information. This restricts the power of analysis to the detection of major changes.

1.5 Reproducibility and reliability challenges

The lack of standards for Hi-C data formats and processing causes a general fragmentation of bioinformatic tools, with many redundant softwares 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 validate each logic block of the software using inputs with known truths, as such, they could be viewed as an equivalent to control experiments in molecular biology. Software lacking these controls is more likely to contain undetected bugs that could impact results and potentially lead to false conclusions.

Some general practices can be adopted to address these issues, such as writing comprehensive documentation, solid tests and ensure long term software maintenance, but as it stands there is little incentive to do so in academia. Such incentives could come in the form of dedicated fundings for open source bioinformatics software development, for which candidates would be evaluated on the quality of their tools rather than traditional bibliometric indicators which are poorly correlated with software quality [190]. Ultimately, some of these quality criteria should be also enforced globally in the publishing process to ensure that tools meet quality

standards for publication, as was done for example by the Journal of Open Source Software [191].

Although adopting such practices would increase the effort and time required to develop methods, the resulting tools would be more reliable, easier to use and more widely adopted, thus benefiting the global research community in the long run. Fortunately, recent years have seen an increasing adoption of good practices in bioinformatic software. One such example is the *nf-core* ecosystem backed by SciLifeLab, a public institution dedicated to open-source scientific software development https://www.scilifelab.se. The generalization of similar initiatives could mean that the quality of bioinformatics software will undergo major improvements in the foreseeable future [192, 193]. The general scientific community has seen other open science successes, providing a more positive outlook on the future of scientific software quality. Notable examples include the Zenodo platform for the sharing and long term archival of scientific software and data [194], or the non-profit NumFOCUS [195] which supports general scientific software development.

2 Perspectives of genomics for infection biology

2.1 The 3D genome and the advent of deep learning

It can look attractive to produce a model of the 3D genome, ideally a predictive one, that would be able to infer how the structure reacts to specific changes. However, in many organisms, the rules governing genome organization are intricate and it would be unwieldy to model them explicitly. Deep learning provides an attractive framework to produce such a model without knowing all the rules involved. There are already successful applications of deep learning in biology for various different tasks such as gene annotation [196, 197], variant calling [198], classification of coding RNA [199], prediction of nucleosome positioning [200] and perhaps most importantly, protein structure prediction [201]. More recently progress has also been made for the prediction prediction gene expression and promoter-enhancer interactions solely from the DNA sequence [202].

Generally, applications of deep learning in the 3D genome field have been limited to denoising or improving the resolution of Hi-C matrices. Recently however, there have been successful attempts at predicting the structure of mammalian genomes from the DNA sequence, including (and most importantly) the prediction of conformational changes induced by mutations [203, 204]. In the future, these approaches could be helpful to identify mutations or regions to focus on, and one could imagine it being used to model the consequences of infection on the host.

Unfortunately, several limitations must be overcome before deep learning methods can become an amenable tool to understand the relationship between biological processes and genome structure. First, it requires tremendous amounts of training data, which in the case of Hi-C remains expensive to generate. Then, such model would also require information about all the factors at play in the process, which we do not know. Even in the event that we manage to obtain a model that effectively predicts conformational changes, in most cases this is still unsatisfying: The general scientific interest is usually to understand the rules and logic that connect the biological process (e.g. infection) to structural changes. In deep learning models these rules are obscured, taking the form of large weight matrices, and extracting biological

meaning from them would require consequent advances in model interpretability [205].

3 Perspectives

As 3C protocols improve, as I have observed during the course of the last three years, and the cost of sequencing decreases, it becomes possible to probe finer details of spatial regulation during bacterial infection. While current projects are mostly limited to analyzing major changes, higher sequencing depth and increasing numbers of replicates will allow for more contrast and with it, the detection of more subtle changes in spatial interactions [206].

Another exciting perspective is the advent of single-cell omics methods. This is especially interesting for infection genomics, where bulk Hi-C signal contains a mixture of cells at different infection stage and cell cycle phase. These single-cell methodologies may allow further refining the analysis and deconvolute different effects obscuring the signal of interest.

In future years, we expect to see major developments in the use of 3D genomics to understand the deregulation induced by infection. There is still much to be learnt in the interplay of the various layers of regulation, and spatial organization will likely become an integrative part of many projects aiming to understand it. This work allowed us to observe the general chromosomal biology of *A. castellanii*, but it would be interesting to study the behaviour of specific features in more details, such as the role of subtelomeric or rDNA clustering. This will also require additional effort to resolve repeated sequences in the assembly. Exploring different infection systems, such as different amoebae hosts, bacteria, or even megaviruses would also provide more insights into whether there are conserved hallmarks of spatial chromosomal changes during infection.

More generally, as more infection studies integrate Hi-C data with epigenetic marks and expression, it will be interesting to study in more details the interplay of structural changes with regulatory information to better understand the factors determining the importance of long range interactions in gene regulation.

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 [207]. 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 matrix, 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}

B Chromosight case study

B.1 Quantification of metaphasic loops in yeast

Example use of Chromosight: Loops during yeast metaphase

August 27, 2021

In this notebook, we demonstrate how `chromosight quantify` can be used to compare chromatin loops between *S. cerevisiae* cultures arrested in G1 phase vs metaphase. In this notebook, we re-analyse Hi-C data from Garcia-Luis, J., Lazar-Stefanita, L., Gutierrez-Escribano, P. et al., 2019.

Input data:

Files used in this analysis are the output from `chromosight quantify`. Loop scores were computed on all 2-way combinations from a set of high confidence RAD21 binding sites separated by 10 to 50kb, on two Hi-C datasets at 2kb resolution: One with G1-arrested cells and the other with metaphase-arrested cells.

- `scer_w303_g1_2kb_SRR8769554.cool`: Hi-C matrix of cells stopped in G1 phase, at 2kb resolution. From Dauban et al. 2020
- `scer_w303_mitotic_2kb_merged.cool`: Hi-C matrix of metaphasic cells, at 2kb resolution. From Garcia-Luis et al. 2019
- `rad21.bed2d`: bed file containing all pairs of positions of RAD21 (cohesin) peaks in metaphasic *S. cerevisiae* separated by 10-50kb.

  Note: see the end of this notebook for an explanation on how to generate a bed2d file from a ChIP-seq bed file.

Getting loop scores

Loop scores at all pairs of positions can be computed using `chromosight quantify`. However, to ensure scores are comparable, the number of contacts should be similar between matrices. When using cool files, cooler can be used for this operation:

```
$ cooler info input/scer_w303_mitotic_2kb_merged.cool | grep sum
    "sum": 44048750

$ cooler info input/scer_w303_g1_2kb_SRR8769554.cool | grep sum
    "sum": 5862820
```

The G1 matrix has around 5.8M contacts whereas the metaphase matrix has 44M. Fortunately, chromosight has a `--subsample` option, which can be used to bring both matrices to the same coverage before computing scores:

```
chromosight quantify --pattern loops \
                     --subsample 5862820 \
                     --win-fmt npy \
                     scer_cohesin_peaks.bed2d \
                     input/scer_w303_g1_2kb_SRR8769554.cool \
                     quantify/rad21_g1

chromosight quantify --pattern loops \
                     --subsample 5862820 \
                     --win-fmt npy \
                     input/scer_cohesin_peaks.bed2d \
                     input/scer_w303_mitotic_2kb_merged.cool \
                     quantify/rad21_metaphase
```

For each condition, `chromosight quantify` generates 2 files:

- A table containing the coordinates and pattern matching scores of all input coordinates.
- A numpy binary file containing a stack of images around the input coordinates. Those images are stored in the same order as the coordinates from the table.

```
quantify/
├── rad21_g1.npy
├── rad21_g1.tsv
├── rad21_metaphase.npy
└── rad21_metaphase.tsv
```

Analysing loop scores

We can now use python to load and compare results from chromosight quantify. Below are a series of analyses showing some examples of downstream processing that can be performed on chromosight results.

[203]:
```
%config InlineBackend.figure_format = 'svg'
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.ndimage as ndi
import chromosight.kernels as ck
import scipy.stats as st
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

res = 2000
```

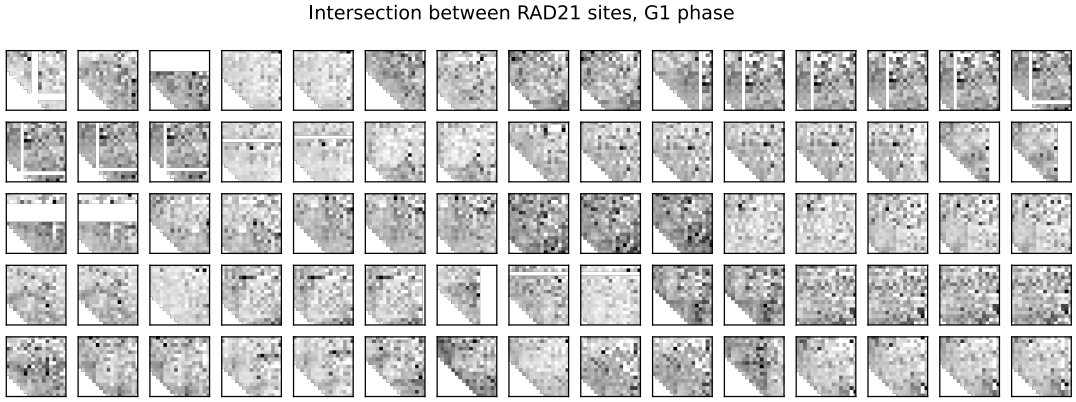
[204]:
```
# Load images (vignettes) around RAD21 interactions coordinates
images_g = np.load('quantify/rad21_g1.npy')
images_m = np.load('quantify/rad21_metaphase.npy')

# Load lists of RAD21 interactions coordinates with their loop scores
# Compute loop size (i.e. anchor distance) for each RAD21 combination
get_sizes = lambda df: np.abs(df.start2 - df.start1)
loops_g = pd.read_csv('quantify/rad21_g1.tsv', sep='\t')
loops_g['loop_size'] = get_sizes(loops_g)
loops_m = pd.read_csv('quantify/rad21_metaphase.tsv', sep='\t')
loops_m['loop_size'] = get_sizes(loops_m)

# Merge data from both conditions into a single table
loops_g['condition'] = 'g1'
loops_m['condition'] = 'metaphase'
loops_df = pd.concat([loops_g, loops_m]).reset_index(drop=True)
images = np.concatenate([images_g, images_m])

# Remove NaN scores (e.g. in repeated regions or overlap the matrix edge)
nan_mask = ~np.isnan(loops_df['score'])
loops_df = loops_df.loc[nan_mask, :]
images = images = images[nan_mask, :, :]

# The loop kernel can be loaded using chromosight.kernels.loops
kernel = np.array(ck.loops['kernels'][0])
pileup_kw = {'vmin': -1, 'vmax': 1, 'cmap': 'seismic'}
```

Peeking at the input coordinates

Images around RAD21 sites 2-way combinations extracted by chromosight can be viewed using numpy and matplotlib. Note there are series off overlapping and slightly shifted images. This is because of adjacent RAD21 sites which are closer in the genome than the size of the vignettes.

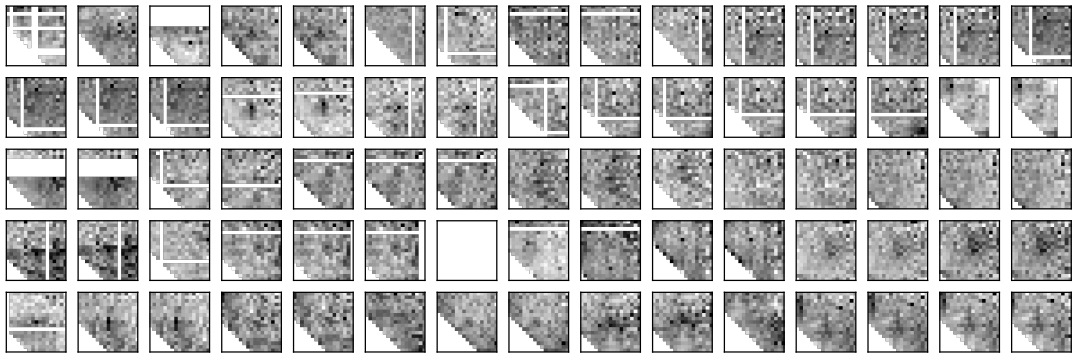
[193]:
```
# Decide how many rows and columns of images to show
r, c = 5, 15
valid_imgs = np.where(~loops_g.score.isnull() & ~loops_g.score.isnull())[0]
fig, axes = plt.subplots(r, c, figsize=(12, 4), subplot_kw={'xticks':[],␣
↪'yticks':[]})
# Show each image as a greyscale vignette
for i, ax in zip(valid_imgs, axes.flat):
    img = images_g[i, :, :] # Showing examples from the end of the image stack␣
↪(M phase)
    ax.imshow(img, cmap=plt.cm.gray_r, interpolation='nearest')
plt.suptitle("Intersection between RAD21 sites, G1 phase")
```

[193]: Text(0.5, 0.98, 'Intersection between RAD21 sites, G1 phase')



[194]:
```
fig, axes = plt.subplots(r, c, figsize=(12, 4), subplot_kw={'xticks':[],␣
↪'yticks':[]})
first_m = np.where(loops_df.condition == 'metaphase')[0][0]
# Show each image as a greyscale vignette
for i, ax in zip(valid_imgs, axes.flat):
    img = images_m[i, :, :] # Showing examples from the end of the image stack␣
↪(M phase)
    ax.imshow(img, cmap=plt.cm.gray_r, interpolation='nearest')
plt.suptitle("Intersection between RAD21 sites, Metaphase")
```

[194]: Text(0.5, 0.98, 'Intersection between RAD21 sites, Metaphase')

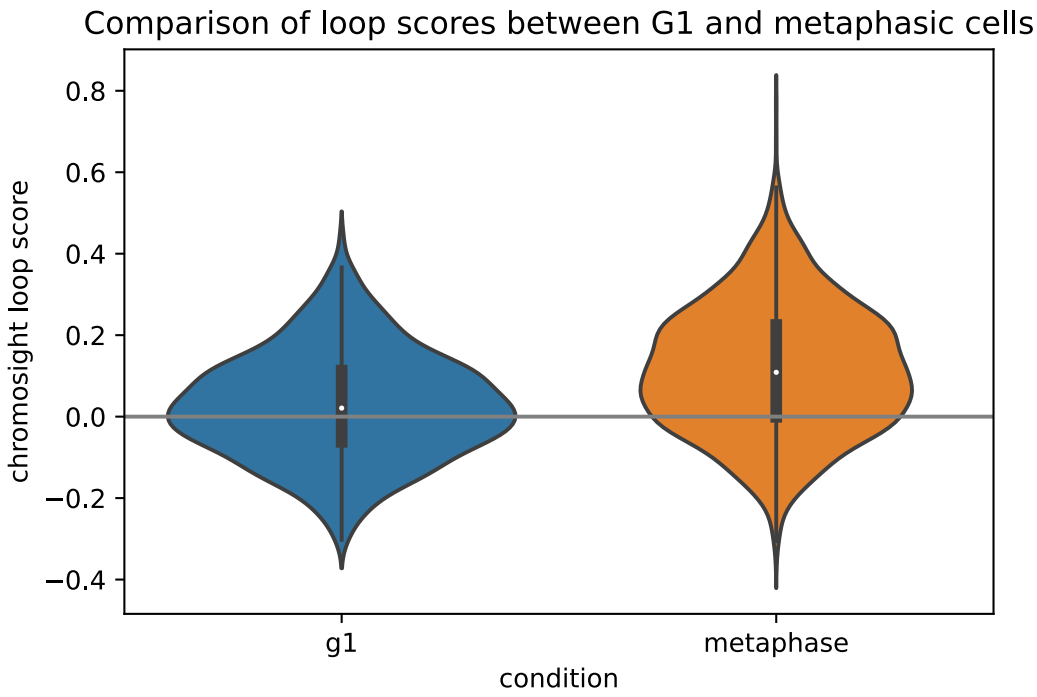


Comparing the distribution of scores

The distribution of chromosight scores (i.e. correlation coefficients with the loop kernel) can be compared between the 2 conditions, revealing that metaphasic cells tend to have stronger loops.

[196]:
```
sns.violinplot(data=loops_df, x='condition', y='score')
plt.ylabel('chromosight loop score')
plt.title('Comparison of loop scores between G1 and metaphasic cells')
plt.axhline(0, c='grey')
```

[196]: <matplotlib.lines.Line2D at 0x7f53f4892a50>

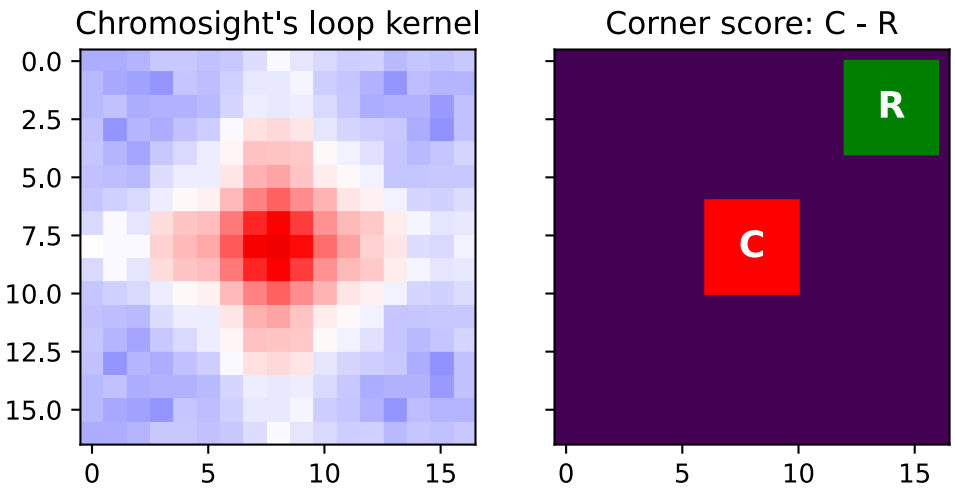


Using different metrics

Chromosight scores loops using their pearson correlation with a "loop kernel" (see below). However, one might want to use another metric than chromosight's score to rank loops. One such metric commonly used in the litterature is the "corner score", which uses the contrast between the center of the image (C) and the corner (R).

[197]:
```
import matplotlib.patches as patches
fig, axes = plt.subplots(1, 2, sharex=True, sharey=True)
axes[0].imshow(np.log(kernel), **pileup_kw)
axes[0].set_title("Chromosight's loop kernel")
axes[1].imshow(np.zeros((17, 17)))
center_rect = patches.Rectangle(
    (8-2, 8-2), 4, 4, linewidth=1, edgecolor='r', facecolor='r'
)
corner_rect = patches.Rectangle(
    (17-5, 0), 4, 4, linewidth=1, edgecolor='g', facecolor='g'
)
axes[1].annotate('C', (8, 8), color='w', weight='bold', fontsize=14,␣
 ↪ha='center', va='center')
axes[1].annotate('R', (14, 2), color='w', weight='bold', fontsize=14,␣
 ↪ha='center', va='center')
axes[1].add_patch(center_rect)
axes[1].add_patch(corner_rect)
axes[1].set_title("Corner score: C - R")
```

[197]:
```
Text(0.5, 1.0, 'Corner score: C - R')
```



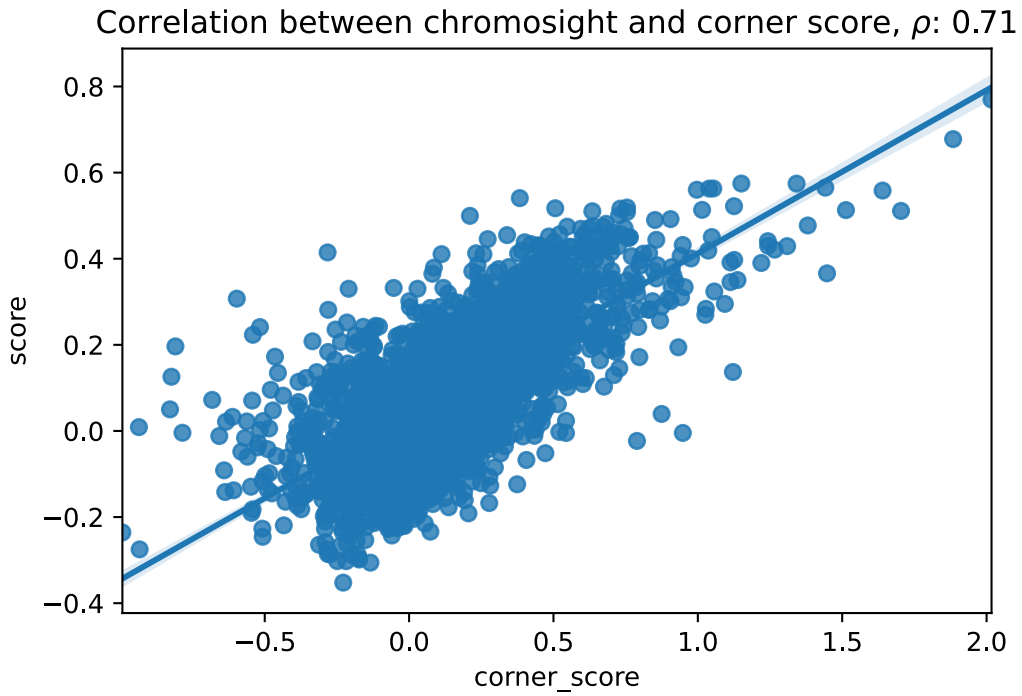
The function defined below could be used to compute the corner score. It computes the difference between the average of contacts in the center and top right corner. Using the rop right corner is better to avoid contacts enrichments for due to the diagonal. This is a pretty intuitive metric tailored based on expectations we have about loops. Here, we define center and corner radii as 10% of the image radius. For our 17x17 images, this means both regions will be 2+1 = 3x3 pixels.

[198]:
```
def corner_score(image, prop_radius=0.1):
    """
    Compute a loop intensity score from a pileup

    Parameters
    ----------
    image : numpy.array of floats
        2D array representing the window around a pattern.
    prop_radius : float
        Proportion of image radius used when selecting
        center and corner contacts.

    Returns
    -------
    float :
        Corner score, defined as mean(center) - mean(corner).
    """
    n, m = image.shape
    center = int(prop_radius * n)
    half_h = n // 2
    half_w = m // 2
    le = half_h - center
    ri = half_h + center + 1
    hi = half_w - center
    lo = half_w + center + 1
    center_mean = np.nanmean(image[hi:lo, le:ri])
    top_right_mean = np.nanmean(image[:hi, ri:])
    return center_mean - top_right_mean
```

This homemade corner score correlates well with chromosight's pearson score:

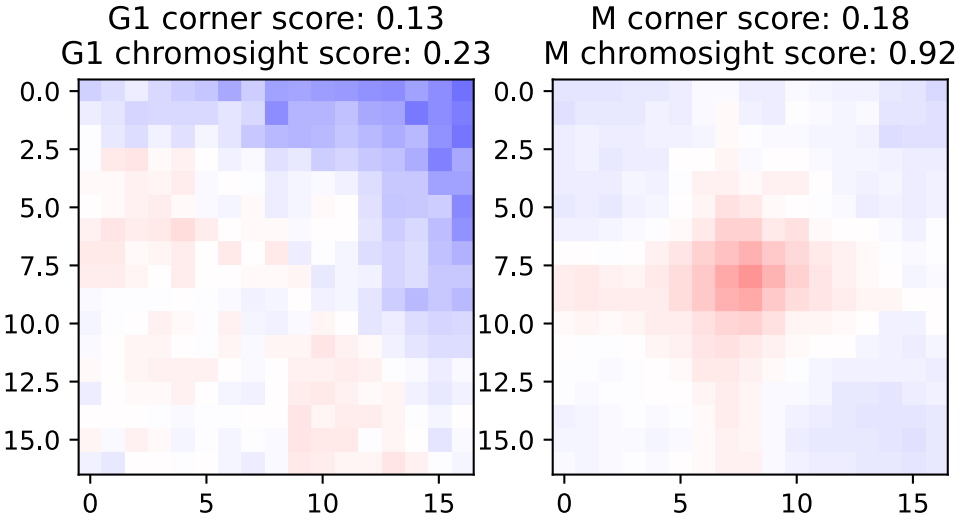
[199]:
```
import scipy.stats as st
loops_df['corner_score'] = [corner_score(m) for m in images]
comp_df = loops_df.loc[
    ~np.isnan(loops_df.corner_score) & ~np.isnan(loops_df.score), :
]
sns.regplot(data=comp_df, x='corner_score', y='score')
plt.title(
    r'Correlation between chromosight and corner score, $\rho$: '
    f'{np.round(st.pearsonr(comp_df.corner_score, comp_df.score)[0], 2)}')
```

[199]:
```
Text(0.5, 1.0, 'Correlation between chromosight and corner score, $\\rho$:
0.71')
```



By computing the pileup (average) of all patterns separately for G1 and M conditions, we can visually appreciate the stronger loop signal in metaphasic cells (M) compared to G1. Computing the chromosight and corner score directly on those pileups shows that the chromosight score makes it easier to discriminate the two conditions. The [-1,1] range is also convenient to interpret results. Note that the chromosight score below is just the pearson coefficient of the pileup with the loop kernel.

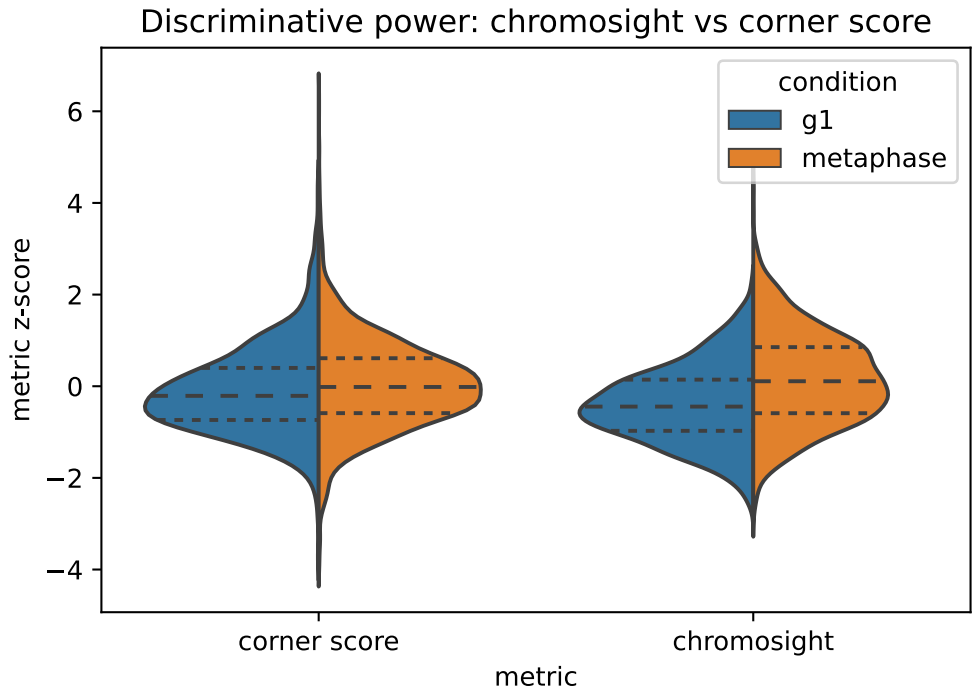
[200]:
```
centroid_g1 = np.apply_along_axis(np.nanmean, 0, images[loops_df.condition ==␣
 ↪'g1'])
centroid_m = np.apply_along_axis(np.nanmean, 0, images[loops_df.condition ==␣
 ↪'metaphase'])

fig, ax = plt.subplots(1, 2)
ax[0].imshow(np.log(centroid_g1), **pileup_kw)
ax[0].set_title(
    f'G1 corner score: {corner_score(centroid_g1):.2f}\n'
    f'G1 chromosight score: {np.round(st.pearsonr(centroid_g1.flat, kernel.
 ↪flat)[0], 2)}'
)
ax[1].imshow(np.log(centroid_m), **pileup_kw)
ax[1].set_title(
    f'M corner score: {corner_score(centroid_m):.2f}\n'
    f'M chromosight score: {st.pearsonr(centroid_m.flat, kernel.flat)[0]:.2f}'
)
plt.show()
```



Instead of summarizing the 2 conditions using only pileups, we can compare the ability of both score to separate the G1 and metaphasic cells based on the distribution of all patterns. Note that both scores are z-transformed to make their ranges comparable.

[201]:
```
corner = comp_df.drop('score', axis=1).rename(columns={'corner_score': 'score'})
corner['metric'] = 'corner score'
corner['score'] = st.zscore(corner['score'])
chromo = comp_df.drop('corner_score', axis=1)
chromo['metric'] = 'chromosight'
chromo['score'] = st.zscore(chromo['score'])
comp_scores = pd.concat([corner, chromo]).reset_index(drop=True)
sns.violinplot(data=comp_scores, x='metric', y='score', split=True,␣
 ↪hue='condition', inner='quartile')
plt.ylabel('metric z-score')
plt.title('Discriminative power: chromosight vs corner score')
```

[201]: `Text(0.5, 1.0, 'Discriminative power: chromosight vs corner score')`



Comparison of loop footprints

For vizualisation purposes, each window can be summarized to a 1D band representing the sum of columns or rows. Here, we compute both the average of rows and columns, and use the element-wise average of both 1D vectors. This gives a good approximation of a 'loop footprint' and is convenient for visualisation.

Each image is centered to its mean to homogenize the overall contact counts in windows. This avoid having globally darker or lighter images and emphasizes relative contrasts within the im-

ages.

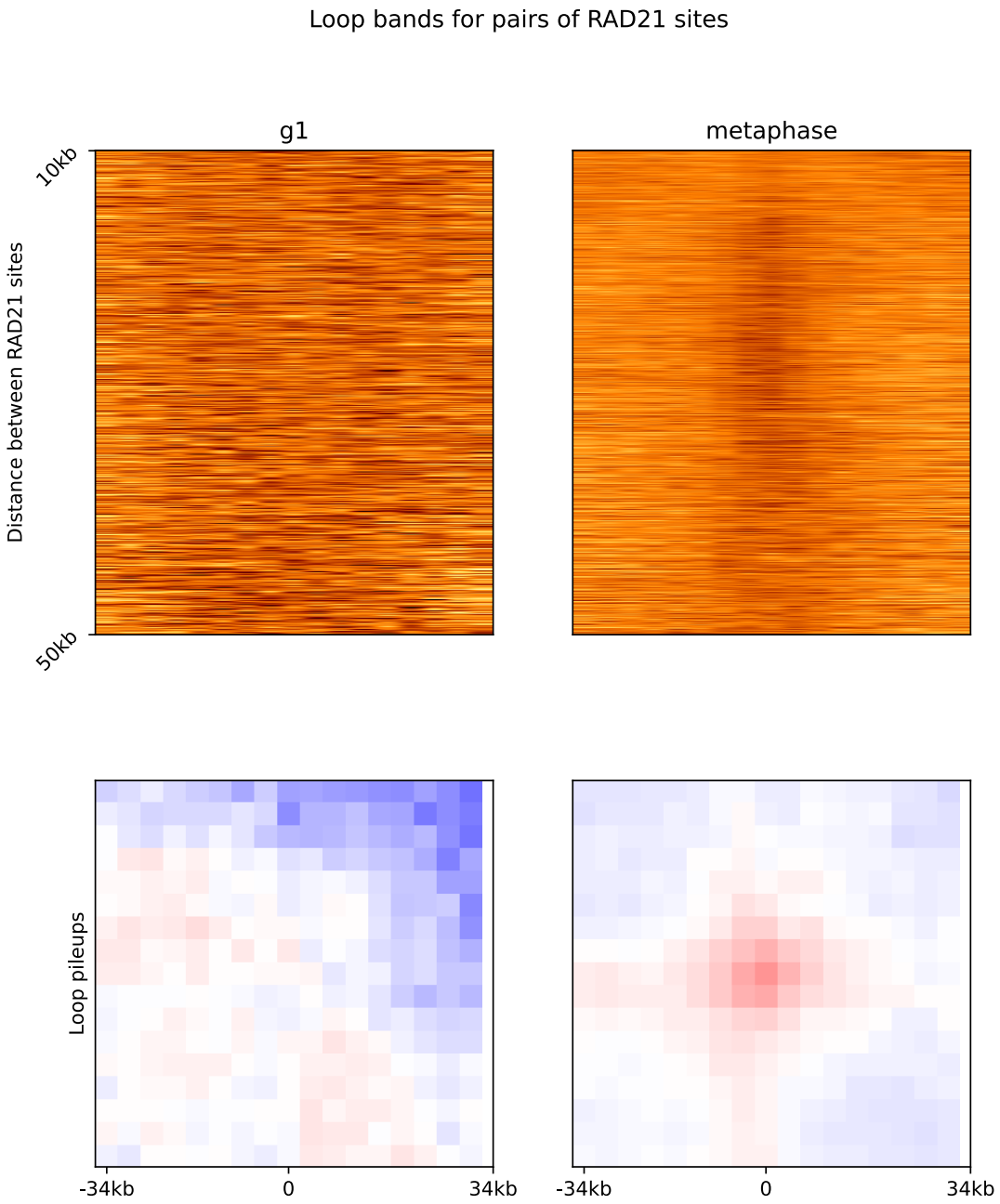
Bands are then sorted by loop size (i.e. distance between anchors) and plotted as a stack from shortest to longest distance interactions.

[ ]:
```
# Center images by subtracting their mean
centered = images.copy()
for img in range(centered.shape[0]):
    centered[img] -= np.nanmean(centered[img])

# Summarise each image by taking the average of its row and col sums.
bands = (np.nansum(centered, axis=1) + np.nansum(centered, axis=2)) / 2


# Reorder bands by distance between anchors
sort_var = 'loop_size'
sorted_bands = bands[np.argsort(loops_df[sort_var]), :]
sorted_cond = loops_df.condition.iloc[np.argsort(loops_df[sort_var])]
sorted_centered = centered[np.argsort(loops_df[sort_var])]

# Define a subset to visualise (too many images so see them all at once)
#smallest_group = np.min(np.unique(sorted_cond, return_counts=True)[1])-1
#smallest_group = 500

# Define saturation threshold for the colormaps
vmax_bands = np.percentile(bands, 99.9)
vmax_img = np.percentile(centered, 99)
```

[202]:
```
fig, axes = plt.subplots(2, 2, figsize=(8, 10))
for i, cond in enumerate(['g1', 'metaphase']):
    axes[0, i].imshow(
        sorted_bands[sorted_cond == cond, :],
        cmap='afmhot_r',
        vmax=vmax_bands,
    )
    axes[0, i].set_title(cond)
    # Compute pileup by averaging all windows for each condition
    centroid = np.apply_along_axis(
        np.nanmean,
        0,
        images[loops_df.condition == cond],
    )
    axes[1, i].imshow(np.log(centroid), **pileup_kw)
    axes[0,i].set_aspect('auto')
    # The rest is just to improve figure aesthetics
    axes[0, i].set_xticks([])
    axes[1, i].set_yticks([])
    if i > 0:
```

```
        axes[0, i].set_yticks([])
    else:
        #axes[0, i].set_yticklabels([], ["10kb", "25kb", "50kb"])
        axes[0, i].set_yticks(
            [0, sorted_bands[sorted_cond == cond, :].shape[0]]
        )
        axes[0, i].set_yticklabels(
            ['10kb', '50kb'],
            minor=False,
            rotation=45
        )

    axes[1, i].set_xticks([0, centroid.shape[0] // 2, centroid.shape[0]])
    half_w = int((res * centroid.shape[0] // 2) / 1000)
    half_w_bp = int(half_w * res / 1000)
    axes[1, i].set_xticklabels([f"{-half_w_bp}kb", "0", f"{half_w_bp}kb"])
    #axes[1, i].set_title(f"corner score: {np.round(corner_score(centroid), 2)}")

axes[0, 0].set_ylabel('Distance between RAD21 sites')
axes[1, 0].set_ylabel('Loop pileups')
plt.suptitle(f'Loop bands for pairs of RAD21 sites')
#plt.savefig('figs/bands_pileup_prots.svg')
```

[202]: Text(0.5, 0.98, 'Loop bands for pairs of RAD21 sites')



Note: Generating a BED2D file

ChIP-seq peaks are often stored as BED files, containing genomic intervals where DNA-binding proteins are enriched. Such files can be used to generate a BED2D file for `chromosight quantify`. This is done by generating all possible 2-ways combinations of peaks that follow desired criteria. In the example below, we use bedtools and awk to generate all intrachromosomal combinations

where peaks are separated by more than 10kb and less than 50kb.

```
MINDIST=10000
MAXDIST=50000
bedtools window -a input/scer_cohesin_peaks.bed \
                -b input/scer_cohesin_peaks.bed \
                -w $MAXDIST \
    | awk -vmd=$MINDIST '$1 == $4 && ($5 - $2) >= md {print}' \
    | sort -k1,1 -k2,2n -k4,4 -k5,5n \
    > input/scer_cohesin_peaks.bed2d
```

B.2 Output visualisation

Plotting Chromosight's output

August 27, 2021

Chromosight generates tabular text files with loops coordinates and scores. This file can be loaded into your favorite scripting language for visualization. For the purpose of this demonstration, we show how to plot the contact maps with detected coordinates using python, pandas and cooler.

The data shown here was generated with the following commands:

```
chromosight detect data_test/example.cool -m8000 -M50000 -p0.35 detect/example_loops
chromosight detect data_test/example.cool --pattern borders detect/example_borders
chromosight detect data_test/example.cool --pattern hairpins detect/example_hairpins
```

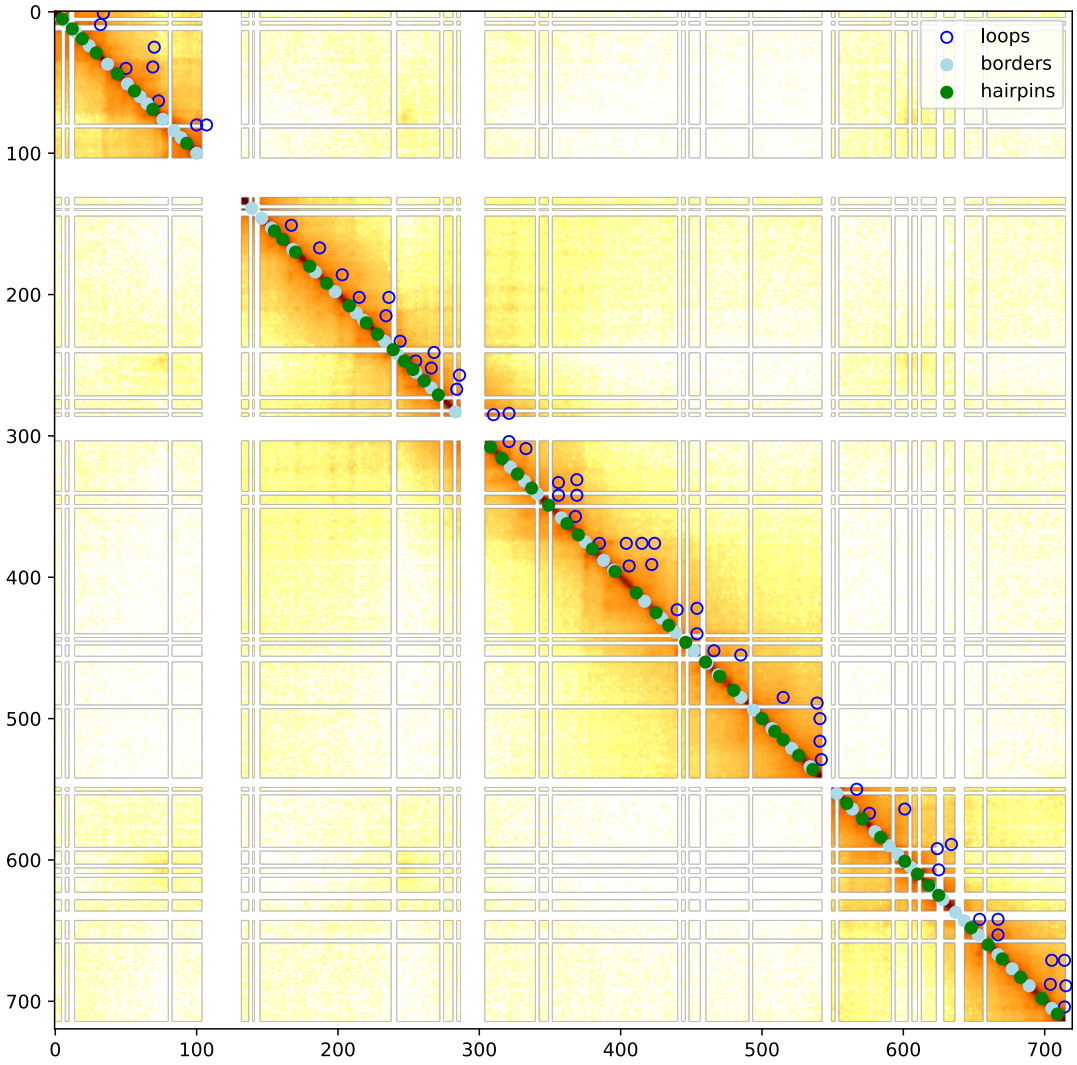
Which will detect all loops of size 8-50kb in `example.cool` and filter those with a score above 0.35. The output files will be located in the `detect/` folder.

[35]:
```
%config InlineBackend.figure_format = 'svg'
import re
import json
import numpy as np
import pandas as pd
import cooler
import matplotlib.pyplot as plt
import chromosight.utils.detection as cud

# Load detected patterns' tables
loops = pd.read_csv('detect/example_loops.tsv', sep='\t')
borders = pd.read_csv('detect/example_borders.tsv', sep='\t')
hairpins = pd.read_csv('detect/example_hairpins.tsv', sep='\t')

# Load Hi-C data in cool format
c = cooler.Cooler("../../data_test/example.cool")
```

View the whole genome matrix

To plot the whole matrix with patterns, the matrix is extracted from the cool file and columns bin1 and bin2 are used. Those columns contain the genome-wide bin number of pattern coordinates, and matches the whole genome matrix. Plotting the whole genome is straightforward, but likely to take too much memory for larger genomes.

[36]:
```
%matplotlib inline
# Plot the whole matrix
plt.figure(figsize=(10, 10))
mat = c.matrix(sparse=False, balance=True)[:]
plt.imshow(mat ** 0.2, cmap='afmhot_r')
plt.scatter(loops.bin2, loops.bin1, edgecolors='blue', facecolors='none',␣
 ↪label='loops')
plt.scatter(borders.bin2, borders.bin1, c='lightblue', label='borders')
plt.scatter(hairpins.bin2, hairpins.bin1, c='green', label='hairpins')
plt.legend()
plt.show()
```



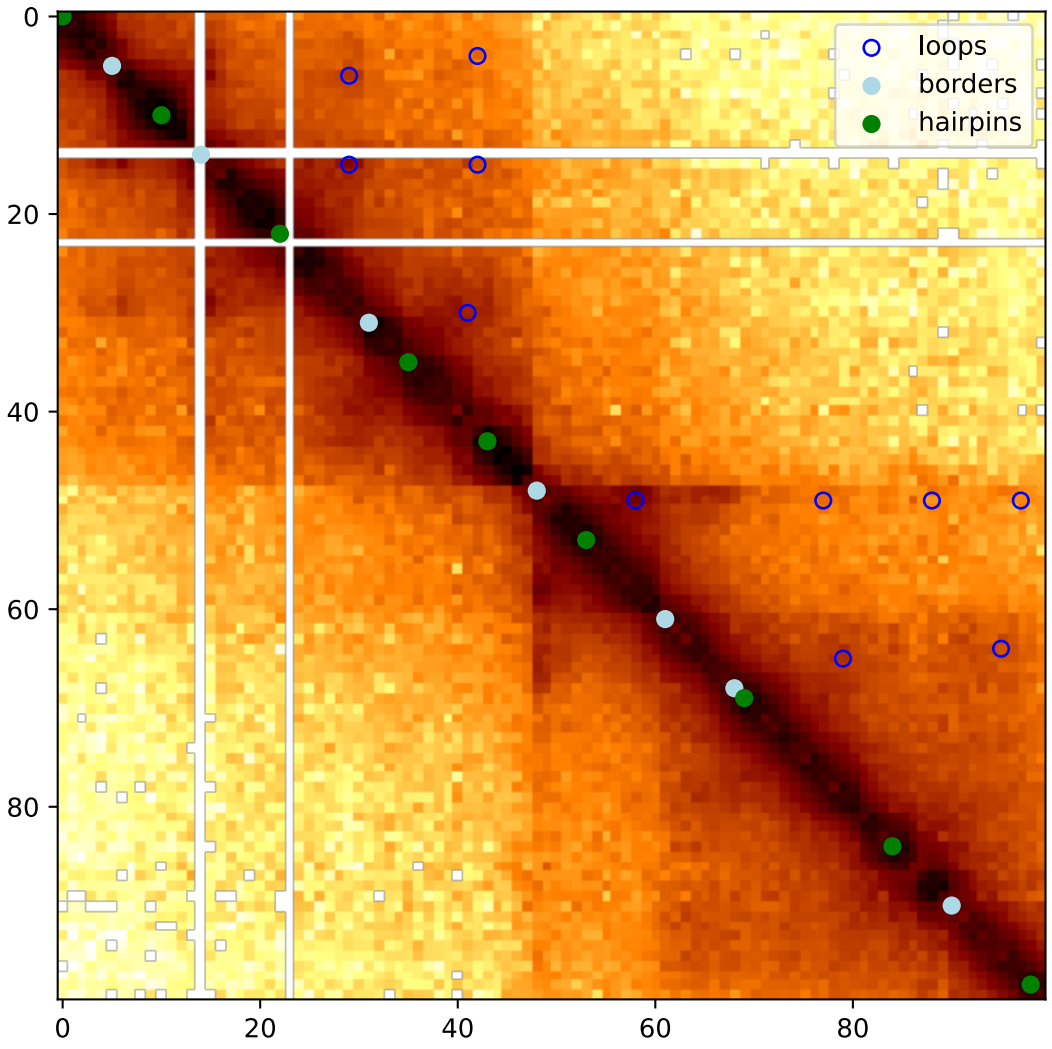
View a matrix region

To reduce the amount of memory required, we can define a region of interest. The corresponding matrix region can be fetched from the cool file using cooler, and patterns falling within that region can be filtered using pandas. Since we want to overlay the patterns on top of the region matrix, the bin1 and bin2 columns should be adjusted to be relative to the region's start instead of the genome.

[ ]:
```
def subset_region(df, region):
    """
    Given a pattern dataframe and UCSC region string, retrieve only patterns in␣
 ↪that region.
    """
    # Split the region string at each occurence of - or : (yields 3 elements)
    chrom, start, end = re.split('[-:]', region)
    start, end = int(start), int(end)
    # Only keep patterns on the same chromosome as the region and
    # within the start-end interval
    subset = df.loc[
        (df.chrom1 == chrom) &
        (df.chrom2 == chrom) &
        (df.start1 >= start) &
        (df.start2 >= start) &
        (df.end1 < end) &
        (df.end2 < end), :
    ]
    return subset
```

[37]:
```
# Select a region of interest
region = 'chr2:200000-300000'
mat = c.matrix(sparse=False, balance=True).fetch(region)


loops_sub = subset_region(loops, region)
borders_sub = subset_region(borders, region)
hairpins_sub = subset_region(hairpins, region)

# Make genome-based bin numbers relative to the region
for df in [loops_sub, borders_sub, hairpins_sub]:
    df.bin1 -= c.extent(region)[0]
    df.bin2 -= c.extent(region)[0]
```

[38]:
```
%matplotlib inline
plt.figure(figsize=(7, 7))
plt.imshow(np.log10(mat), cmap='afmhot_r')
plt.scatter(loops_sub.bin2, loops_sub.bin1, edgecolors='blue',␣
 ↪facecolors='none', label='loops')
plt.scatter(borders_sub.bin2, borders_sub.bin1, c='lightblue', label='borders')
plt.scatter(hairpins_sub.bin2, hairpins_sub.bin1, c='green', label='hairpins')
plt.legend()
plt.show()
```



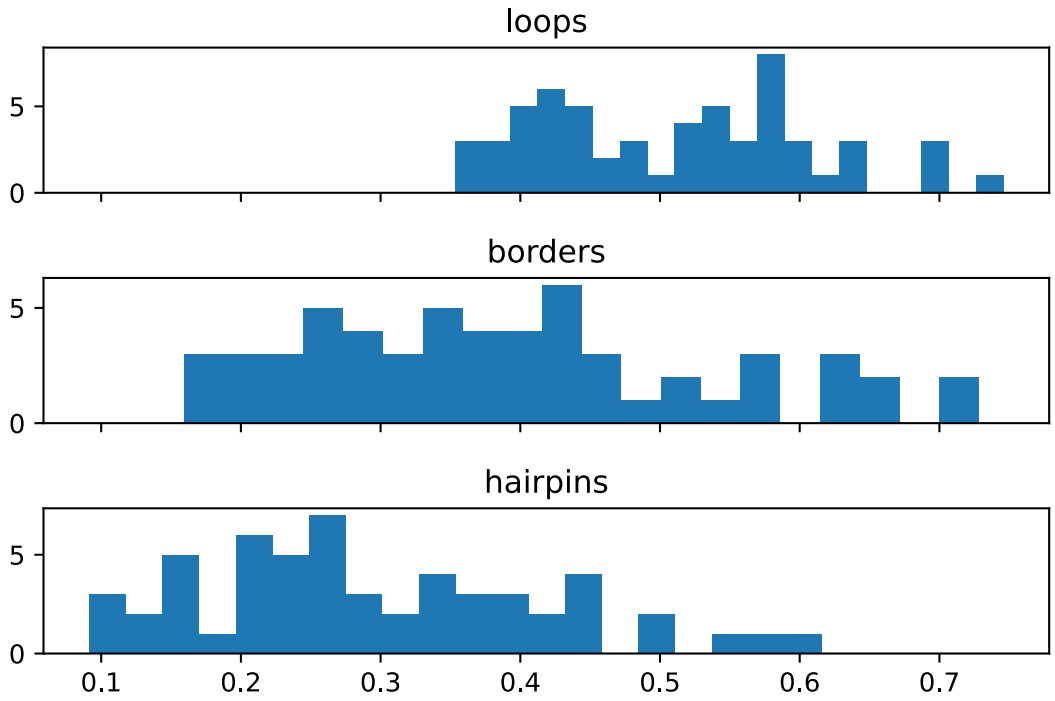
Plot the distribution of scores

Scores of detected patterns are provided as Pearson correlation coefficient with the template and are stored in the 'score' column of the tabular output. Their distribution can be viewed with regular histogram functions. Since we use a threshold for detection (the `--pearson` option in the command line interface), the score lower end of the distribution will be truncated at this threshold.

Different patterns will have different score distributions and default thresholds.

[39]:
```
%matplotlib inline
plt.figure(figsize=(8, 8))
fig, ax = plt.subplots(3, 1, sharex=True)
for i, (df, pat) in enumerate(zip([loops, borders, hairpins], ['loops',␣
 ↪'borders', 'hairpins'])):
    ax[i].hist(df.score, 20)
    ax[i].set_title(pat)
plt.tight_layout()
```

```
<Figure size 576x576 with 0 Axes>
```



Looking at detected patterns

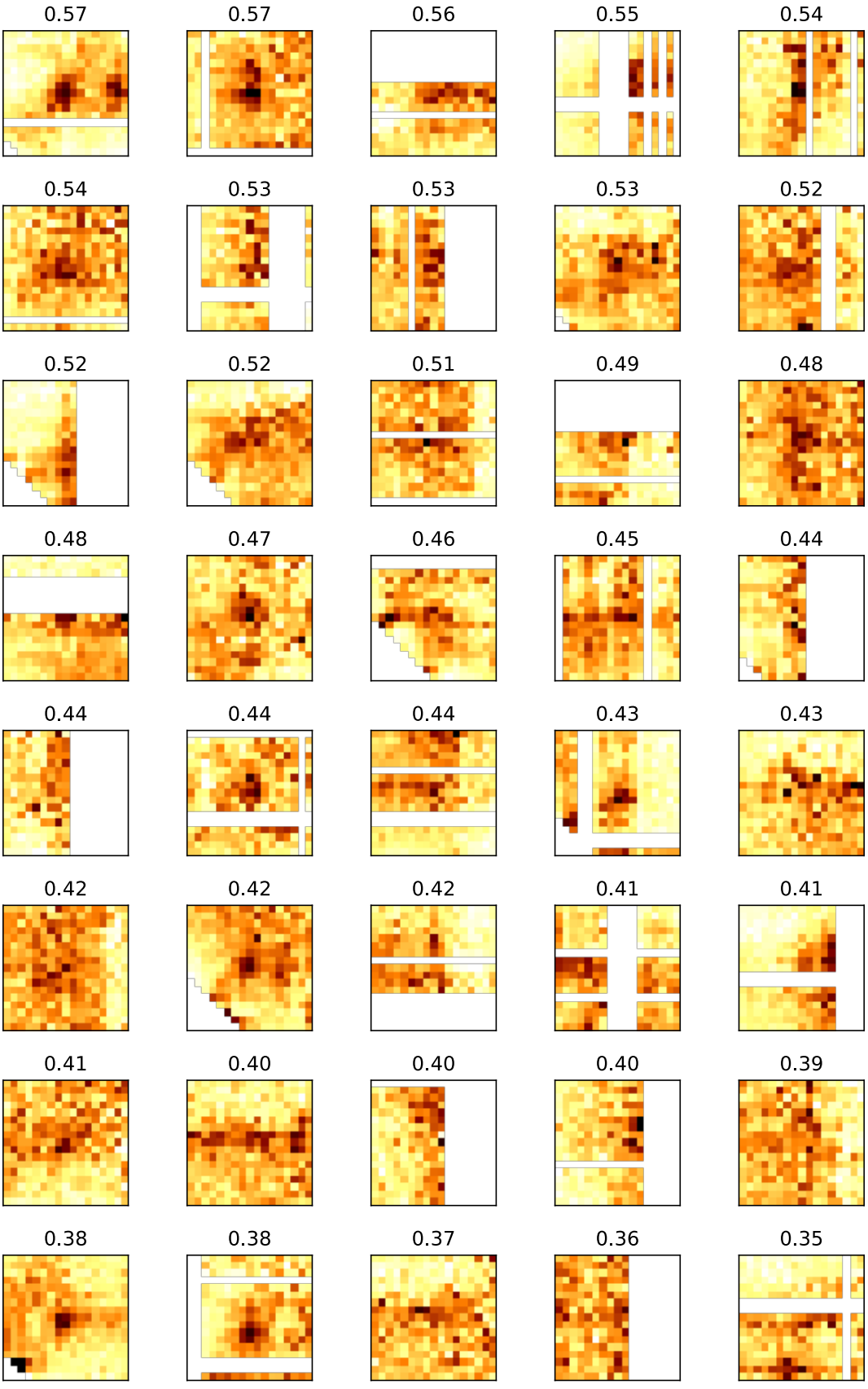
Windows around detected patterns in the processed matrix are stored in the JSON / npy file when running chromosight's detect or quantify commands. These windows are in the same order as the coordinates in the output table.

[40]:
```
# Load input json file into a dictionary
loop_wins = json.load(open('detect/example_loops.json', 'r'))
# Note that keys are string, as required by the JSON format,
# so we convert them to int() for convenience
loop_wins = {int(i): np.array(w) for i, w in loop_wins.items()}
# Make an empty 3D array of shape N_coords x height x width
wins = np.zeros((len(loop_wins.items()), *loop_wins[0].shape))
# Fill the 3D array with windows values
for i, w in loop_wins.items(): wins[i] = w
```

For example, we can plot the best 40 windows around detected loops ordered by score:

[41]:
```
%matplotlib inline
plt.figure(figsize=(10, 10))

fig, ax = plt.subplots(8, 5, figsize=(8, 12))

for i, n in enumerate(np.argsort(loops.score)[39::-1]):
    m, s = np.nanmean(loop_wins[n]), np.nanstd(loop_wins[n])
    ax.flat[i].imshow((loop_wins[n] - m) / s, cmap='afmhot_r', vmax=4)
    ax.flat[i].set_title(f'{loops.score[n]:.2f}')
    ax.flat[i].set_xticks([])
    ax.flat[i].set_yticks([])
plt.tight_layout()
```

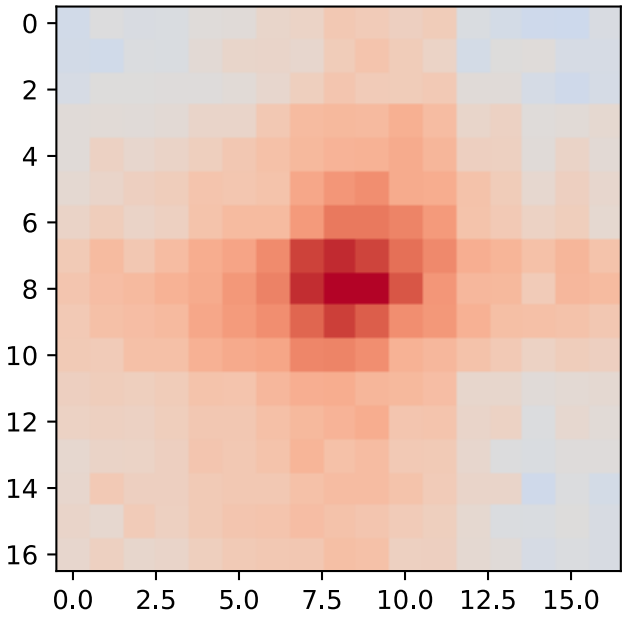
```
<Figure size 720x720 with 0 Axes>
```


0.57
0.57
0.56
0.55
0.54
0.54
0.53
0.53
0.53
0.52
0.52
0.52
0.51
0.49
0.48
0.48
0.47
0.46
0.45
0.44
0.44
0.44
0.44
0.43
0.43
0.42
0.42
0.42
0.41
0.41
0.41
0.40
0.40
0.40
0.39
0.38
0.38
0.37
0.36
0.35

The pileup can also be re-computed from these windows using chromosight's helper function. This is useful to plot the pileup for a subset of the detected patterns, or just to generate the pileup plot with different aesthetics.

[42]:
```
%matplotlib inline
plt.figure(figsize=(4, 4))
pileup = cud.pileup_patterns(wins)
plt.imshow(pileup, cmap='coolwarm', vmax=1.8, vmin=0)
```

[42]:
```
<matplotlib.image.AxesImage at 0x7f571dcb6f50>
```



C Walkthrough of Pareidolia's algorithm

Change detection in pattern intensities

September 9, 2021

This notebook details the inner working of Pareidolia and walks through each step with code and visualization. We show the intermediate steps, parameters involved and use a small region of mouse chromosome 14 as an example to visualize the results.

Throughout the notebook, we call internal functions of pareidolia and chromosight to show the transformations going on inside the program. In a normal use case, however, this whole process is not neccessary and the user can simply execute the program as explained in the documentation, either through the command line interface or the python API.

Background

We compare several samples issued from 2 different timepoints (t_0, t_1). Multiple samples (replicates) $(r_1, r_2, ..., r_R)$ can share the same timepoint. Each sample has a matrix $M_{r,t}$ where $M_{r,t}[i, j]$ is a Pearson correlation coefficient with a kernel K representing the pattern of interest. If the kernel was of size $K_m x K_n$, the correlation coefficient was computed as:

$$M_{r,t}[i, j] = Corr(K, H_{r,t}[i - \frac{K_m}{2} : i + \frac{K_m}{2}, j - \frac{K_n}{2} : j + \frac{K_n}{2}])$$

Where $H_{r,t}$ is the Hi-C matrix of the sample.

The method is inspired by median filtering-based background formation. We start by generating a background matrix for each condition (timepoint), whose values are defined as the median of all replicates in that condition:

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

Next, we compute global background matrix, defined as:

$$B = median(B_1, B_2, ..., B_T)$$

Finally we extract the change as the difference between each condition 's background.

$$D = B_{t_1} - B_{t_0}$$

Changes can then filtered on various criteria to extract patches of strong differences

[486]:
```
from typing import Iterable, Optional, Tuple, Iterator, Set
from skimage.filters import threshold_otsu
import matplotlib.pyplot as plt
import scipy.sparse as sp
import chromosight.kernels as ck
import chromosight.utils.preprocessing as cup
import chromosight.utils.detection as cud
import pareidolia.detection as pad
import pareidolia.preprocess as pap
import pareidolia.hic_utils as pah
import numpy as np
import pandas as pd
import cooler
RES = 40000
region = 'chr14:12900000-14000000'
```

Example data

In this case, we are using mice bone macrophage infected by a bacterium at different timepoints. Each timepoint includes several replicates. We will be comparing infected samples to non-infected samples.

[487]:
```
def get_cool(sample: str) -> cooler.Cooler:
    """Given a sample name, load the corresponding cool file."""
    cool_path = f'./data/output/cool/{sample}.mcool::/resolutions/{RES}'
    try:
        cool = cooler.Cooler(cool_path)
    except OSError:
        cool = None
    return cool


# Load references to cool files from each sample into the dataframe
samples = pd.read_csv('samples.tsv', sep='\t', usecols=[0, 4])
samples["cool"] = samples.library.apply(get_cool)
samples['cond'] = 'control'
samples.loc[samples.infection_time > 0, 'cond'] = 'treat'
# Remove samples without Hi-C data
samples = samples.loc[~samples.cool.isnull(), :]
samples = samples.set_index('library')
samples.head(20)
```

[487]:

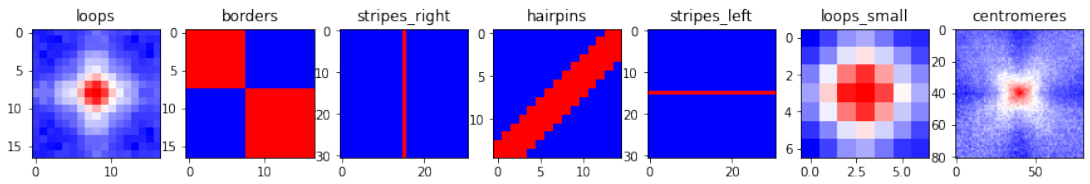
library	infection_time	cool	cond
PM51	0	<Cooler "PM51.mcool::/resolutions/40000">	control
PM52	20	<Cooler "PM52.mcool::/resolutions/40000">	treat
PM53	20	<Cooler "PM53.mcool::/resolutions/40000">	treat
PM54	2	<Cooler "PM54.mcool::/resolutions/40000">	treat
PM55	2	<Cooler "PM55.mcool::/resolutions/40000">	treat
PM121	0	<Cooler "PM121.mcool::/resolutions/40000">	control
PM122	20	<Cooler "PM122.mcool::/resolutions/40000">	treat
PM123	20	<Cooler "PM123.mcool::/resolutions/40000">	treat
PM124	2	<Cooler "PM124.mcool::/resolutions/40000">	treat
PM125	2	<Cooler "PM125.mcool::/resolutions/40000">	treat
PM126	20	<Cooler "PM126.mcool::/resolutions/40000">	treat
PM127	20	<Cooler "PM127.mcool::/resolutions/40000">	treat

Preprocessing matrices We first need to convert Hi-C matrices into convolution maps for a pattern of interest (e.g. loops). We also need to make sure we are comparing the same positions between maps by keeping the same nonzero values in sparse matrices. We also work on a subregion of the matrices to make computations faster.

Below, we perform the steps which are done internally in Pareidolia to explore the intermediates used to compute and filter changes. The examples are visualized on a small region to make visualizations more clear.

Pareidolia measures changes relative to the a pattern of interest, represented by the kernel K. Default chromosight kernels shown below can be used by providing the kernel name. Alternatively, a user defined matrix can be provided.

[488]:
```
%matplotlib inline

fig, ax = plt.subplots(1, len(ck.kernel_names), figsize=(15, 20))
for a, name in zip(ax, ck.kernel_names):
    kernel_mat = np.array(getattr(ck, name)['kernels'][0])
    a.imshow(np.log1p(kernel_mat), cmap='bwr')
    a.set_title(name)
```

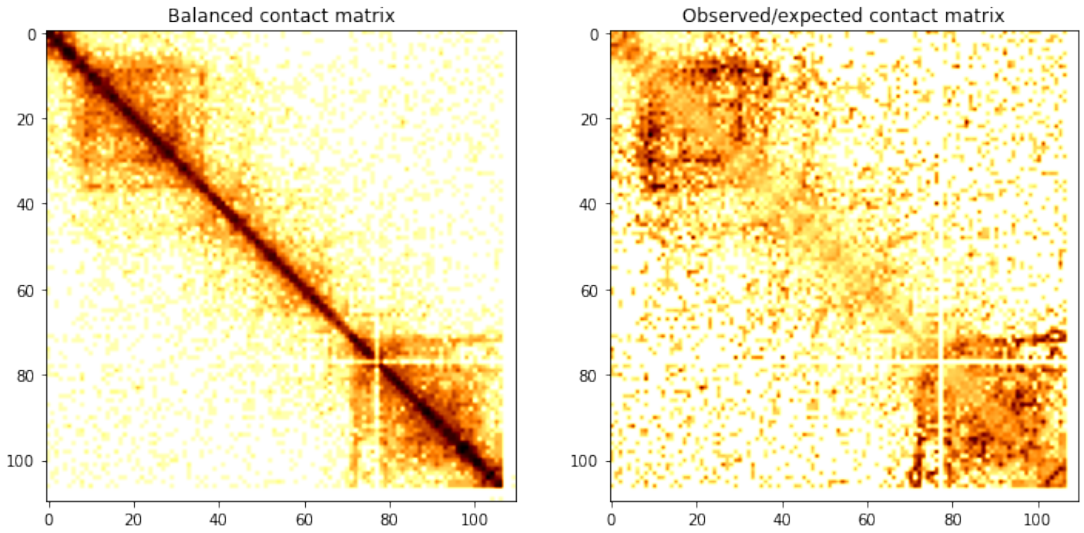


The first step is to preprocess Hi-C matrices. We work on sparse matrices to reduce memory usage. In order for samples to be comparable, they must have the same sparsity structure (the positions of explicitely stored values).

First, we subsample contacts in all matrices to ensure they have the same coverage (i.e. the coverage of the lowest sample). We also detrend the matrix for the distance-contact decay gradient. The resulting matrix (also called observed/expected) is computed by dividing each value by the average of its diagonal.

[489]:
```
%matplotlib inline
region = 'chr14:21600000-26000000'
# Compute lowest amount of contacts among all matrices
min_contacts = pah.get_min_contacts(samples.cool, region=region)

# Subsample, preprocess and subset matrices
samples['mat'] = samples.cool.apply(
    lambda c: pah.preprocess_hic(c, min_contacts=min_contacts, region=region)
)
fig, ax = plt.subplots(1, 2, figsize=(12, 8))
ax[0].imshow(np.log1p(samples.cool['PM51'].matrix(sparse=False, balance=False).
 ↪fetch(region)), cmap='afmhot_r')
ax[0].set_title('Balanced contact matrix')
ax[1].set_title('Observed/expected contact matrix')
ax[1].imshow(np.log1p(samples.mat['PM51'].toarray()), cmap='afmhot_r')
```

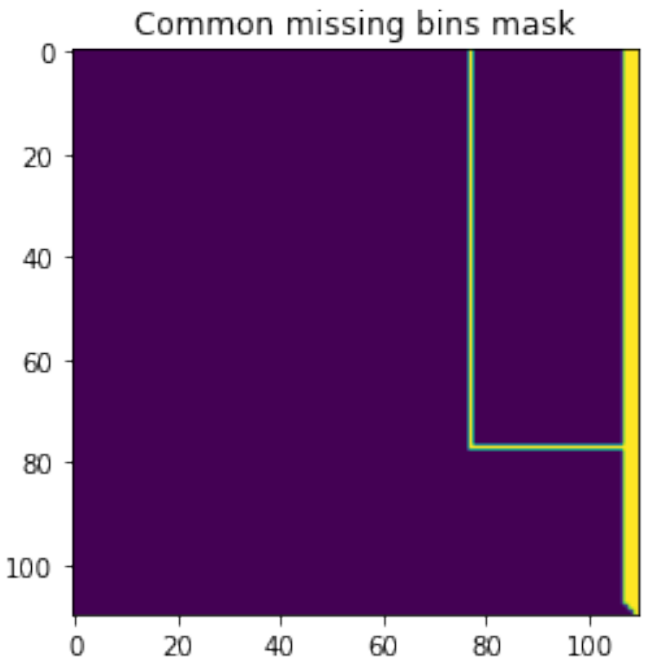
[489]:
```
<matplotlib.image.AxesImage at 0x7f86369de890>
```



Then, we compute the missing bins in each sample (genomic bins of low coverage) and take the union of those bins across samples. The resulting mask is applied on all samples.

[490]:
```
%matplotlib inline

# Get bins valid in all matrices
common_valid = pap.get_common_valid_bins(samples.mat)

# Generate mask to remove all bins missing in any matrix
common_mask = cup.make_missing_mask(
    samples.mat[0].shape,
    common_valid,
    common_valid,
    max_dist=None,
    sym_upper=True
)

# Make sure missing bins are set to 0 in all matrices and discard lower triangle
samples.mat = samples.mat.apply(lambda m: cup.erase_missing(sp.triu(m),␣
 ↪common_valid, common_valid))


plt.imshow(common_mask.toarray())
plt.title("Common missing bins mask")
```

[490]:
```
Text(0.5, 1.0, 'Common missing bins mask')
```



Each sample's preprocessed matrix is fed to Chromosight's convolution engine to return a matrix

of identical dimension whose values are the correlation coefficient with the kernel matrix at each position.

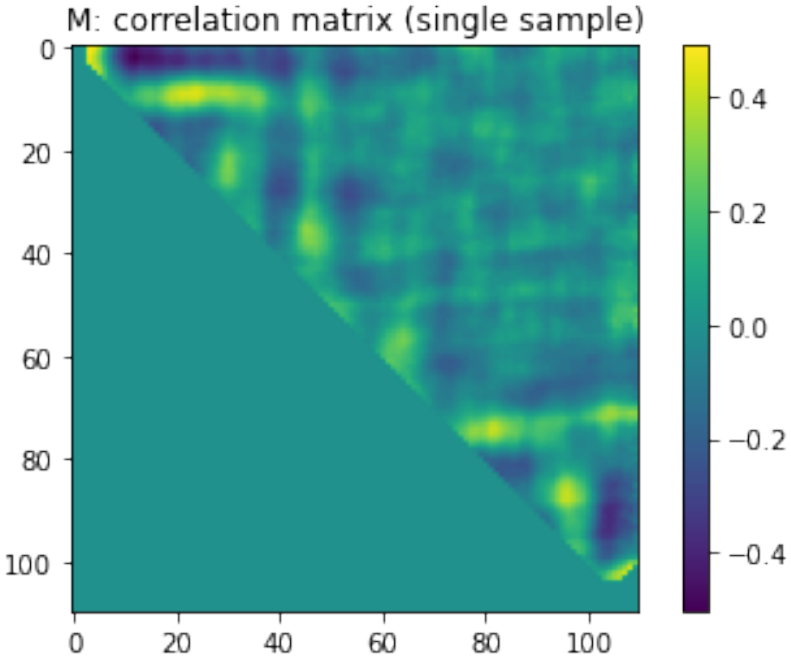
[491]:
```
# Compute correlation maps, using the same mask for all samples
# Note normxcorr2 returns a map of correlations, and one of p-values
# (None by default). Hence the slicing of the output
samples['corr'] = samples.mat.apply(lambda m: cud.normxcorr2(
    m,
    ck.loops['kernels'][0],
    full=True,
    missing_mask=common_mask,
    sym_upper=True
)[0])
```

Then, we take the union of nonzero positions across all samples' correlation matrices and enforce explicit storage of those positions across samples. This will ensure all samples have the same sparsity, so that we can perform operations directly on the aligned nonzero values.

[492]:
```
# First option: Keep values that are present in any matrix to zero
total_nnz_set = pap.get_nnz_union(samples['corr'])
samples['corr_union'] = samples['corr'].apply(lambda c: pap.fill_nnz(c,␣
 ↪total_nnz_set))
```

[493]:
```
%matplotlib inline
plt.imshow(samples.corr_union['PM51'].toarray())
plt.title("M: correlation matrix (single sample)")
plt.colorbar()
```

[493]:
```
<matplotlib.colorbar.Colorbar at 0x7f863684c350>
```



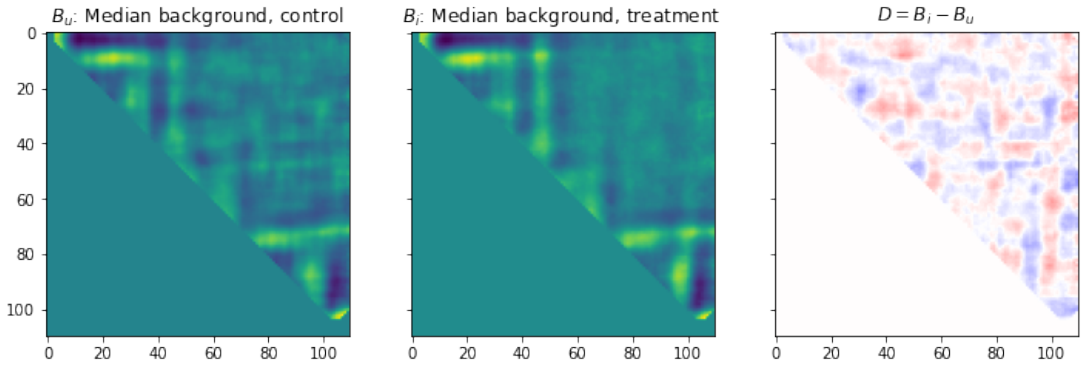
[494]:
```
# Take one infected (t=20h) and one uninfected matrix. Visualise the␣
 ↪median-filter
# based background accumulation
u_union = samples['corr_union'].loc[samples.cond == 'control'].values
i_union = samples['corr_union'].loc[samples.cond == 'treat'].values
bg_uni = pad.median_bg(u_union)
bg_inf = pad.median_bg(i_union)
```

[495]:
```
bg_inf
```

[495]:
```
<110x110 sparse matrix of type '<class 'numpy.float64'>'
        with 6067 stored elements in Compressed Sparse Row format>
```

[496]:
```
%matplotlib inline
diff = bg_inf.toarray() - bg_uni.toarray()

f, ax = plt.subplots(1, 3, sharex=True, sharey=True, figsize=(12, 4))
ax[0].imshow(bg_uni.toarray(), cmap='viridis')
ax[0].set_title("$B_u$: Median background, control")
ax[1].imshow(bg_inf.toarray(), cmap='viridis')
ax[1].set_title("$B_i$: Median background, treatment")
ax[2].imshow(diff, cmap='seismic', vmin=-1, vmax=1)
ax[2].set_title("$D= B_i - B_u$")
```

[496]: `Text(0.5, 1.0, '$D= B_i - B_u$')`



Filtering changes

We can clearly identify regions that change between conditions, however this is only based on the median and does not account for variability. It would be better to penalize regions with technical variability.

Pareidolia uses a combination of 3 filters to select region with differential pattern intensity.

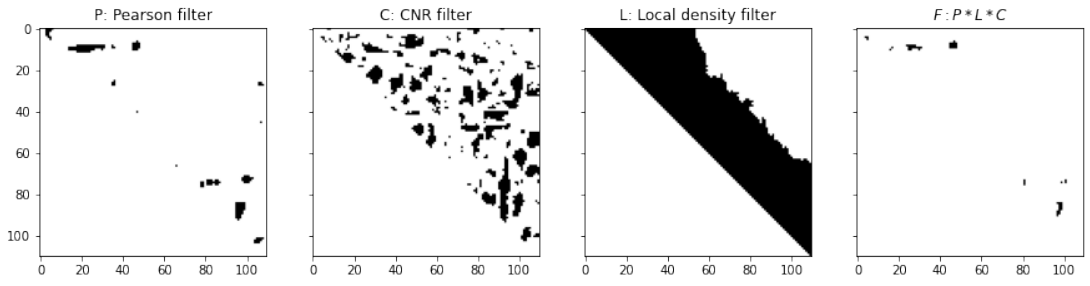
- Pearson score: At least 1 sample must have a pearson score (pattern similarity) above T_p.
- Local contact density: All samples must have a local contact density (nonzero contacts in surrounding window) above T_d
- Contrast-to-noise ratio: The contrast-to-noise-ratio $\frac{D}{\sigma}$ with sigma being within-condition position-wise standard errors must be above T_c

[497]:
```
Td, Tp, Tc = 0.2, 0.35, 0.2

pearson_fail = [
    (m.data < Tp).astype(bool) for m in samples["corr_union"]
]
pearson_fail = np.bitwise_and.reduce(pearson_fail)


# Get thresholded pearson matrix
ex = samples['corr_union']['PM51'].copy().tocoo()
P = sp.coo_matrix(
    ([True for _ in range(sum(~pearson_fail))], (ex.row[~pearson_fail], ex.
 ↪col[~pearson_fail])),
    shape=ex.shape,
    dtype=bool)

# Get thresholded CNR matrix
```

```
_, C = pah._median_bg_subtraction(
    pd.DataFrame(
        {
            'cond': ['ctrl' if c else 'treat' for c in samples.cond.values ==␣
 ↪'control'],
            'mat': samples.corr_union
        },
    ),
    control='ctrl',
    cnr_thresh=Tc
)
C.data[C.data < Tc] = 0.0
C.data[C.data > 0] = 1

# Get thresholded density matrix
L = pah.make_density_filter(samples.mat, Td, win_size=17, sym_upper=True)

# Convert everything to dense boolean array for visualization
filt = lambda m: m.toarray().astype(bool)
P, C, L = filt(P), filt(C), filt(L)
```

[504]:
```
%matplotlib inline
fig, ax = plt.subplots(1, 4, figsize=(15, 4), sharex=True, sharey=True)
ax[0].imshow(P, cmap='Greys')
ax[0].set_title('P: Pearson filter')
ax[1].imshow(C, cmap='Greys')
ax[1].set_title('C: CNR filter')
ax[2].imshow(L, cmap='Greys')
ax[2].set_title('L: Local density filter')
ax[3].imshow(P * C * L, cmap='Greys')
ax[3].set_title('$F: P * L * C$')
```

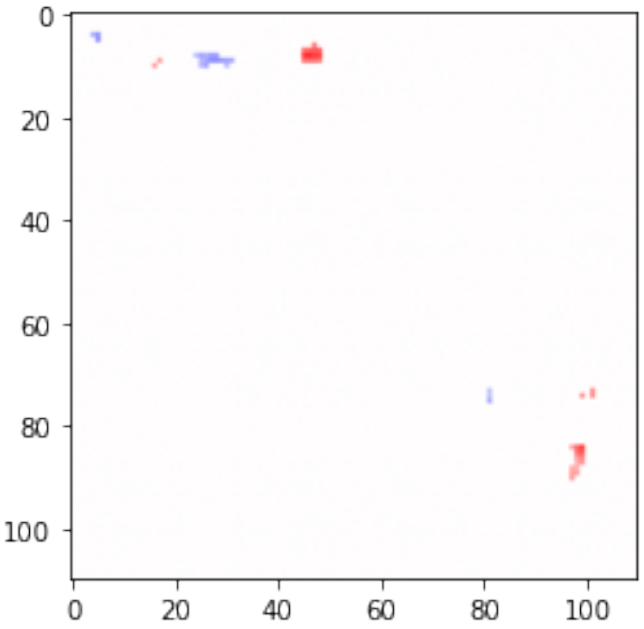
[504]:
```
Text(0.5, 1.0, '$F: P * L * C$')
```



We can then use the resulting filter F to mask the pattern chang matrix.

[505]:
```
%matplotlib inline
diff_f = diff * P * C * L
plt.imshow(diff_f, cmap='seismic', vmin=-.5, vmax=.5)
```

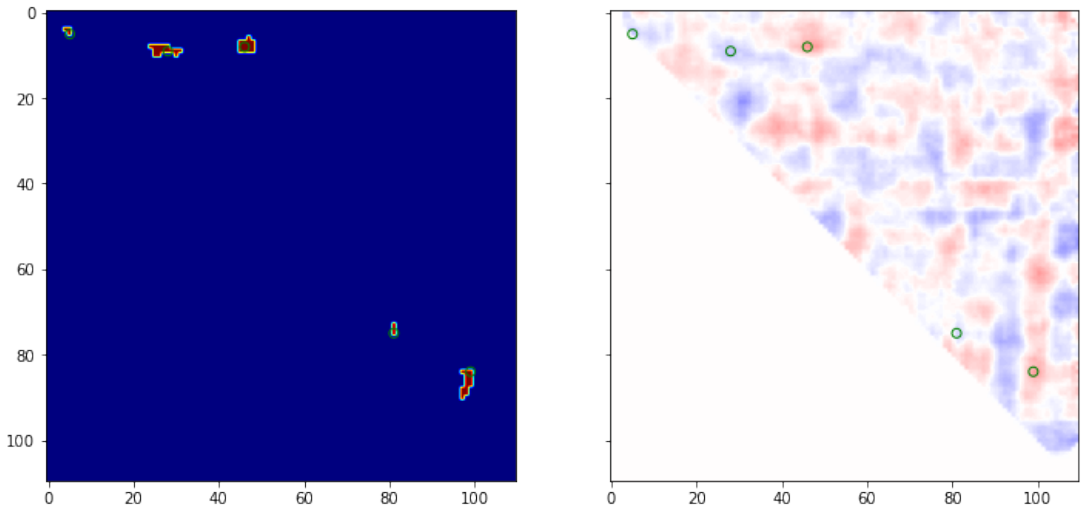
[505]: <matplotlib.image.AxesImage at 0x7f8635e01bd0>



Finally, discrete patches of change (or foci) are extracted from the matrix, and the coordinate of the local maximum of change value in each patch is returned.

[506]:
```
%matplotlib inline
foci, foci_mat = cud.pick_foci(sp.csr_matrix(np.abs(diff_f)), pearson=0,␣
 ↪min_size=3)
foci_mat= foci_mat.toarray().astype(float)
foci_mat[foci_mat != 0] = 10
fig, ax = plt.subplots(1, 2, sharex=True, sharey=True, figsize=(12, 8))
ax[0].imshow(foci_mat, cmap='jet')
ax[0].scatter(foci[:, 1], foci[:, 0], edgecolors='green', facecolors='none')
ax[1].imshow(diff, cmap='seismic', vmax=1, vmin=-1)
ax[1].scatter(foci[:, 1], foci[:, 0], edgecolors='green', facecolors='none')
```

[506]: <matplotlib.collections.PathCollection at 0x7f8635cd4ed0>

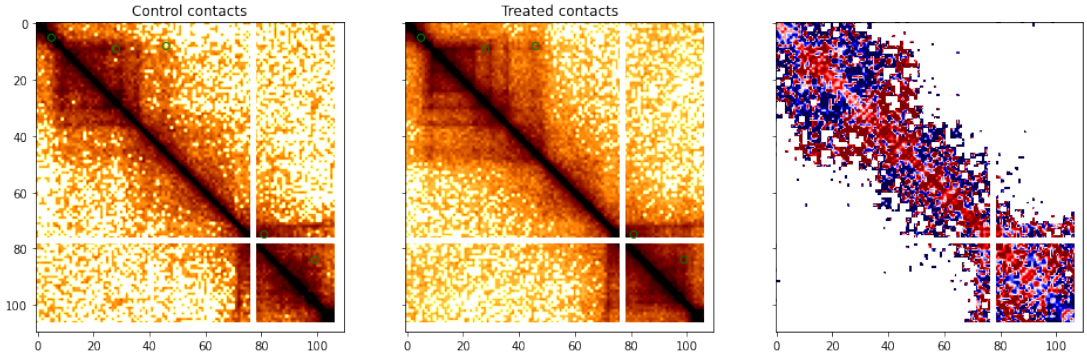


We can now visualize the original contact maps (averaged by condition) with the overlay of detected differential loop positions:

[507]:
```
control_contacts = np.dstack(
    [clr.matrix(balance=True, sparse=False).fetch(region) for clr in samples.
 ↪cool[samples.cond=='control']]
).mean(axis=2)
treatment_contacts = np.dstack(
    [clr.matrix(balance=True, sparse=False).fetch(region) for clr in samples.
 ↪cool[samples.cond=='treat'][:2]]
).mean(axis=2)
```

[508]:
```
%matplotlib inline
fig, ax = plt.subplots(1, 3, sharex=True, sharey=True, figsize=(16, 8))
vmax = np.nanpercentile(control_contacts**0.2, 98)
ax[0].imshow(control_contacts**0.2, cmap='afmhot_r', vmax=vmax)
ax[0].scatter(foci[:, 1], foci[:, 0], edgecolors='green', facecolors='none')
ax[0].set_title('Control contacts')
ax[1].imshow(treatment_contacts**0.2, cmap='afmhot_r', vmax=vmax)
ax[1].set_title('Treated contacts')
ax[1].scatter(foci[:, 1], foci[:, 0], edgecolors='green', facecolors='none')
ax[2].imshow(np.log2(treatment_contacts/control_contacts), cmap='seismic',␣
 ↪vmin=-0.5, vmax=0.5)
```

[508]:
```
<matplotlib.image.AxesImage at 0x7f8635ba4f90>
```


Control contacts
Treated contacts
0
20
40
60
80
100

Bibliography

[1] N. C. Johnson, J. H. Graham, and F. A. Smith. „Functioning of Mycorrhizal Associations along the Mutualism-Parasitism Continuum“. 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?“ In: *Trends in Ecology & Evolution* 21.11 (Nov. 2006), pp. 621–628 (cit. on p. 2).

[3] Jonathan Knight. „Meet the Herod Bug“. 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“. 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“. 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“. 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“. 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“. In: *Nature Reviews Microbiology* 10.1 (Jan. 2012), pp. 13–26 (cit. on p. 3).

[9] Leigh Van Valen. „A NEW EVOLUTIONARY LAW.“ In: *Evolutionary Theory* 1.1 (1973), pp. 1–30 (cit. on p. 3).

[10] Alicia M. Holmgren, Cameron A. McConkey, and Sunny Shin. „Outrunning the Red Queen: Bystander Activation as a Means of Outpacing Innate Immune Subversion by Intracellular Pathogens“. In: *Cellular & Molecular Immunology* 14.1 (Jan. 2017), pp. 14–21 (cit. on p. 3).

[11] Soma Ghosh and Tamara J. O'Connor. „Beyond Paralogs: The Multiple Layers of Redundancy in Bacterial Pathogenesis“. In: *Frontiers in Cellular and Infection Microbiology* 7 (Nov. 2017), p. 467 (cit. on pp. 3, 4).

[12] U. Bergthorsson, D. I. Andersson, and J. R. Roth. „Ohno's Dilemma: Evolution of New Genes under Continuous Selection“. In: *Proceedings of the National Academy of Sciences* 104.43 (Oct. 2007), pp. 17004–17009 (cit. on p. 4).

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

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

[15] 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).

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

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

[18] 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“. In: *Genome Biology* 14.2 (2013), R11 (cit. on pp. 4, 5).

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

[20] 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).

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

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

[23] 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“. In: *Journal of Bacteriology* 187.22 (Nov. 2005), pp. 7716–7726 (cit. on p. 5).

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

[25] 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).

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

[27] Holger Brüggemann, Arne Hagman, Matthieu Jules, et al. „Virulence Strategies for Infecting Phagocytes Deduced from the in Vivo Transcriptional Program of Legionella Pneumophila“. In: *Cellular Microbiology* 8.8 (2006), pp. 1228–1240 (cit. on pp. 6, 7).

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

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

[30] Hagen Wieland, Susanne Ullrich, Florian Lang, and Birgid Neumeister. „Intracellular Multiplication of Legionella Pneumophila Depends on Host Cell Amino Acid Transporter SLC1A5“. In: *Molecular Microbiology* 55.5 (2005), pp. 1528–1537 (cit. on p. 7).

[31] 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“. In: *PLOS Pathogens* 5.12 (Dec. 2009), e1000704 (cit. on pp. 7, 136).

[32] 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“. In: *Cell Reports* 21.8 (Nov. 2017), pp. 2031–2038 (cit. on p. 7).

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

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

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

[36] 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“. In: *Proceedings of the National Academy of Sciences* 103.49 (Dec. 2006), pp. 18745–18750 (cit. on p. 7).

[37] 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“. In: *Cell Host & Microbe* 13.4 (Apr. 2013), pp. 395–405 (cit. on pp. 7, 75).

[38] Jesse R. Dixon, David U. Gorkin, and Bing Ren. „Chromatin Domains: The Unit of Chromosome Organization“. In: *Molecular Cell* 62.5 (June 2016), pp. 668–680 (cit. on p. 7).

[39] Robert Schneider and Rudolf Grosschedl. „Dynamics and Interplay of Nuclear Architecture, Genome Organization, and Gene Expression“. In: *Genes & Development* 21.23 (Dec. 2007), pp. 3027–3043 (cit. on p. 7).

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

[41] Doris L. LaRock, Anu Chaudhary, and Samuel I. Miller. „Salmonellae Interactions with Host Processes“. In: *Nature Reviews Microbiology* 13.4 (Apr. 2015), pp. 191–205 (cit. on pp. 8, 116).

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

[43] Andreas J. Bäumler. „The Record of Horizontal Gene Transfer in Salmonella“. In: *Trends in Microbiology* 5.8 (Aug. 1997), pp. 318–322 (cit. on p. 8).

[44] Marcelo B. Sztein, Andrea C. Bafford, and Rosângela Salerno-Goncalves. „Salmonella Enterica Serovar Typhi Exposure Elicits Ex Vivo Cell-Type-Specific Epigenetic Changes in Human Gut Cells“. In: *Scientific Reports* 10.1 (Aug. 2020), p. 13581 (cit. on p. 8).

[45] Yuanmei Wang, Liying Liu, Min Li, et al. „Chicken Cecal DNA Methylome Alteration in the Response to Salmonella Enterica Serovar Enteritidis Inoculation“. In: *BMC Genomics* 21.1 (Dec. 2020), pp. 1–14 (cit. on p. 8).

[46] Zhongyong Gou, Ranran Liu, Guiping Zhao, et al. „Epigenetic Modification of TLRs in Leukocytes Is Associated with Increased Susceptibility to Salmonella Enteritidis in Chickens“. In: *PLOS ONE* 7.3 (Mar. 2012), e33627 (cit. on p. 8).

[47] 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).

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

[49] Michal Levo and Eran Segal. „In Pursuit of Design Principles of Regulatory Sequences“. In: *Nature Reviews Genetics* 15.7 (July 2014), pp. 453–468 (cit. on p. 11).

[50] Michael S Brown and Joseph L Goldstein. „The SREBP Pathway: Regulation of Cholesterol Metabolism by Proteolysis of a Membrane-Bound Transcription Factor“. In: *Cell* 89.3 (May 1997), pp. 331–340 (cit. on p. 11).

[51] David G. Christensen, Xueshu Xie, Nathan Basisty, et al. „Post-Translational Protein Acetylation: An Elegant Mechanism for Bacteria to Dynamically Regulate Metabolic Functions“. In: *Frontiers in Microbiology* 10 (2019), p. 1604 (cit. on p. 11).

[52] Brian D. Strahl and C. David Allis. „The Language of Covalent Histone Modifications“. In: *Nature* 403.6765 (Jan. 2000), pp. 41–45 (cit. on p. 11).

[53] Jialiang Huang, Eugenio Marco, Luca Pinello, and Guo-Cheng Yuan. „Predicting Chromatin Organization Using Histone Marks“. In: *Genome Biology* 16.1 (Dec. 2015), p. 162 (cit. on p. 12).

[54] Liang Wang, Yifei Gao, Xiangdong Zheng, et al. „Histone Modifications Regulate Chromatin Compartmentalization by Contributing to a Phase Separation Mechanism“. In: *Molecular Cell* 76.4 (Nov. 2019), 646–659.e6 (cit. on p. 12).

[55] Elizaveta V. Benevolenskaya. „Histone H3K4 Demethylases Are Essential in Development and differentiationThis Paper Is One of a Selection of Papers Published in This Special Issue, Entitled 28th International West Coast Chromatin and Chromosome Conference, and Has Undergone the Journal's Usual Peer Review Process." In: *Biochemistry and Cell Biology* 85.4 (Aug. 2007), pp. 435–443 (cit. on p. 13).

[56] Gangning Liang, Joy C. Y. Lin, Vivian Wei, et al. „Distinct Localization of Histone H3 Acetylation and H3-K4 Methylation to the Transcription Start Sites in the Human Genome". In: *Proceedings of the National Academy of Sciences* 101.19 (May 2004), pp. 7357–7362 (cit. on p. 13).

[57] Christoph M. Koch, Robert M. Andrews, Paul Flicek, et al. „The Landscape of Histone Modifications across 1% of the Human Genome in Five Human Cell Lines". In: *Genome Research* 17.6 (June 2007), pp. 691–707 (cit. on p. 13).

[58] Andrey Poleshko, Cheryl L Smith, Son C Nguyen, et al. „H3K9me2 Orchestrates Inheritance of Spatial Positioning of Peripheral Heterochromatin through Mitosis". In: *eLife* 8 (Oct. 2019). Ed. by Andrés Aguilera, Jessica K Tyler, and Andrew S Belmont, e49278 (cit. on p. 13).

[59] Jeffrey A Rosenfeld, Zhibin Wang, Dustin E Schones, et al. „Determination of Enriched Histone Modifications in Non-Genic Portions of the Human Genome". In: *BMC Genomics* 10 (Mar. 2009), p. 143 (cit. on p. 13).

[60] Artem Barski, Suresh Cuddapah, Kairong Cui, et al. „High-Resolution Profiling of Histone Methylations in the Human Genome". In: *Cell* 129.4 (May 2007), pp. 823–837 (cit. on p. 13).

[61] Menno P. Creyghton, Albert W. Cheng, G. Grant Welstead, et al. „Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.50 (Dec. 2010), pp. 21931–21936 (cit. on p. 13).

[62] Paulina Kolasinska-Zwierz, Thomas Down, Isabel Latorre, et al. „Differential Chromatin Marking of Introns and Expressed Exons by H3K36me3". In: *Nature Genetics* 41.3 (Mar. 2009), pp. 376–381 (cit. on p. 13).

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

[64] Boyan Bonev and Giacomo Cavalli. „Organization and Function of the 3D Genome". In: *Nature Reviews Genetics* 17.11 (Nov. 2016), pp. 661–678 (cit. on pp. 13, 75).

[65] Nick Gilbert, Shelagh Boyle, Heike Fiegler, et al. „Chromatin Architecture of the Human Genome: Gene-Rich Domains Are Enriched in Open Chromatin Fibers". In: *Cell* 118.5 (Sept. 2004), pp. 555–566 (cit. on p. 13).

[66] Boryana Doyle, Geoffrey Fudenberg, Maxim Imakaev, and Leonid A. Mirny. „Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions". In: *PLOS Computational Biology* 10.10 (Oct. 2014), e1003867 (cit. on p. 13).

[67] Jill M. Dowen, Steve Bilodeau, David A. Orlando, et al. „Multiple Structural Maintenance of Chromosome Complexes at Transcriptional Regulatory Elements". In: *Stem Cell Reports* 1.5 (Nov. 2013), pp. 371–378 (cit. on p. 13).

[68] Dale Dorsett and Matthias Merkenschlager. „Cohesin at Active Genes: A Unifying Theme for Cohesin and Gene Expression from Model Organisms to Humans“. In: *Current Opinion in Cell Biology*. Cell Nucleus 25.3 (June 2013), pp. 327–333 (cit. on p. 13).

[69] Xiong Ji, Daniel B. Dadon, Benjamin E. Powell, et al. „3D Chromosome Regulatory Landscape of Human Pluripotent Cells“. In: *Cell Stem Cell* 18.2 (Feb. 2016), pp. 262–275 (cit. on p. 13).

[70] Tsung-Han S. Hsieh, Assaf Weiner, Bryan Lajoie, et al. „Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C“. In: *Cell* 162.1 (July 2015), pp. 108–119 (cit. on p. 13).

[71] Keerthi T. Chathoth and Nicolae Radu Zabet. „Chromatin Architecture Reorganization during Neuronal Cell Differentiation in Drosophila Genome“. In: *Genome Research* 29.4 (Apr. 2019), pp. 613–625 (cit. on p. 13).

[72] J. Dekker. „Capturing Chromosome Conformation“. In: *Science* 295.5558 (Feb. 2002), pp. 1306–1311 (cit. on p. 14).

[73] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, et al. „Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome“. In: *Science* 326.5950 (Oct. 2009), pp. 289–293 (cit. on pp. 15, 16, 32).

[74] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, et al. „Formation of Chromosomal Domains by Loop Extrusion“. In: *Cell Reports* 15.9 (May 2016), pp. 2038–2049 (cit. on pp. 16, 17).

[75] Pierre Therizols, Tarn Duong, Bernard Dujon, Christophe Zimmer, and Emmanuelle Fabre. „Chromosome Arm Length and Nuclear Constraints Determine the Dynamic Relationship of Yeast Subtelomeres“. In: *Proceedings of the National Academy of Sciences* 107.5 (Feb. 2010), pp. 2025–2030 (cit. on p. 16).

[76] Margit Schardin, T. Cremer, H. D. Hager, and M. Lang. „Specific Staining of Human Chromosomes in Chinese Hamster x Man Hybrid Cell Lines Demonstrates Interphase Chromosome Territories“. In: *Human Genetics* 71.4 (Dec. 1985), pp. 281–287 (cit. on p. 16).

[77] Thomas Cremer and Marion Cremer. „Chromosome Territories“. In: *Cold Spring Harbor Perspectives in Biology* 2.3 (Mar. 2010), a003889 (cit. on p. 16).

[78] Angela Taddei and Susan M Gasser. „Structure and Function in the Budding Yeast Nucleus“. In: *Genetics* 192.1 (Sept. 2012), pp. 107–129 (cit. on p. 16).

[79] Pengfei Dong, Xiaoyu Tu, Po-Yu Chu, et al. „3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments“. In: *Molecular Plant* 10.12 (Dec. 2017), pp. 1497–1509 (cit. on p. 16).

[80] M. Jordan Rowley, Michael H. Nichols, Xiaowen Lyu, et al. „Evolutionarily Conserved Principles Predict 3D Chromatin Organization“. In: *Molecular Cell* 67.5 (Sept. 2017), 837–852.e7 (cit. on p. 16).

[81] Nils Krietenstein, Sameer Abraham, Sergey V. Venev, et al. „Ultrastructural Details of Mammalian Chromosome Architecture“. In: *Molecular Cell* 78.3 (May 2020), 554–565.e7 (cit. on p. 16).

[82] Bas van Steensel and Andrew S. Belmont. „Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression“. In: *Cell* 169.5 (May 2017), pp. 780–791 (cit. on p. 16).

[83] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, et al. „Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions“. In: *Nature* 485.7398 (May 2012), pp. 376–380 (cit. on p. 17).

[84] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, et al. „Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre“. In: *Nature* 485.7398 (May 2012), pp. 381–385 (cit. on p. 17).

[85] Chunhui Hou, Li Li, Zhaohui S. Qin, and Victor G. Corces. „Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains“. In: *Molecular Cell* 48.3 (Nov. 2012), pp. 471–484 (cit. on p. 17).

[86] Suhas S.P. Rao, Su-Chen Huang, Brian Glenn St Hilaire, et al. „Cohesin Loss Eliminates All Loop Domains“. In: *Cell* 171.2 (Oct. 2017), 305–320.e24 (cit. on pp. 17, 72).

[87] Tim J. Stevens, David Lando, Srinjan Basu, et al. „3D Structures of Individual Mammalian Genomes Studied by Single-Cell Hi-C“. In: *Nature* 544.7648 (Apr. 2017), pp. 59–64 (cit. on p. 17).

[88] Jonathan A. Beagan and Jennifer E. Phillips-Cremins. „On the Existence and Functionality of Topologically Associating Domains“. In: *Nature Genetics* 52.1 (Jan. 2020), pp. 8–16 (cit. on p. 17).

[89] Li-Hsin Chang, Sourav Ghosh, and Daan Noordermeer. „TADs and Their Borders: Free Movement or Building a Wall?“ In: *Journal of Molecular Biology*. Perspectives on Chromosome Folding 432.3 (Feb. 2020), pp. 643–652 (cit. on p. 17).

[90] Elzo de Wit. „TADs as the Caller Calls Them“. In: *Journal of Molecular Biology*. Perspectives on Chromosome Folding 432.3 (Feb. 2020), pp. 638–642 (cit. on p. 17).

[91] Edward J. Banigan and Leonid A. Mirny. „Loop Extrusion: Theory Meets Single-Molecule Experiments“. In: *Current Opinion in Cell Biology*. Cell Nucleus 64 (June 2020), pp. 124–138 (cit. on p. 17).

[92] Riccardo Calandrelli, Qiuyang Wu, Jihong Guan, and Sheng Zhong. „GITAR: An Open Source Tool for Analysis and Visualization of Hi-C Data“. In: *Genomics, Proteomics & Bioinformatics*. Bioinformatics Commons (II) 16.5 (Oct. 2018), pp. 365–372 (cit. on p. 17).

[93] Mattia Forcato, Chiara Nicoletti, Koustav Pal, et al. „Comparison of Computational Methods for Hi-C Data Analysis“. In: *Nature Methods* 14.7 (July 2017), pp. 679–685 (cit. on p. 17).

[94] Bryan R. Lajoie, Job Dekker, and Noam Kaplan. „The Hitchhiker's Guide to Hi-C Analysis: Practical Guidelines“. In: *Methods* 72 (Jan. 2015), pp. 65–75 (cit. on pp. 17, 33).

[95] Sergey Venev, Nezar Abdennur, Anton Goloborodko, et al. *Open2c/Cooltools: V0.4.0*. Zenodo. Apr. 2021 (cit. on pp. 18, 37).

[96] Elisa Salviato, Vera Djordjilović, Judith Mary Hariprakash, et al. „Leveraging Three-Dimensional Chromatin Architecture for Effective Reconstruction of Enhancer–Target Gene Regulatory Interactions“. In: *Nucleic Acids Research* gkab547 (July 2021) (cit. on p. 20).

[97] Emily Crane, Qian Bian, Rachel Patton McCord, et al. „Condensin-Driven Remodelling of X Chromosome Topology during Dosage Compensation“. In: *Nature* 523.7559 (July 2015), pp. 240–244 (cit. on p. 20).

[98] 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. 20, 21).

[99] Neva C. Durand, James T. Robinson, Muhammad S. Shamim, et al. „Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom“. In: *Cell Systems* 3.1 (July 2016), pp. 99–101 (cit. on p. 20).

[100] Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, et al. „High-Resolution TADs Reveal DNA Sequences Underlying Genome Organization in Flies“. In: *Nature Communications* 9.1 (Jan. 2018), p. 189 (cit. on p. 20).

[101] 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. 20).

[102] Guillaume Andrey, Thomas Montavon, Bénédicte Mascrez, et al. „A Switch Between Topological Domains Underlies *HoxD* Genes Collinearity in Mouse Limbs“. In: *Science* 340.6137 (June 2013), p. 1234167 (cit. on p. 20).

[103] Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, et al. „Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions“. In: *Cell* 161.5 (May 2015), pp. 1012–1025 (cit. on p. 20).

[104] „Cohesins and Condensins Orchestrate the 4D Dynamics of Yeast Chromosomes during the Cell Cycle“. In: *The EMBO Journal* 36.18 (Sept. 2017), pp. 2684–2697 (cit. on p. 20).

[105] 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. 21).

[106] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, et al. „HiCRep: Assessing the Reproducibility of Hi-C Data Using a Stratum-Adjusted Correlation Coefficient“. In: *Genome Research* 27.11 (Nov. 2017), pp. 1939–1949 (cit. on pp. 21, 138).

[107] 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. 21).

[108] 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 pp. 21, 69, 73).

[109] 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 pp. 21, 73).

[110] 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“. In: *Molecular Cell* 38.4 (May 2010), pp. 576–589 (cit. on p. 21).

[111] A. Zemach, I. E. McDaniel, P. Silva, and D. Zilberman. „Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation“. In: *Science* 328.5980 (May 2010), pp. 916–919 (cit. on p. 22).

[112] Miao Wang, Chen Guo, Liang Wang, et al. „Long Noncoding RNA GAS5 Promotes Bladder Cancer Cells Apoptosis through Inhibiting EZH2 Transcription“. In: *Cell Death & Disease* 9.2 (Feb. 2018), p. 238 (cit. on p. 22).

[113] 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. 22).

[114] 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. 22).

[115] J. Omar Yáñez-Cuna, Evgeny Z. Kvon, and Alexander Stark. „Deciphering the Transcriptional Cis-Regulatory Code“. In: *Trends in Genetics* 29.1 (Jan. 2013), pp. 11–22 (cit. on p. 22).

[116] Julien Mozziconacci, Mélody Merle, and Annick Lesne. „The 3D Genome Shapes the Regulatory Code of Developmental Genes“. In: *Journal of Molecular Biology*. Perspectives on Chromosome Folding 432.3 (Feb. 2020), pp. 712–723 (cit. on p. 22).

[117] O. Symmons, V. V. Uslu, T. Tsujimura, et al. „Functional and Topological Characteristics of Mammalian Regulatory Domains“. In: *Genome Research* 24.3 (Mar. 2014), pp. 390–400 (cit. on p. 22).

[118] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, et al. „MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data“. In: *Genome Biology* 21.1 (Dec. 2020), pp. 1–17 (cit. on p. 23).

[119] Anshul Kundaje, Wouter Meuleman, Jason Ernst, et al. „Integrative Analysis of 111 Reference Human Epigenomes“. In: *Nature* 518.7539 (Feb. 2015), pp. 317–330 (cit. on p. 23).

[120] Jun Yoshimura, Kazuki Ichikawa, Massa J. Shoura, et al. „Recompleting the Caenorhabditis Elegans Genome“. In: *Genome Research* 29.6 (June 2019), pp. 1009–1022 (cit. on p. 24).

[121] H. W. Mewes, K. Albermann, M. Bähr, et al. „Overview of the Yeast Genome“. In: *Nature* 387.S6632 (May 1997), pp. 7–8 (cit. on p. 24).

[122] F. Sanger, S. Nicklen, and A. R. Coulson. „DNA Sequencing with Chain-Terminating Inhibitors“. In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977), pp. 5463–5467 (cit. on p. 24).

[123] A. M. Maxam and W. Gilbert. „A New Method for Sequencing DNA“. In: *Proceedings of the National Academy of Sciences* 74.2 (Feb. 1977), pp. 560–564 (cit. on p. 24).

[124] F. Sanger, A.R. Coulson, G.F. Hong, D.F. Hill, and G.B. Petersen. „Nucleotide Sequence of Bacteriophage λ DNA“. In: *Journal of Molecular Biology* 162.4 (Dec. 1982), pp. 729–773 (cit. on p. 24).

[125] R. Baer, A. T. Bankier, M. D. Biggin, et al. „DNA Sequence and Expression of the B95-8 Epstein—Barr Virus Genome“. In: *Nature* 310.5974 (July 1984), pp. 207–211 (cit. on p. 24).

[126] Peter M. G. F. van Wezenbeek, Theo J. M. Hulsebos, and John G. G. Schoenmakers. „Nucleotide Sequence of the Filamentous Bacteriophage M13 DNA Genome: Comparison with Phage Fd“. In: *Gene* 11.1 (Oct. 1980), pp. 129–148 (cit. on p. 24).

[127] S. G. Oliver, Q. J. M. van der Aart, M. L. Agostoni-Carbone, et al. „The Complete DNA Sequence of Yeast Chromosome III“. In: *Nature* 357.6373 (May 1992), pp. 38–46 (cit. on pp. 24, 25).

[128] 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“. In: *Yeast* 6.6 (1990), pp. 521–534 (cit. on p. 24).

[129] John D. McPherson, Marco Marra, La Deana Hillier, et al. „A Physical Map of the Human Genome“. In: *Nature* 409.6822 (Feb. 2001), pp. 934–941 (cit. on p. 25).

[130] F Collins and D Galas. „A New Five-Year Plan for the U.S. Human Genome Project“. In: *Science* 262.5130 (Oct. 1993), pp. 43–46 (cit. on p. 25).

[131] M. D. Adams. „The Genome Sequence of Drosophila Melanogaster“. In: *Science* 287.5461 (Mar. 2000), pp. 2185–2195 (cit. on p. 25).

[132] J. Craig Venter, Mark D. Adams, Eugene W. Myers, et al. „The Sequence of the Human Genome“. In: *Science* 291.5507 (Feb. 2001), pp. 1304–1351 (cit. on p. 25).

[133] Phillip Compeau, P. Pevzner, and G. Tesler. „How to Apply de Bruijn Graphs to Genome Assembly.“ In: *Nature biotechnology* (2011) (cit. on pp. 25, 26).

[134] J. T. Simpson, K. Wong, S. D. Jackman, et al. „ABySS: A Parallel Assembler for Short Read Sequence Data“. In: *Genome Research* 19.6 (June 2009), pp. 1117–1123 (cit. on p. 25).

[135] D. R. Zerbino and E. Birney. „Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs“. In: *Genome Research* 18.5 (Feb. 2008), pp. 821–829 (cit. on p. 25).

[136] Jared T. Simpson and Mihai Pop. „The Theory and Practice of Genome Sequence Assembly“. In: *Annual Review of Genomics and Human Genetics* 16.1 (Aug. 2015), pp. 153–172 (cit. on p. 25).

[137] Karen H. Miga, Sergey Koren, Arang Rhie, et al. „Telomere-to-Telomere Assembly of a Complete Human X Chromosome“. In: *Nature* 585.7823 (Sept. 2020), pp. 79–84 (cit. on p. 26).

[138] Glennis A. Logsdon, Mitchell R. Vollger, PingHsun Hsieh, et al. „The Structure, Function and Evolution of a Complete Human Chromosome 8“. In: *Nature* (Apr. 2021), pp. 1–7 (cit. on p. 26).

[139] 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. 26).

[140] Miten Jain, Sergey Koren, Karen H. Miga, et al. „Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads“. In: *Nature Biotechnology* 36.4 (Apr. 2018), pp. 338–345 (cit. on p. 26).

[141] Pierre Morisse, Camille Marchet, Antoine Limasset, Thierry Lecroq, and Arnaud Lefebvre. „Scalable Long Read Self-Correction and Assembly Polishing with Multiple Sequence Alignment“. In: *Scientific Reports* 11.1 (Jan. 2021), p. 761 (cit. on p. 26).

[142] 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. 26).

[143] 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. 26).

[144] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. „Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads“. In: *Genome Research* 27.5 (May 2017), pp. 737–746 (cit. on p. 27).

[145] Bruce J. Walker, Thomas Abeel, Terrance Shea, et al. „Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement“. In: *PLOS ONE* 9.11 (Nov. 2014), e112963 (cit. on p. 27).

[146] Ritu Kundu, Joshua Casey, and Wing-Kin Sung. *HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies*. Biorxiv. Bioinformatics, Dec. 2019 (cit. on p. 27).

[147] Aaron M. Wenger, Paul Peluso, William J. Rowell, et al. „Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome“. In: *Nature Biotechnology* 37.10 (Oct. 2019), pp. 1155–1162 (cit. on p. 27).

[148] Dandan Lang, Shilai Zhang, Pingping Ren, et al. „Comparison of the Two Up-to-Date Sequencing Technologies for Genome Assembly: HiFi Reads of Pacific Biosciences Sequel II System and Ultralong Reads of Oxford Nanopore“. In: *GigaScience* 9.12 (Nov. 2020) (cit. on p. 27).

[149] Haoyu Cheng, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. „Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm“. In: *Nature Methods* 18.2 (Feb. 2021), pp. 170–175 (cit. on p. 27).

[150] Barış Ekim, Bonnie Berger, and Rayan Chikhi. „Minimizer-Space de Bruijn Graphs: Whole-Genome Assembly of Long Reads in Minutes on a Personal Computer“. In: *Cell Systems* (Sept. 2021) (cit. on p. 27).

[151] Ernest T Lam, Alex Hastie, Chin Lin, et al. „Genome Mapping on Nanochannel Arrays for Structural Variation Analysis and Sequence Assembly“. In: *Nature Biotechnology* 30.8 (Aug. 2012), pp. 771–776 (cit. on p. 27).

[152] Matt Ravenhall, Nives Škunca, Florent Lassalle, and Christophe Dessimoz. „Inferring Horizontal Gene Transfer“. In: *PLOS Computational Biology* 11.5 (May 2015). Ed. by Shoshana Wodak, e1004095 (cit. on p. 28).

[153] Genome 10K Community of Scientists. „Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species“. In: *Journal of Heredity* 100.6 (Nov. 2009), pp. 659–674 (cit. on p. 29).

[154] Monica Poelchau, Christopher Childers, Gary Moore, et al. „The I5k Workspace@NAL - Enabling Genomic Data Access, Visualization and Curation of Arthropod Genomes“. In: *Nucleic Acids Research* 43.D1 (Jan. 2015), pp. D714–D719 (cit. on p. 29).

[155] *Darwin Tree Of Life*. https://www.darwintreeoflife.org/ (cit. on p. 29).

[156] Deanna M Church, Valerie A Schneider, Karyn Meltz Steinberg, et al. „Extending Reference Assembly Models“. In: *Genome Biology* 16.1 (Dec. 2015), p. 13 (cit. on p. 29).

[157] Heng Li, Xiaowen Feng, and Chong Chu. „The Design and Construction of Reference Pangenome Graphs with Minigraph“. In: *Genome Biology* 21.1 (Dec. 2020), p. 265 (cit. on p. 29).

[158] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. „WASP: Allele-Specific Software for Robust Molecular Quantitative Trait Locus Discovery“. In: *Nature Methods* 12.11 (Nov. 2015), pp. 1061–1063 (cit. on p. 29).

[159] Charlotte Cockram, Agnès Thierry, and Romain Koszul. „Generation of Gene-Level Resolution Chromosome Contact Maps in Bacteria and Archaea“. In: *STAR Protocols* 2.2 (June 2021), p. 100512 (cit. on p. 32).

[160] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. „Normalization of a Chromosomal Contact Map“. In: *BMC Genomics* 13.1 (2012), p. 436 (cit. on p. 33).

[161] Nezar Abdennur and Leonid A Mirny. „Cooler: Scalable Storage for Hi-C Data and Other Genomically Labeled Arrays“. In: *Bioinformatics* 36.1 (Jan. 2020), pp. 311–316 (cit. on p. 34).

[162] 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. 34).

[163] Joachim Wolff, Rolf Backofen, and Björn Grüning. *Loop Detection Using Hi-C Data with HiCExplorer*. Biorxiv. Mar. 2020 (cit. on p. 37).

[164] 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“. In: *Cell* 159.7 (Dec. 2014), pp. 1665–1680 (cit. on p. 37).

[165] Murat İlsever and Cem Ünsalan. *Two-Dimensional Change Detection Methods*. SpringerBriefs in Computer Science. London: Springer London, 2012 (cit. on p. 70).

[166] Elphège P. Nora, Anton Goloborodko, Anne-Laure Valton, et al. „Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization“. In: *Cell* 169.5 (May 2017), 930–944.e22 (cit. on pp. 72, 74).

[167] Merve Sahin, Wilfred Wong, Yingqian Zhan, et al. „HiC-DC+ Enables Systematic 3D Interaction Calls and Differential Analysis for Hi-C and HiChIP“. In: *Nature Communications* 12.1 (June 2021), p. 3366 (cit. on p. 73).

[168] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q. Zhang. „FIND: difFerential Chromatin INteractions Detection Using a Spatial Poisson Process“. In: *Genome Research* 28.3 (Mar. 2018), pp. 412–422 (cit. on p. 73).

[169] Kate B Cook, Borislav H Hristov, Karine G Le Roch, Jean Philippe Vert, and William Stafford Noble. „Measuring Significant Changes in Chromatin Conformation with ACCOST“. In: *Nucleic Acids Research* 48.5 (Mar. 2020), pp. 2303–2311 (cit. on p. 73).

[170] Pedro Escoll, Sonia Mondino, Monica Rolando, and Carmen Buchrieser. „Targeting of Host Organelles by Pathogenic Bacteria: A Sophisticated Subversion Strategy“. In: *Nature Reviews Microbiology* 14.1 (Jan. 2016), pp. 5–19 (cit. on p. 75).

[171] Pedro Escoll, Ok-Ryul Song, Flávia Viana, et al. „Legionella Pneumophila Modulates Mitochondrial Dynamics to Trigger Metabolic Repurposing of Infected Macrophages“. In: *Cell Host & Microbe* 22.3 (Sept. 2017), 302–316.e7 (cit. on p. 75).

[172] Robert J. Neff. „Purification, Axenic Cultivation, and Description of a Soil Amoeba, Acanthamoeba Sp.“ In: *The Journal of Protozoology* 4.3 (Aug. 1957), pp. 176–182 (cit. on p. 75).

[173] Rolf Michel and Bärbel Hauröder. „Isolation of an Acanthamoeba Strain with Intracellular Burkholderia Pickettii Infection“. In: *Zentralblatt für Bakteriologie* 285.4 (Apr. 1997), pp. 541–557 (cit. on p. 75).

[174] D. M. Emms and S. Kelly. *STAG: Species Tree Inference from All Genes*. Biorxiv. Feb. 2018 (cit. on p. 112).

[175] David M Emms and Steven Kelly. „STRIDE: Species Tree Root Inference from Gene Duplication Events“. In: *Molecular Biology and Evolution* 34.12 (Dec. 2017), pp. 3267–3278 (cit. on p. 112).

[176] Paschalia Kapli, Ziheng Yang, and Maximilian J. Telford. „Phylogenetic Tree Building in the Genomic Age“. In: *Nature Reviews Genetics* 21.7 (July 2020), pp. 428–444 (cit. on p. 112).

[177] Anthony D. Schmitt, Ming Hu, Inkyung Jung, et al. „A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome“. In: *Cell Reports* 17.8 (Nov. 2016), pp. 2042–2059 (cit. on p. 116).

[178] Ibrahim Ahmed and Nahed Ismail. „M1 and M2 Macrophages Polarization via mTORC1 Influences Innate Immunity and Outcome of Ehrlichia Infection“. In: *Journal of Cellular Immunology* 2.3 (May 2020) (cit. on p. 116).

[179] Jörg Mages, Harald Dietrich, and Roland Lang. „A Genome-Wide Analysis of LPS Tolerance in Macrophages“. In: *Immunobiology* 212.9-10 (Jan. 2008), pp. 723–737 (cit. on p. 116).

[180] Chiara Porta, Monica Rimoldi, Geert Raes, et al. „Tolerance and M2 (Alternative) Macrophage Polarization Are Related Processes Orchestrated by P50 Nuclear Factor kappaB“. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.35 (Sept. 2009), pp. 14978–14983 (cit. on p. 117).

[181] Hnin Thanda Aung, Kate Schroder, Stewart R. Himes, et al. „LPS Regulates Proinflammatory Gene Expression in Macrophages by Altering Histone Deacetylase Expression“. In: *The FASEB Journal* 20.9 (2006), pp. 1315–1327 (cit. on p. 117).

[182] Scott Hotaling, Joanna L. Kelley, and Paul B. Frandsen. *Towards a Genome Sequence for Every Animal: Where Are We Now?* Biorxiv. Oct. 2021 (cit. on p. 136).

[183] Kyle T David, Alan E Wilson, and Kenneth M Halanych. „Sequencing Disparity in the Genomic Era“. In: *Molecular Biology and Evolution* 36.8 (Aug. 2019), pp. 1624–1627 (cit. on p. 136).

[184] Matthieu Legendre, Audrey Lartigue, Lionel Bertaux, et al. „In-Depth Study of *Mollivirus Sibericum* , a New 30,000-y-Old Giant Virus Infecting *Acanthamoeba*“. In: *Proceedings of the National Academy of Sciences* 112.38 (Sept. 2015), E5327–E5335 (cit. on p. 136).

[185] Lena König, Alexander Siegl, Thomas Penz, et al. „Biphasic Metabolism and Host Interaction of a Chlamydial Symbiont“. In: *mSystems* 2.3 (2017 May-Jun), e00202–16 (cit. on p. 136).

[186] Ari B. Molofsky and Michele S. Swanson. „Differentiate to Thrive: Lessons from the Legionella Pneumophila Life Cycle: Regulation of Legionella Differentiation“. In: *Molecular Microbiology* 53.1 (June 2004), pp. 29–40 (cit. on p. 136).

[187] J. P. Vogel. „Conjugative Transfer by the Virulence System of Legionella Pneumophila“. In: *Science* 279.5352 (Feb. 1998), pp. 873–876 (cit. on p. 137).

[188] Seungsoo Kim, Maitreya J Dunham, and Jay Shendure. „A Combination of Transcription Factors Mediates Inducible Interchromosomal Contacts“. In: *eLife* 8 (May 2019). Ed. by Jeannie T Lee, Kevin Struhl, Peter J Fraser, and Hsueh ping Chu, e42499 (cit. on p. 137).

[189] 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?“ In: *RNA* 22.6 (June 2016), pp. 839–851 (cit. on p. 138).

[190] Paul P. Gardner, James M. Paterson, Stephanie McGimpsey, et al. *Sustained Software Development, Not Number of Citations or Journal Choice, Is Indicative of Accurate Bioinformatic Software*. Biorxiv. Sept. 2021 (cit. on p. 138).

[191] *Review Criteria — JOSS Documentation*. https://joss.readthedocs.io/ en/latest/review_criteria.html (cit. on p. 139).

[192] Philip A. Ewels, Alexander Peltzer, Sven Fillinger, et al. „The Nf-Core Framework for Community-Curated Bioinformatics Pipelines“. In: *Nature Biotechnology* 38.3 (Mar. 2020), pp. 276–278 (cit. on p. 139).

[193] Laura Wratten, Andreas Wilm, and Jonathan Göke. „Reproducible, Scalable, and Shareable Analysis Pipelines with Bioinformatics Workflow Managers“. In: *Nature Methods* (Sept. 2021), pp. 1–8 (cit. on p. 139).

[194] *Zenodo - Research. Shared.* https:// about.zenodo.org/ (cit. on p. 139).

[195] *NumFOCUS: A Nonprofit Supporting Open Code for Better Science*. https:// numfocus.org/ (cit. on p. 139).

[196] Felix Stiehler, Marvin Steinborn, Stephan Scholz, et al. „Helixer: Cross-Species Gene Annotation of Large Eukaryotic Genomes Using Deep Learning“. In: *Bioinformatics* 36.22-23 (Dec. 2020), pp. 5291–5298 (cit. on p. 140).

[197] Ghazaleh Khodabandelou, Etienne Routhier, and Julien Mozziconacci. „Genome Annotation across Species Using Deep Convolutional Neural Networks“. In: *PeerJ Computer Science* 6 (June 2020), e278 (cit. on p. 140).

[198] Ryan Poplin, Pi-Chuan Chang, David Alexander, et al. „A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks“. In: *Nature Biotechnology* 36.10 (Nov. 2018), pp. 983–987 (cit. on p. 140).

[199] Steven T Hill, Rachael Kuintzle, Amy Teegarden, et al. „A Deep Recurrent Neural Network Discovers Complex Biological Rules to Decipher RNA Protein-Coding Potential“. In: *Nucleic Acids Research* 46.16 (Sept. 2018), pp. 8105–8113 (cit. on p. 140).

[200] Etienne Routhier, Edgard Pierre, Ghazaleh Khodabandelou, and Julien Mozziconacci. „Genome-Wide Prediction of DNA Mutation Effect on Nucleosome Positions for Yeast Synthetic Genomics“. In: *Genome Research* 31.2 (Feb. 2021), pp. 317–326 (cit. on p. 140).

[201] John Jumper, Richard Evans, Alexander Pritzel, et al. „Highly Accurate Protein Structure Prediction with AlphaFold“. In: *Nature* (July 2021), pp. 1–7 (cit. on p. 140).

[202] Žiga Avsec, Vikram Agarwal, Daniel Visentin, et al. *Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions*. Biorxiv. Apr. 2021 (cit. on p. 140).

[203] Geoff Fudenberg, David R. Kelley, and Katherine S. Pollard. „Predicting 3D Genome Folding from DNA Sequence with Akita“. In: *Nature Methods* 17.11 (Nov. 2020), pp. 1111–1117 (cit. on p. 140).

[204] Ron Schwessinger, Matthew Gosden, Damien Downes, et al. „DeepC: Predicting 3D Genome Folding Using Megabase-Scale Transfer Learning“. In: *Nature Methods* 17.11 (Nov. 2020), pp. 1118–1124 (cit. on p. 140).

[205] Amlan Talukder, Clayton Barham, Xiaoman Li, and Haiyan Hu. „Interpretation of Deep Learning in Genomics and Epigenomics“. In: *Briefings in Bioinformatics* 22.3 (May 2021) (cit. on p. 141).

[206] H. Muller, V. Scholari, N. Agier, et al. „Characterizing Meiotic Chromosomes' Structure and Pairing Using a Designer Sequence Optimized for Hi-C“. In: *Molecular Systems Biology* 14.7 (July 2018), e8293 (cit. on p. 142).

[207] *Neural Network - 2-D Convolution as a Matrix-Matrix Multiplication*. https:// stackoverflow.com/q/16798888 (cit. on p. 144).

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 16 décembre 2021

Cyril Matthey-Doret