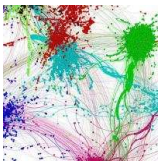
Investigating chromosome dynamics through Hi-C assembly

Lyam Baudry



Sorbonne Université

Complexité du Vivant

Unité de Régulation Spatiale des Génomes

Investigating chromosome dynamics through Hi-C assembly

Par Lyam Baudry

Thèse de doctorat de Bioinformatique

Dirigée par Romain Koszul

Présentée et soutenue publiquement le 13 juin 2019

Devant un jury composé de :

Olivier Jaillon, directeur de recherche	Rapporteur
Thomas Sexton, directeur de recherche	Rapporteur
Gilles Fischer, directeur de recherche	Examinateur
Claire Lemaitre, chargée de recherche	Examinatrice
Nicolas Servant, chargé de recherche	Examinateur
Romain Koszul, directeur de recherche	Directeur de thèse

Contents

List of Figures

Acknowledgements

This PhD project could not have been achieved without the support and help from many people, each of whom I would like to thank personally.

Firstly, I would like to thank my supervisor Romain Koszul, to whom I am grateful for this unique and rich learning experience.

It has been an exhilarating four years. Over the course of this PhD work I have been granted unparalleled opportunities and autonomy in my research that let me explore a great diversity of areas, strike up a wide range of collaborations and learn in many fields. I like to believe these experiences have transformed me for the better and helped me mature as a person and a scientist.

I am also grateful to my other supervisor Martial Marbouty for his careful guidance and direction, and appreciate the time and effort he took to help me bring many projects to completion.

I would also like to thank members of my PhD committee as well as my thesis reporters for taking the time to review this manuscript.

Many thanks to all members of the RSG lab: Axel, Pierrick and Vittore, with whom I've enjoyed many a fruitful (and not-so-fruitful) discussion over drinks (across the coffee machine, at the CCC, or at the many *pots* we've had), exchanged some scientific (and not-so-scientific) insights, which sometimes blossomed to a full-blown project. Thanks Cyril, Nadège and Théo for bearing with me and all the buggy code I've written. Thanks to Charlie for helping me proofread and correct the manuscript. And thanks to Agnès, Aurèle, Brenna and Rémi, for contributing to the lab's uniquely enjoyable atmosphere.

I would also like to thank many people from Pasteur, past and present, without whom my time as a PhD would have been much less interesting: Alicia, Antoine, Célia, Coralie, Guillaume, Héloïse, Hervé, Luciana, Nina, Maria, Marwah...the list goes on!

Grazie mille alla mia biologa marina preferita per la tua incredibile dolcezza.

شكراً لأفضل الأمهات لوجودهن دائمًا من أجلي ودعمك لي دون قيد أو شرط.

Merci à mes soeurs adorées qui m'ont toujours soutenu ou rassuré quoi qu'il arrive. Ce manuscrit leur est dédié, ainsi qu'à ma mère.

Abstract

The advent of high-throughput DNA sequencing technologies has set off an expanding trend in genome assembling and scaffolding. Once limited to a few model organisms, chromosome-level assemblies for an ever expanding range of species are now made possible by these technologies. Such genome quality is an essential preliminary to understand interactions between and among chromosomes.

We built upon a computational and technological framework that let us tackle genome assembly problems of increasing complexity. Our methods are mainly based on chromosome conformation capture technologies such as Hi-C. In a Hi-C experiment, DNA molecules are cross-linked with the surrounding proteins and form a large, static protein-DNA complex. This captures the spatial conformation by trapping together molecules that are physically close to each other. Therefore, Hi-C is very suitable for 3D genome structure analysis, which lets us infer a wealth of information about the genome. It was indeed shown that the tridimensional structure of the genome can be unambiguously linked to its 1D structure thanks to the physical properties of DNA polymers. Moreover, such 3D proximity also gives access to cell compartment information, thus opening the way for an additional approach for metagenomic binning. Both of these methods were implemented as proof-of-concepts or on simple model organisms where prior information was easily available for reference.

In this work, we expand upon these methods and apply them to use cases with more and more complexity.

1 Introduction

One of the most salient characteristics of our modern era is its current fascination with DNA. Due to the tremendous advances in sequencing technologies made over the recent decade, coupled with the ever decreasing costs of computational resources, all fields related to the study of DNA have received increased attention and have expanded beyond early expectations: from comparative genomics to metagenomics to evolutionary genomics, the scale and pace at which new discoveries are made have dramatically accelerated and show no signs of slowing down.

These advances have thoroughly transformed the way we think about DNA: no longer a static string of code binding the individual according to some linear phenotype-to-genotype dogma, it is indeed a dynamic molecule in a constant state of interaction: with itself, as befits any polymer subject to random looping as well as internal rearrangements; with its cell environment, as various proteins bind to specific regions and in turn affect transcription levels and functional ability; with other DNA molecules, as more dramatic rearrangements occur, such as DNA transfers; and lastly, with time, as evolution runs its course and its additional mechanisms further complicate our model of understanding.

One may thus understand the need to access the entirety of a species' genome in order to draw definite conclusions about its global picture, whether it be evolutionary, structural, or functional. Yet, to this day, relatively few genomes have been fully characterized, with respect to the estimated 11 million species present on the planet; in fact, with few exceptions, the genome of virtually all species that have been sequenced is known in a more or less fragmented state. Still, many independent efforts have been made to bring the genome of many species ever closer to a state of full completeness. With this present work, we hope to contribute an additional step into that direction.

In the following sections, we will present a brief overview of genome sequencing technologies and the corresponding genome assembly strategies exploiting such sequence data, from historical practices to current state-of-the-art techniques. We will explain the approach we have chosen (chromosome conformation capture), and how its mathematical and computational framework fits into global picture of genome assembly, adding to it and bringing extra insights into chromosome dynamics from species of interest.

1.1 Overview of genome sequencing

DNA sequencing technologies are evolving quickly and have been traditionally separated into *generations*; the first concerns early efforts to obtain the sequences of model organisms, whereas the second may be argued to have kicked off the so-called 'genomics revolution' at the turn of the millennium. The third generation usually refers to ad-

vances made over the last decade, mostly regarding *long read* technology, which we will detail below.

1.1.1 First generation: early genomes

In this section we will cover the first achievements and long-spanned projects that characterized the first generation of sequencing.

1.1.1.1 Early efforts

Early landmarks on DNA sequencing were achieved in the sixties, in the trail of the discovery of double-helix 3D structure of DNA the decade before [1]. Available techniques were mostly focused on proteins, *i.e.* shorter sequences whose base units were very different from one another [2]. Early protocols, inspired from analytical chemistry, could not determine sequence order, and were not adapted to DNA [3].

The very first efforts to infer the order of nucleotide base pairs were derived from RNA-related techniques, notably due to their single-stranded nature and shorter length, making them simpler to analyze by techniques at the time [4], involving specific treatments to partially degrade RNA fragments. For instance, in 1965, the very first fully characterized acid nucleic sequence was that of alanine tRNA from the baker yeast *Saccharomyces cerevisiae*, by Robert Holley [5].

Gradually, a number of landmarks were made. The first protein-coding gene sequence was determined in 1972 using a two-dimensional fractionation method by Walter Friers [6]. It was the coat protein of the bacteriophage MS2, whose full genome (3,569 bp) would be characterized in 1976, making it the first genome ever sequenced [7]. However, its sequence had still been determined at the RNA level, the phage being a single-stranded RNA virus.

DNA-specific sequencing methods began to arise with the use of DNA polymerase [8] [9]. In 1970, using the *Enterobacteria* phage λ as a target, Ray Wu and Dale Kaiser added radioactive nucleotides one by one with the enzyme, measuring each time the composition to infer the actual order of incorporation [10]. The use of location-specific oligonucleotides to help prime the polymerase would enable the sequencing at any region in the molecule [11]. Although these primer-extension methods formed the basis for future advances and helped sequence more genes [12], they were still time and resource consuming, as they involved 2D fractionation, and could not scale beyond very short molecules [13].

1.1.1.2 The *plus and minus* method

A number of changes simplified the sequencing protocols, and helped achieve further landmarks. The use of polyacrylamide gels made 2D fractionation unnecessary, as their separation power during an electrophoresis was much more resolutive [13]. However, the first design shift came in 1975 with Sanger's and Coulson's *plus and minus* method [14]. Like above, a primer and DNA polymerase are used to incorporate DNA in the presence of ^{32}P-labelled nucleotides. A separate mix was created for each nucleotide

radiollabeled this way. Then, each mix would be then split for two joint reactions: the first one would only use the specific nucleotide that was labelled (the *plus* reaction) and the second would use all other three (the *minus* reaction). The principle is illustrated in figure 1. A set of eight reactions, two for each nucleotide, was thus run. For each nucleotide, there would be extension sequences only ending with that base, and a set of sequences that terminate right before that nucleotide's position. A polyacrylamide gel featuring all eight runs could help infer the position of every single nucleotide in a genome [15]. This technique was successfully used by Sanger to sequence the genome of bacteriophage ϕX174 (or *PhiX*), the first ever DNA genome [16].

In parallel, in 1976, Maxam and Gilbert would use specific chemicals instead of DNA polymerase that break up the strand at specific positions (DNA sequencing by chemical degradation). Given these, as well as the fragments generated this way, it was possible to infer the exact order of the sequence [17]. However, the technical complexity of the procedure, as well as the dangerous chemicals it required handling, meant it rapidly fell into disuse with the advent of Sanger sequencing ([18], see the following section).

1.1.1.3 Chain-termination, or *Sanger sequencing*

The first breakthrough was made by Sanger in 1977, whose eponymous protocol would later define first-generation sequencing for decades to come [19]. Also called *chain-termination* or *dideoxy* method, it uses so-called *dideoxynucleotides* (ddNTPs), which are deoxyribonucleotide (dNTPs) analogues lacking the hydroxyl group in 3' (figure 2).

That group is necessary to extend the DNA chain, and a chain ending with a ddNTP cannot bind with the 5' phosphate group of another dNTP [20]. A mix of radioactive ddNTPs and dNTP is prepared for a DNA extension reaction, with ddNTPs present in smaller amounts than dNTP; the end result is that ddNTP will be sometimes randomly selected for incorporation uniformly across the strand, stopping its extension at various points. Eventually chains that stop at every single point in the genome will be generated. Such a reaction is run for each nucleotide and its corresponding dNTP/ddNTP base mix. With the help of a polyacrylamide gel combining all four results, one may deduce the original nucleotide sequence order: each position will be matched by a fragment stopping exactly there [2][13]. The complete protocol is shown in figure 3.

This major advance enabled the sequencing of longer molecules; in 1981, Sanger published the complete, 48 kb long sequence of the λ bacteriophage [22], which was the longest at the time, and in 1984, the Medical Research Council published the 172 kb long sequence of the Epstein-Barr virus [23]. This was a major undertaking that took three years, but progress accelerated with time. The exponential evolution of the amount of publicly available DNA sequences since the first ever genome is shown in figure 4. For this design breakthrough, Sanger was rewarded with the 1980 Nobel price in chemistry, which he shared with Gilbert.

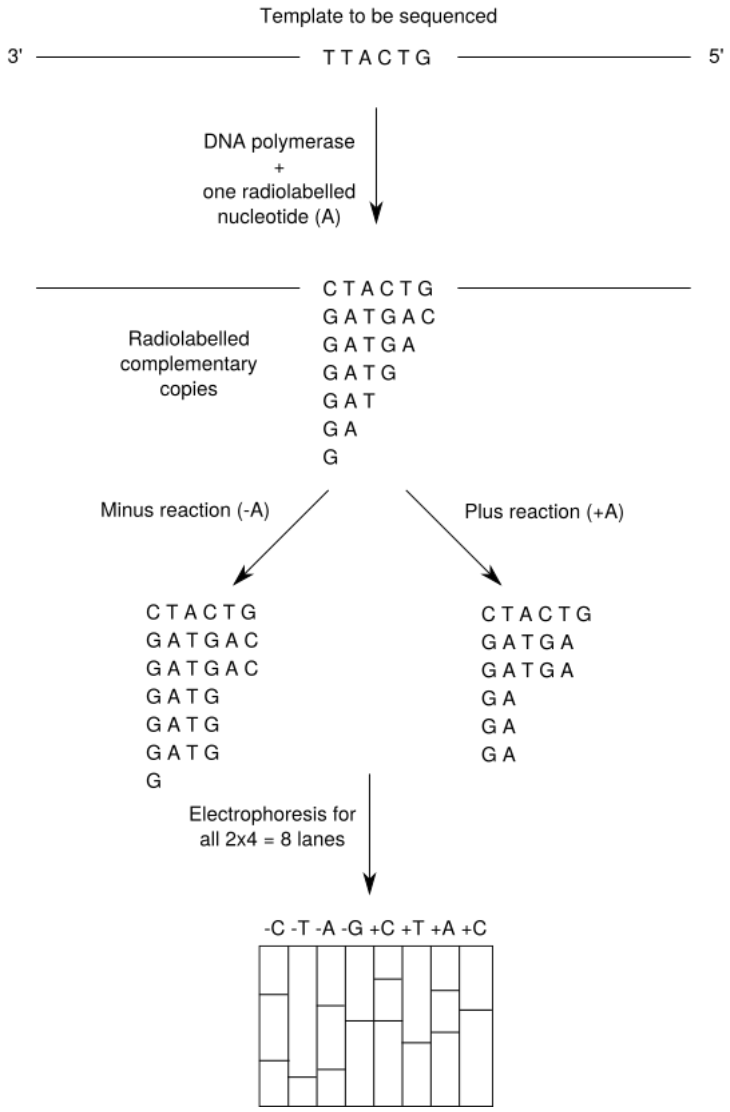


Figure 1: The *plus and minus* method

Source: Inspired from Sanger, Coulson *et al.*, 1975, [14]

1.1.1.4 Industrial developments and further landmarks

Sequencing development was accelerated as the original Sanger protocol was subject to various improvements [24] [13], such as replacing radioactive with fluorescent labelling [25] [26] [27] [28] [29], the use of capillary based electrophoreses [30] [31], or a more suited polymerase [32]. Increasingly, it became possible to automate the sequencing of genomes at relatively cheap costs, thus giving rise to the first commercial DNA sequencing ma-

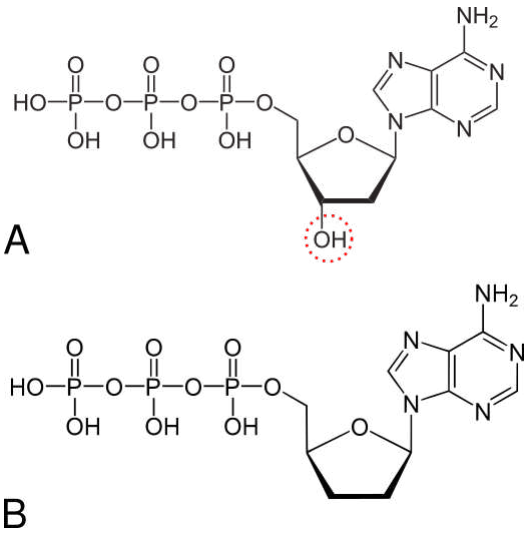


Figure 2: Representation of (A) a deoxyribonucleotide, where the characteristic hydroxyl group has been marked, and (B) a dideoxynucleotide, where it is missing.

chines. The first semi-automated one was announced in 1986 [29], followed by Applied Biosystems releasing the first fully automated sequencing machine (ABI 370) in 1987 [33], which was rapidly put to use to succesfully determine the sequence of a gene [34]. Later, Craig Venter would set up the Institute for Genomic Research (TIGR), putting together 30 ABI 373A automated sequencers and 17 ABI Catalyst 800 robots [35] [36], helping the sequencing efforts gain momentum.

These machines were limited in output and could only produce short chunks called *reads*, which were approximately one kilobase long. In order to infer the sequence of longer genomes, many such fragments had to be cloned and sequenced, and the overlaps were to be assembled *in silico* ([37], see section 1.2); thus, *shotgun sequencing* was born [38]. To that end, very highly concentrated amounts of DNA had to be produced so that the redundancy provided by the read overlaps would make the assembly process easier. This was facilitated with the development of polymerase chain reaction (PCR) [39] [40] and recombinant DNA technologies [41] [42].

As a result of increased automation, landmarks were achieved rapidly at the turn of the 1990's and beyond: following the release of the direct blotting electrophoresis system (GATC 1500) sequencer by GATC Biotech, the complete sequence of *Saccaromyces cerevisiae*'s chrmosome III was published in 1994 [43]; the remaining fifteen chromosomes would be sequenced in 1996 by an international consortium [44], making *S. cerevisiae* the first eukaryotic genome to be ever published. It was, however, predated by the sequencing of the *Haemohilus influenzae* genome in 1995, which, at 1.8 Mb long, was the first free-living organism to be ever sequenced [45]. Other genomes would later follow, such as *Bacillus subtilis* [46], *Escherichia coli* [47], the first animal *Caenorhabditis elegans* [48], and *Drosophila melanogaster* [49].

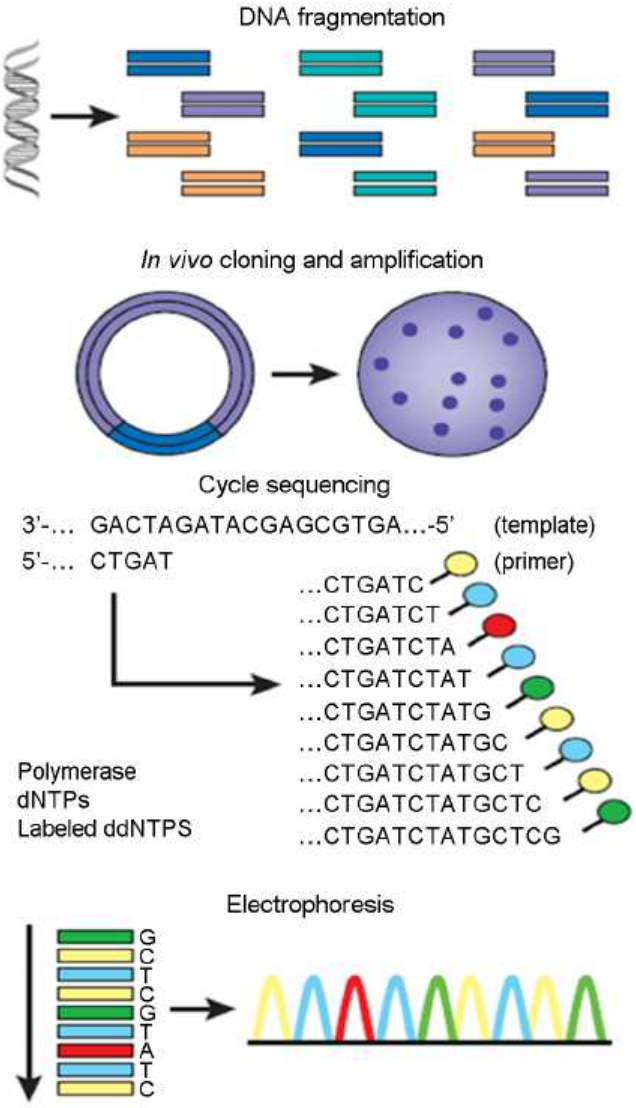


Figure 3: The Sanger protocol

Source: Adapted from Kim *et al.*, 2012, [21]

1.1.1.5 The Human Genome Project

These pioneering projects, and notably the European led consortium on the yeast genome, acted as successful proof-of-concept achievements that paved the way to the ambitious Human Genome Project (HGP), an initiative aiming at characterizing the entire euchromatic regions of the human genome within 15 years. A joint international consortium gathering institutes from Europe and Asia (notably including France's Génoscope - that provided the important genetic maps necessary to scaffold the chromosomes - and Ger-

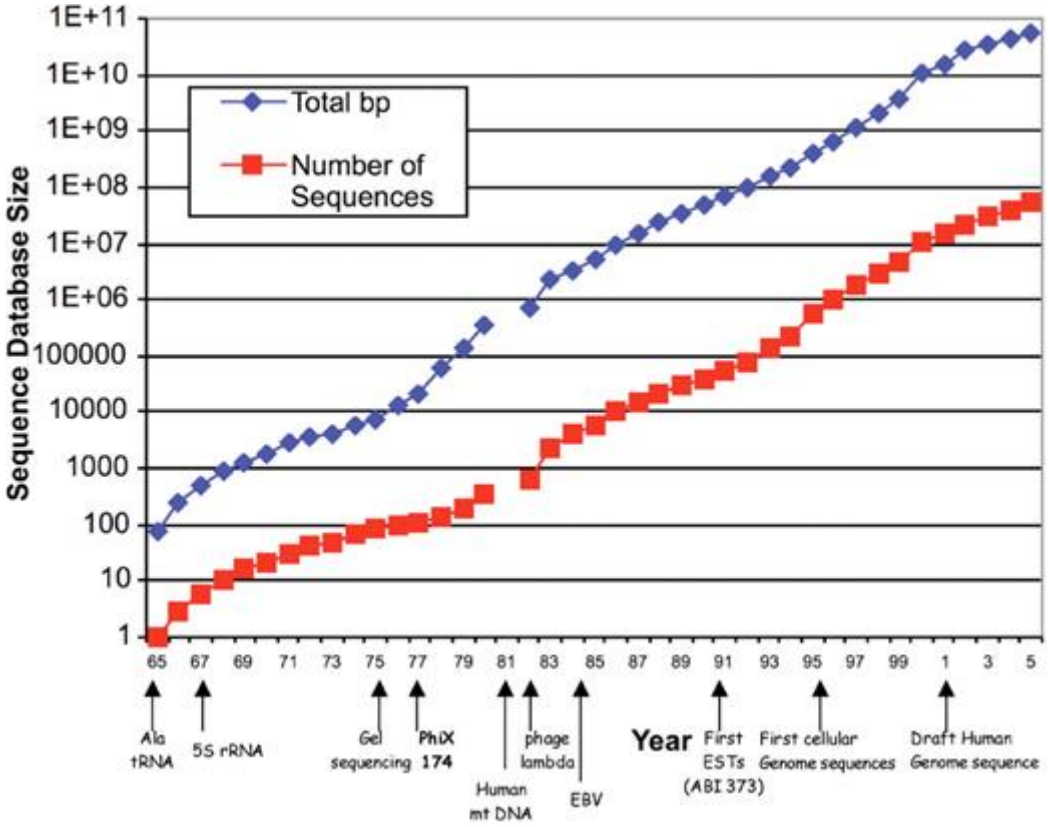


Figure 4: Evolution of publicly available sequences along the first sequencing generation

Source: Hutchison *et al.*, 2007, [2]

many's Max Planck Institute) contributed to the project [50]. It made use of vectors called bacterial artificial chromosomes (BACs); after breaking up the 3 Gb genome into 150 kb chunks, each of these was then incorporated into the vectors, using the bacteria's internal machinery to replicate and produce more DNA. The resulting molecules were then shotgun-sequenced as usual; this two-tiered method was called *hierarchical shotgun sequencing* [51]. Given the enormous size of the human genome, hundreds of such samples needed to be processed this way, and progress was spurred by the release of newer Sanger sequencers [52], such as the ABI PRISM machines from Applied Biosystems in 1998. In the meantime, Craig Venter split off from TIGR to create his own privately-funded company, Celera Genomics. Its business model was the creation of genomic data with shotgun approaches which researchers could access for a fee. It was also known for attempting to patent genes, filing preliminary applications for 6,500 of them. This was however abandoned when the US president and UK prime minister at the time released a joint statement in 2000, arguing that the human genome should not be patented. Overall, the increased competition incentivized the publicly-funded HGP to double down its efforts and the project was completed ahead of time; a "rough draft"

of the human genome was announced in 2000 [53] [54], and the genome was declared complete on April 13, 2003 [55]. Both Celera Genomics and the HGP shared the credit. However, many sequence gaps still remain to this day [56]. The heterochromatic regions (centromeres and telomeres) were outside the scope of the project, and figure 5 shows that as of 2018, hundreds of gaps of various sizes were still extant.

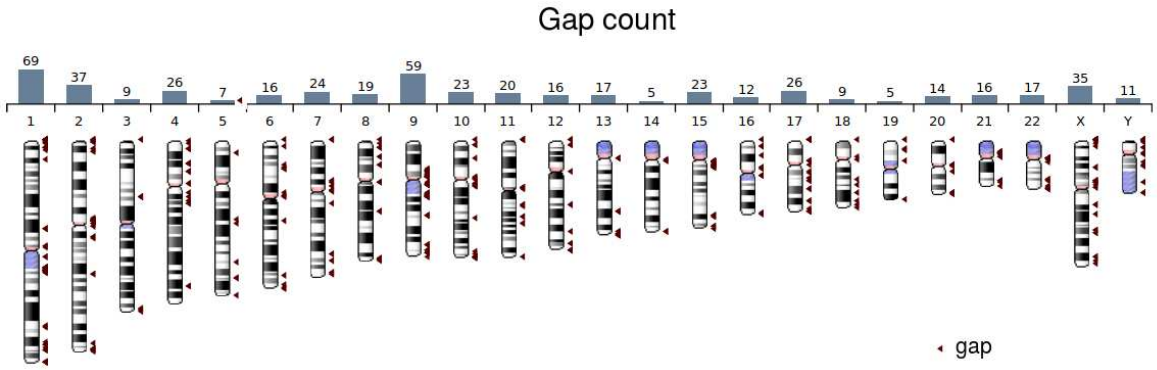


Figure 5: Human genome gap count and locations by chromosome.

Source: Adapted from the Genome Reference Consortium

1.1.2 Second generation: the 'genomic revolution'

In this section we will briefly cover how a radically different approach to sequencing in the early 2000's helped reduce its costs and make it accessible to all.

1.1.2.1 The advent of *high-throughput* sequencers

In parallel with Sanger sequencing development, an independent method was discovered by Pål Nyrén and colleagues from the Royal Institute of Technology in Stockholm, using luciferase: in this protocol, ATP-sulfurylase converts pyrophosphate into ATP, which in turn acts as a substrate for the enzyme. It is remarkable for an emission of light that's proportional to the amount of pyrophosphate it catalyzes [57]. In practice, a template is attached to a solid support [58] and the DNA to be sequenced is flowed over it, and the correct base is inferred by measuring the amount of pyrophosphate *via* the intensity of light produced by the luciferase. This allowed the sequencing output to be directly detected instead of using electrophoreses, and the protocol did not require modified dNTPs [59] [60]. However, the intensity ceases to be proportional to the produced pyrophosphate after a few identical nucleotides are passed through [61]. This led to issues when attempting to sequence such long identical chains.

This method, dubbed *pyrosequencing* and illustrated in figure 6, was patented by 454 Life Sciences [62] and completely changed the way sequencing was thought of, as the

machines they released could run many reactions in parallel, thus allowing much higher amounts of DNA to be sequenced at a time [63]. Numerous improvements included the use of paramagnetic beads for the DNA to be coated onto and PCR-amplified; each bead is fits a well where the dNTPs are washed through [64]. In this setup, millions of such wells could be fit, containing that many beads, thus greatly reducing the necessary effort and cost to sequence ever longer stretches of DNA. This led to researchers coining the term of *high-throughput sequencing* (HTS) to reflect the several orders of magnitude gained in return-to-investment, and the era in which such developments were undertaken (often made concurrently with technological advances in other fields that made these breakthroughs possible in the first place) was referred to as *second-generation sequencing*, to contrast with the first one [65].

1.1.2.2 Industrial competition

454 Life Sciences, which would later be acquired by Roche, released the first commercially available machines designed with high-throughput sequencing in mind, such as the GS 20 or 454 GS FLX [66]. However, several methodologies and associated companies arose in the wake of 454's success. One of the more prominent ones is the Solexa sequencing technique, illustrated in figure 7. In this design, DNA molecules are surrounded by adapters; the molecules are then flowed through a field of oligonucleotides that are complementary to the adapters and affixed to flowcells. After a PCR, each DNA molecule processed this way is surrounded by identical molecules cornering the flowcell [67] [68]. Special fluorescent dNTPs are then used for the sequencing proper, where the fluorophore (or dye) occupies the 3' position and prevents further extension by the polymerase. The nucleotide is detected by exciting the dye with lasers, and the dye itself is removed before the next position is sequenced [69]. A DNA molecule and its replicates can thus been sequenced synchronously, one nucleotide at a time.

Early machines using this design, such as Genome Analyzer (GA), could only produce short reads but were among the first to yield paired-end data. The GA was followed by the MiSeq and HiSeq: HiSeq was designed for longer and more covered reads, whereas MiSeq was optimized for cost and run speed [70].

As the field of HTS opened and flourished, other companies started designing sequencing protocols of their own [13]:

- Ion Torrent was remarkable in that it relied on the pH difference caused by the release of protons during the polymerase reaction. [71] This was enabled by *complementary metal oxide semiconductors* (CMOS), a specific technology used in integrating circuits and microchips [70].
- Applied Biosystems also emerged through the second generation with its SOLiD (Sequencing by Oligonucleotide Ligation and Detection) system, using a ligase instead of a DNA polymerase unlike all of the above methods [72]. It was reported to struggle with palindromic sequences and produced shorter reads than Illumina technologies, although at cheaper rates [73].

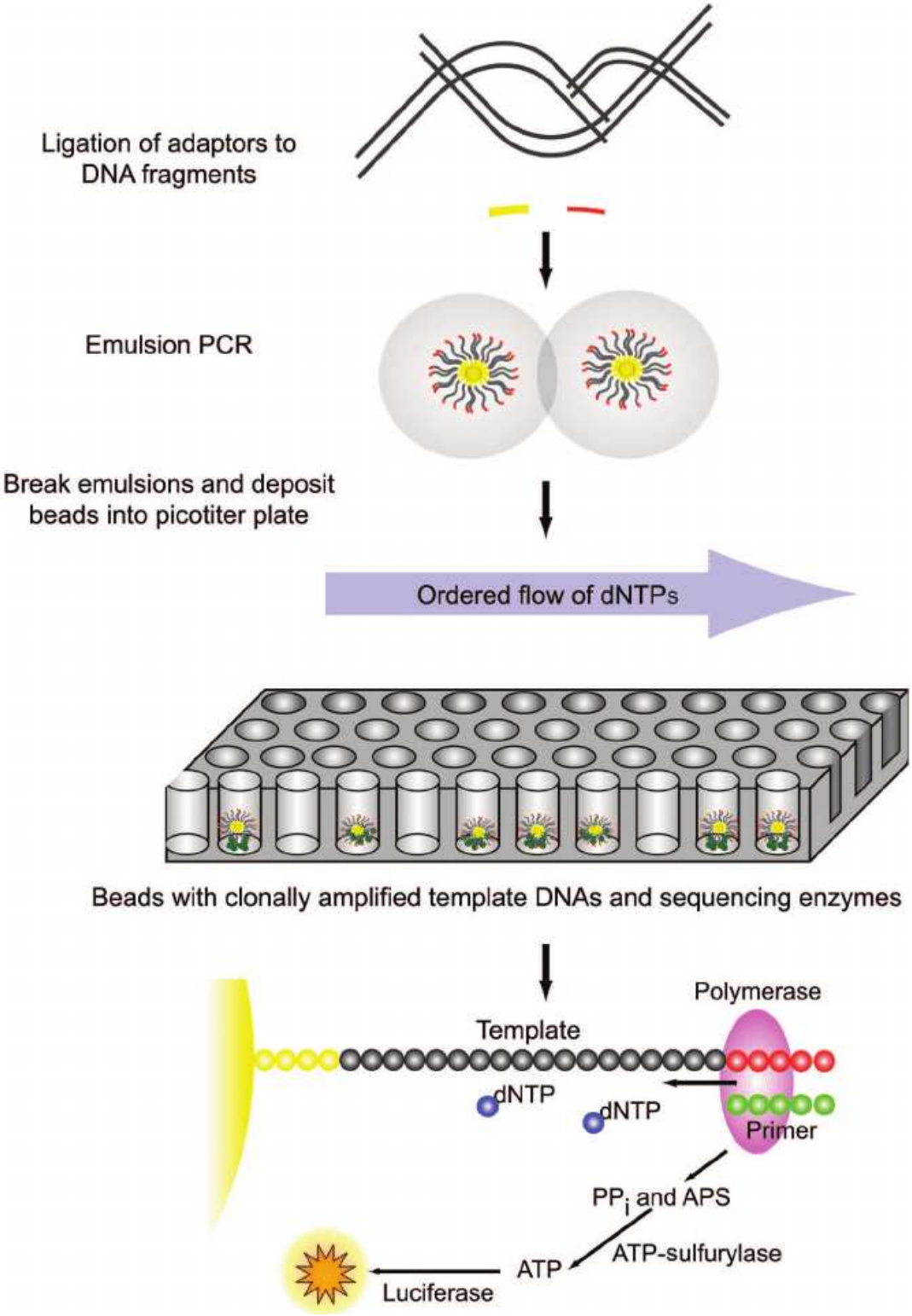


Figure 6: Pyrosequencing protocol by 454 Life Sciences.

Source: Adapted from Voelkerding *et al.*, 2008, [66]

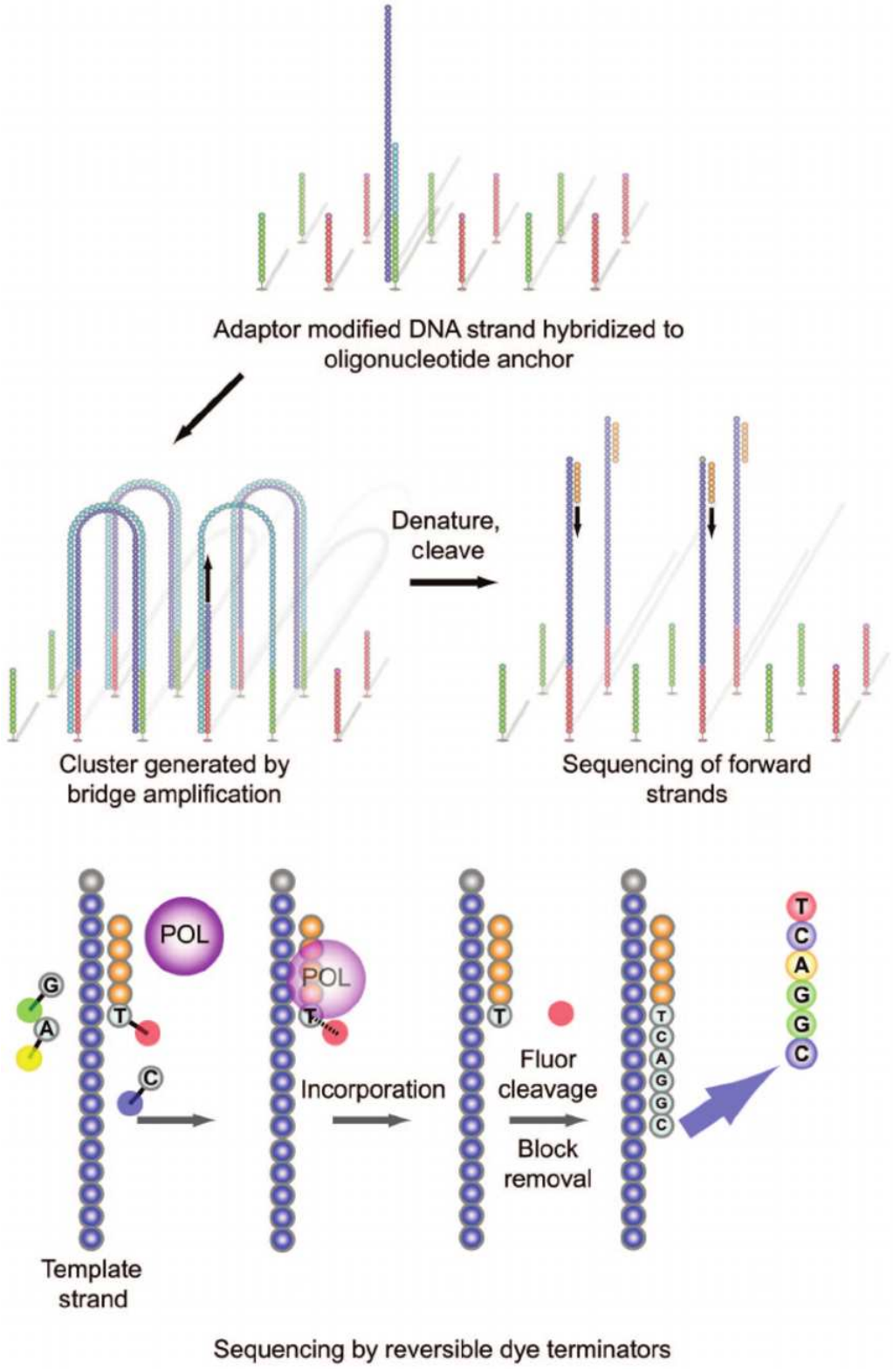


Figure 7: Illumina/Solexa sequencing protocol.

Source: Voelkerding *et al.*, 2008, [66]

- Complete Genomics used *DNA nanoballs*, whereby unknown DNA templates are flanked by known adapters; a round of PCR amplification produces long, linear chains of template-adapter monomers affixed to each other using rolling circle replication; they collapse into "nanoballs" on their own accord before being attached to a flowcell and sequenced in the usual fluorescence-based manner [74].

1.1.2.3 Aftermath

The above competition and breakthroughs in nucleotide sequencing technologies are often described as a *genomic revolution* in that they helped drive down prices immensely, at a several times faster rate than Moore's law that is usually associated with the costs of transistors, as illustrated in figure 8 [75]. Initially a costly and time-consuming endeavor, DNA sequencing became in less than a decade accessible to many labs. In the mid-to-late 2010's, however, the competition seems to have died down, with Illumina emerging as a clear winner and major contributor to the second generation [76].

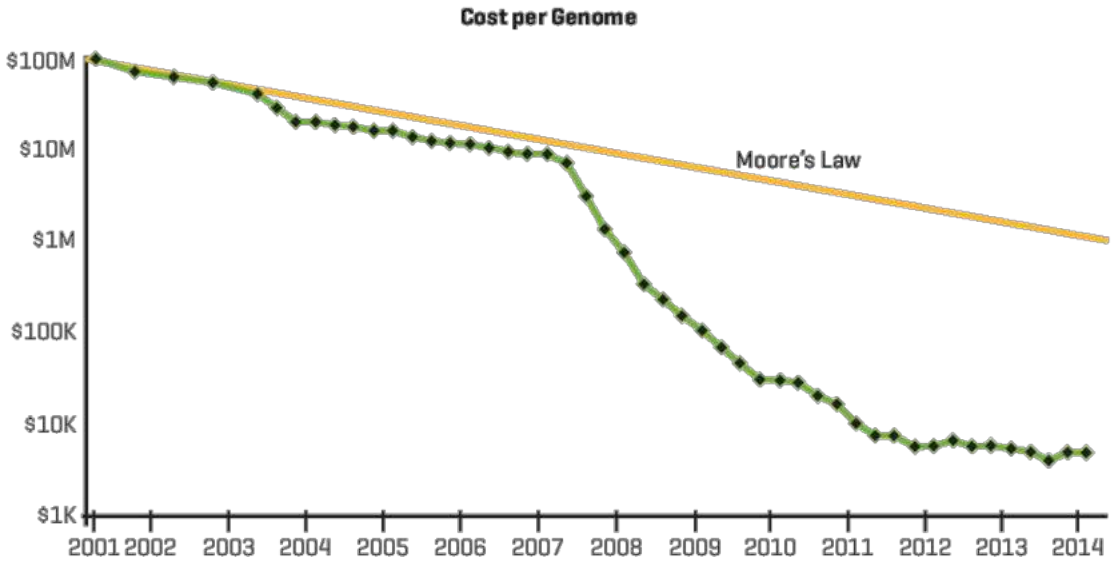


Figure 8: Evolution of sequencing cost compared to Moore's law.

Source: National Human Genome Research Institute (`https://www.genome.gov/sequencingcosts`), *via* Lippert *et al.*, 2015, [75]

The advances described above enabled the sequencing of many species, but the resulting assemblies often could not progress beyond draft form, especially for eukaryotes. The main reason is the presence of repeated sequences that create ambiguity and additional difficulty for traditional short-read based assemblers. Even to this day, Illumina short reads are rarely above a few hundred base pairs long, making them ill-suited for large stretches of repeats [73].

Another consequence of the advent and subsequent widespread access to cheap second-

generation shotgun sequencers is the emergence of the field of metagenomics [77]. It became possible to indiscriminately sequence everything in a live sample and analyze the results, often leading to the discovery of hitherto unknown species that could not be grown in laboratory conditions [78].

1.1.3 Third generation and beyond: long reads

The definition of *third generation* is muddier than the previous two [79], but there seems to be general agreement on its being hallmarked by the advent of single molecule sequencing (SMS) [80] [81]. This technology distances itself from the second generation by not requiring any PCR amplification. Early attempts at SMS had been made by Helicos BioSciences [82]. However, as of the late 2010's, two companies are currently leading the field: Pacific Biosciences (*PacBio*, which was acquired by Illumina) [83] and Oxford Nanopore Technologies (ONT, or simply *nanopore*) [84]. In this section we will briefly cover the technologies at stake.

1.1.3.1 Pacific Biosciences

PacBio machines make use of their single molecule real time (SMRT) platform. The polymerisation reaction occurs in special structures dubbed zero-mode waveguides (ZMWs), which are nanometer-sized holes in a film over a microchip [85]. As light goes through a hole of diameter smaller than its wavelength, it undergoes exponential decay. Therefore, only the bottom of the holes are illuminated. The sequencing makes use of laser-excited dye molecules, and these can be visualized in real time [13]. The protocol is illustrated in figure 9. The speed of the sequencing is very fast and equal to the polymerizing rate itself. Moreover, the molecules being sequenced are noticeably larger than second-generation ones and can reach 10 kb or more, and modified bases can be detected by this method [86].

1.1.3.2 Oxford Nanopore Technologies

Nanopore sequencing is in fact a derivative of general nanopore use geared at any kind of biochemical molecule [87]. In this design, the DNA passes through an ion channel, which halts ion flow. This can be measured by the current difference, which should be proportional to the length of the sequence itself. Very large sequences can be characterized this way. ONT has released a number of nanopore sequencing platforms such as the GridION, PromethION and MinION. The latter is remarkable for being very compact, as small as a normal USB device, and could be used in a decentralized way, further increasing the accessibility of sequencing to the masses. Current limitations include a high error rate, although improvements are being rapidly made [88].

1.1.3.3 Beyond: the fourth generation?

There is little consensus on what constitutes *fourth-generation* sequencing, but it has been used for designating single-cell technologies and *in situ* sequencing [89]. One of

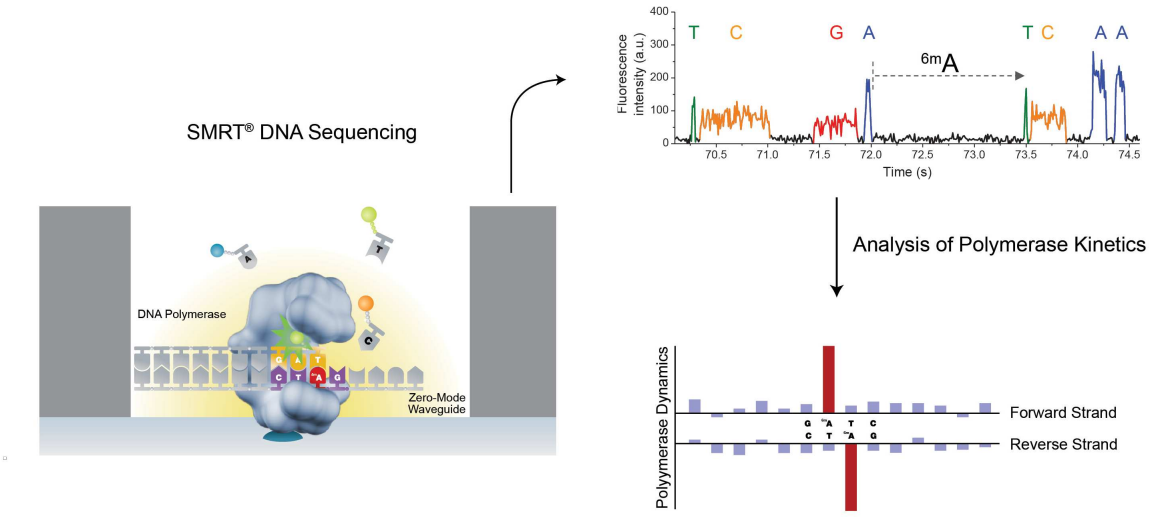


Figure 9: Overview of PacBio sequencing technology.

Source: Adapted from Rhoads *et al.*, 2015, [83]

the cited advantages would be to preserve the original spatial conformation of DNA and RNA molecules and benefit from increased (cellular) resolution, making it suited for cancer research. However, the efforts and technologies involved, such as *in situ* RNA sequencing (ISS) are still in their early stage and haven't reached industrial scale for everyday use yet.

1.2 Genome assembly

With the advent of shotgun sequencing, *i.e.* cutting DNA molecules into chunks, the field of genome assembly arose. It mostly piggybacks that of DNA sequencing technologies, and the underlying strategies naturally reflect the nature of sequence data being produced and the different forms it takes. Most traditional assembly algorithms rely on short accurate paired-end or single-end reads, whether they come from Sanger or second generation sequencing. These range from 30 bp to over 1 kb (especially in the case of Sanger sequencing). In the case of paired-end reads, a predetermined region of known length fixed by the sequencer, called the *insert size*, lies between both ends. Paired-end reads have a number of advantages over single-end data, in that they provide more mapping information, especially regarding repeated sequences [90] [91]. They also come in handy when detecting small-scale rearrangements, splicing or gene fusion.

Two different kind of assembly processes must be distinguished:

- *Reference-guided* or *mapping* assemblies are generated using a set of reads and a reference genome to go from, which is presumed to be similar to the final output. These assembly methods are fast and efficient [92] but they obviously make assumptions about the data and require a reference to be available in the first place. For the purpose of comparative genomics, they have proven successful for detecting rearrangements across similar species, such as the saker falcon, budgeridar and ostrich genomes [93].
- *De novo* assembly refers to the process of putting together sequence reads, often from multiple sources, to newly form the most complete and contiguous sequences representing a species' genome (or more in the case of metagenomics). *De novo* methods are orders of magnitude slower than reference-guided ones but are also free from bias. Also, they're often all one has at disposal.

In this work we will only focus on *de novo* assembly. Indeed, none of our case studies gave any prior knowledge at our disposal about the genomes we worked on. The following sections cover the different stages of a (meta-)genome assembly and their underlying principles.

1.2.1 Assessing assembly stages

The case of a few chromosomes within a single genome needs to be treated separately from metagenomics, and each one gets its own subsection.

1.2.1.1 Single genome

We roughly categorize the quality of an assembly into four states represented in figure 10, which we denominated according to current usage in the literature:

- *Contigs* represent an assembly in its most basic form, consensus sequences deduced from read overlaps. They are generally short but accurate, since their error rate is that of the sequencer itself.
- *Scaffolds* are ordered sets of contigs that have been determined to belong to the same chromosome. Each scaffold is represented by a single sequence, whereby the contigs themselves are separated by *gaps*, unknown sequences (represented by *N*s according to the IUPAC standard for ambiguous nucleotides) whose length can be more or less accurately estimated [94].
- *Chromosome-level* assemblies represent the next improvement stage, whereby all contigs have successfully been scaffolded such that there exists a one-to-one mapping between each chromosome and each scaffold, usually with the help of independent data. In practice, there are often a few short contigs that couldn't be assigned to a chromosome, but the term can still be applied provided those sequences only make up a small portion of the genome's total size. The terminology is applied at the discretion of the genome publisher and curators. Chromosome-level assemblies may still feature gaps.

- *Complete* or *telomere-to-telomere* assemblies have been completely and contiguously characterized, *i.e.* each scaffold uniquely maps to a chromosome and there are no gaps left. There are relatively few eukaryotic genomes in such a state: only model organisms such as *S. cervisiae* (and, through the efforts of the Génolevures consortium [95] [96] [97] [98], many other yeast species) or *C. elegans*. Complete bacterial genomes are more common [99]. However, some large genomes have been partially assembled this way, *e.g.* the human X chromosome.

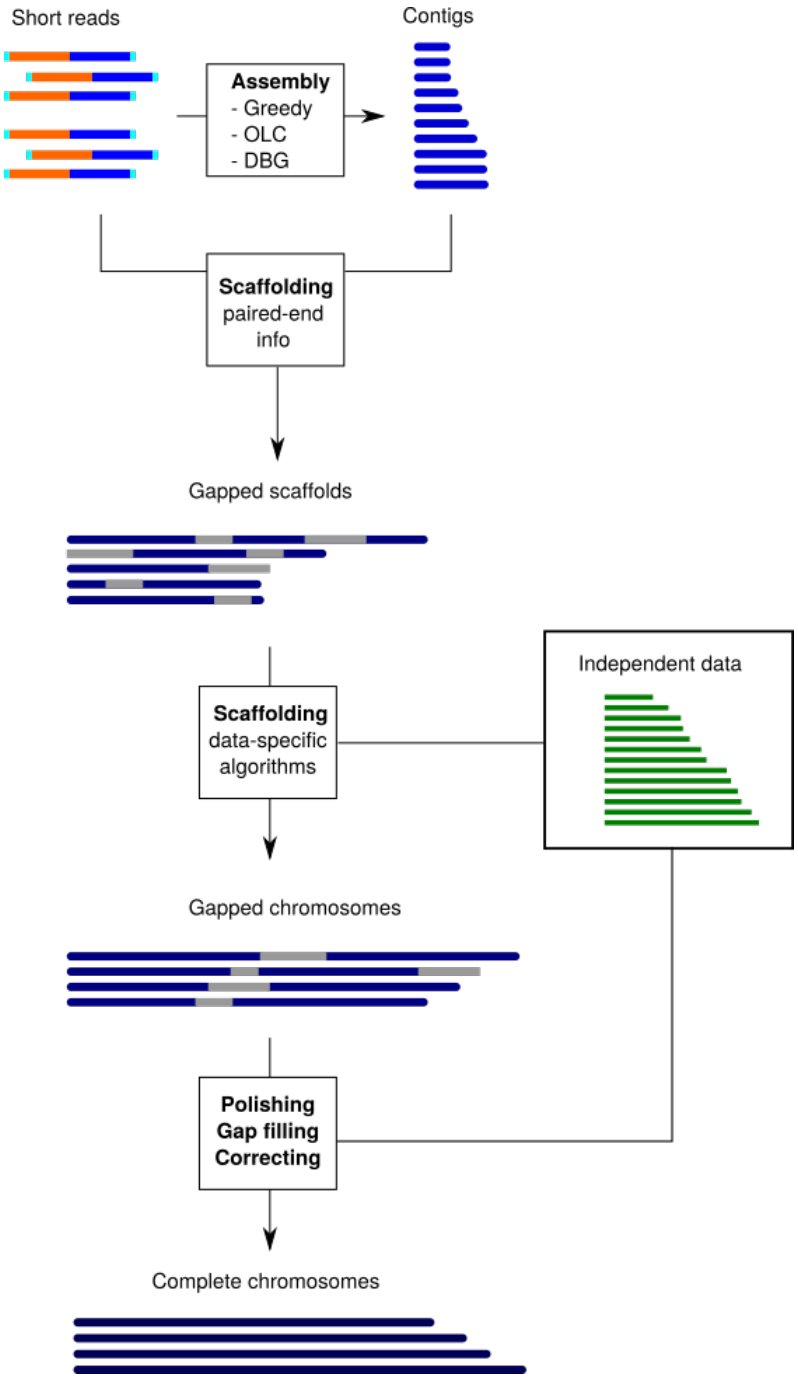


Figure 10: A scaffolding pipeline, from beginning to end. The four stages are *contigs*, *scaffolds* (usually gapped), *chromosome*-level scaffolds (usually gapped) and *complete chromosomes*.

1.2.1.2 Metagenome assembly

In the case of metagenomics, the picture is different, as the genomes themselves are smaller, being mostly bacterial, but much more numerous [100]. Gaps are seldom relevant, but uneven coverage and lack of material often means many genomes are incomplete and impossible to recover in their entirety. We roughly categorize the progression into three steps, illustrated in figure 11:

- *Metagenomic assemblies* are often incomplete, fragmented and act as preliminary steps for further analyses. Because of the uneven coverage across the genome spectrum, specialized software needs to be used, such as Bambus 2 [101]. Many popular assemblers have their metagenomics-specialized counterpart, like metaVelvet [102] or meta-IDBA [103].
- *Metagenomic binning* refers to the process of pooling contigs together so that each *bin* (or *metagenome-assembled genome*, MAG) contains sequences that belong to the same species. Bins are not necessarily complete (they don't represent the entirety of the genome), nor are they ordered; they are a next-best solution, absent complete reconstruction of every single genome in a metagenome. Nevertheless, with enough data from samples, thousands of new draft genomes have been succesfully published this way, greatly expanding the tree of life [104] [105]. Note that reads may be directly binned, skipping the assembly phase.
- *Scaffolding* is the full reconstruction of a species' genome within a metagenome. It is in practice very hard (and sometimes impossible) to fully reconstruct every single genome in a sufficiently complex sample, but many of the most highly covered ones can be characterized this way. This step is often skipped.

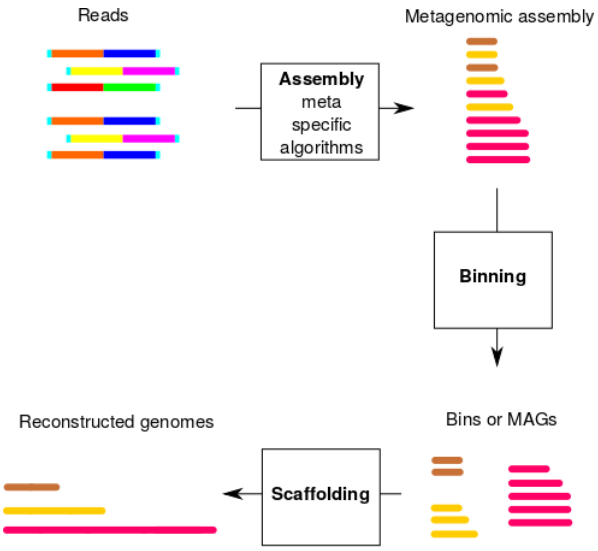


Figure 11: A complete metagenomic pipeline.

1.2.2 Base principle

In its simplest form, genome assembly is an offshoot of the shortest common superstring (SCS) problem: give a set of strings $\mathcal{S} = \{S_1, \ldots, S_n\}$, how to find the shortest string $SCS(\mathcal{S})$ such that each of the $S_1, \ldots, S_n$ is contained within $SCS(\mathcal{S})$?

Unfortunately, this problem is NP-complete [106], and there is no known algorithm running in efficient (polynomial) time for large inputs n that exactly solves this problem. In practice, the community uses heuristics to (hopefully) reach a satisfactory solution.

Current short-read based assembly tools can be broadly sorted into several categories, depending on the underlying fundamental algorithm behind it. Three of which are the most prominent [107]:

- Greedy methods
- Overlap-layout-consensus (OLC) based methods
- De Bruijn graph (DBG) based methods

Other approaches exist, such as methods based on string graphs or hybrid ones, but they have known comparatively less success, historical or current [108].

1.2.2.1 The Lander-Waterman model

Almost all assembly algorithms trace back to the original *Lander-Waterman model*, which first set in 1988 the initial mathematical basis for sequence assembly [109]. In that model, the sequencing depth or coverage c is assumed to be constant, as is the read length L. Moreover, a cutoff threshold T for the overlap length between reads is set, below which overlaps are discarded. If $\mathcal{G}$, the genome size, is known, the Lander-Waterman model is able to predict the exact number of contigs:

$$N_{contigs} = \mathcal{G} \cdot \frac{c}{L} \cdot e^{-c \cdot \frac{L-T}{L}} \tag{1}$$

The quantity $\mathcal{G} \cdot \frac{c}{L}$ is in fact the total read number, whereas the decaying factor $e^{-c \cdot \frac{L-T}{L}}$ corresponds to the probability that a read be the rightmost one within a contig.

Such a model gives an initial glance at the coverage c that would be needed given the strength of the overlaps (related to T) and the length of the reads L. In practice, L is fixed by the platform and c is constrained as well, so algorithms often depend on the overlap threshold T as a parameter [110].

The following methods can be more or less consistent with the model, from greedy algorithms (that often disregard T altogether) to overlap layout consensus and de Bruijn graphs (where an equivalence is often found).

1.2.2.2 Greedy methods

Greedy methods, the earliest assembly algorithms, are the simplest and probably the most naive, and consist in systematically merging the biggest overlaps every time [108]. Simplified steps can be described in Algorithm 1.

Algorithm 1 Simplified greedy *de novo* assembly

Require: $\mathcal{S} = \{r_1, \ldots, r_n\}$, a set of reads
Require: $f(r_1, r_2)$, a function merging r_1 and r_2 if they overlap
  repeat
    for r in $\mathcal{S}$ **do**
      $r_{greedy} \leftarrow \arg\max_{\tilde{r}} |\tilde{r} \cap r|$
      $r \leftarrow f(r, r_{greedy})$
    end for
  until no more overlaps are found

Greedy algorithms are not guaranteed to find the optimal solution, and the problem is exacerbated by the presence of repeated sequences or sequencing errors. They are also very computationally intensive, as they require calculating distances between all read pairs to find overlaps [107].

Examples of implementations of greedy algorithms include SSAKE, the first ever assembler [111], as well as VCAKE [112].

1.2.2.3 Overlap layout consensus

Overlap layout consensus methods proceed by three steps:

- First, a graph of *overlaps* is built. In that weighted graph, each node represents a read and each edge the length of the overlap. A simplified example of such a graph is provided in figure 12.
- Second, the graph is simplified so that redundant edges that could be inferred by simple transition across the graph are removed. Edges where the overlap is below a certain threshold (the T value in the Lander-Waterman model) are also discarded. One thus gets a *layout* of the graph. Deducing contigs then formally reduces to finding Hamiltonian paths (a path going through every node exactly once), which is an NP-complete problem [113].
- Lastly, all the reads making up a putative contig are aligned so that a consensus can be built by simple majority. Ideally this should be superfluous; in practice, sequencing errors and ambiguities created by ploidy or haplotypes make this step necessary [110].

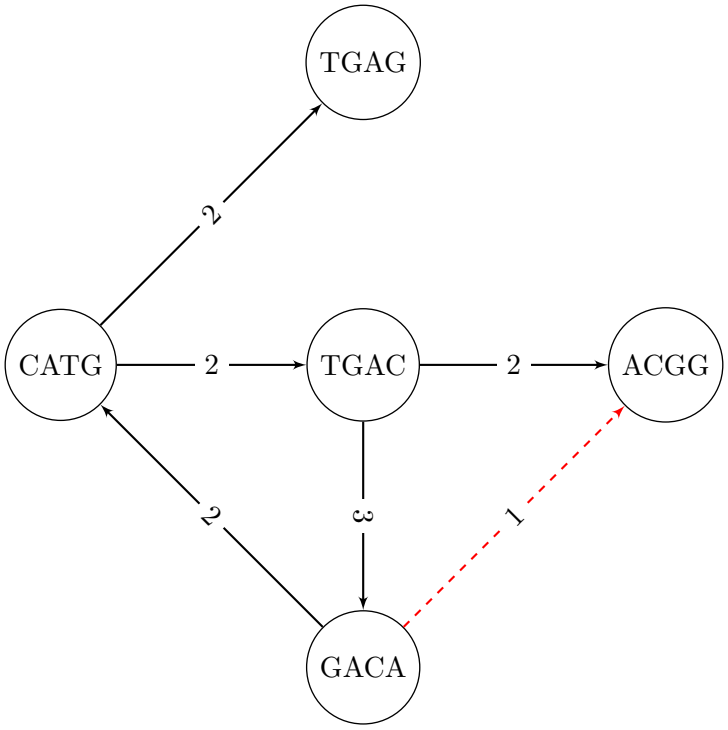


Figure 12: An overlap graph with five four-base reads. Insufficiently large overlaps (below $T = 2$) are marked to be removed.

In order to build the overlap graph, each read is mapped against each other read, which can be computationally intensive. The graph can either be constructed using *suffix trees* or using a scoring function when mapping two reads against each other [114]. The latter approach is more time-consuming but also more flexible since it allows for gaps and mismatches, which are simply given a lower score.

OLC is very consistent with the Lander-Waterman model, as the threshold used to build the overlap directly corresponds to the T parameter described in section 1.2.2.1. Due to their reliability, they dominated the first generation of sequencing technologies [110].

Examples of programs using OLC algorithms include PHRAP [115], TIGR [116], ARACHNE [117] or Celera (from Celera Genomics). The latter was notable for its extensive use during early genome projects such as the assembly of *D. melanogaster* [49].

However, OLC methods have a number of limitations: building the overlap graph can be slow, and the graph itself is huge: one node for each read, and edges grow even quicker. They tend to be ill-suited for second-generation libraries that often feature hundred of millions or even billions of reads [108].

1.2.2.4 De Bruijn graphs

De Bruijn graph (DBG) algorithms have risen to prominence with the advent of very large datasets and they are now the standard for modern assemblers [108]. DBGs take their name after their inventor, Nicolaas Govert de Bruijn. They are formal structures designed to represent overlaps exactly $k-1$ bases long, where k is the total length of the sequence, incidentally making them very suited for genome assembly [78]. A node represents a $k-1$ overlap relationship, whereas an edge represents a sequence of k bases, or k-mer. A simple example with four nodes is shown in figure 13.

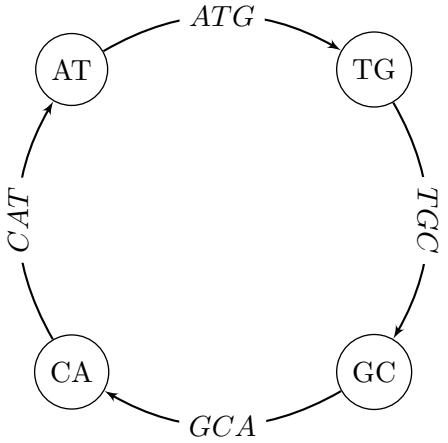


Figure 13: A simple cyclic, four-node de Bruijn graph.

In order to build the graph, a fixed k value is set, normally an odd integer in order to avoid palindromes between a sequence and its reverse complement. Then, each read of length n is converted into a set of $n-k+1$ such overlapping k-mers.

Contrary to OLC graphs, where overlaps are edges and sequences are nodes, DBGs reverses this relationship. This means that solving an assembly, which would previously reduce to the computationally hard problem of finding a Hamiltonian path (going through all nodes exactly once), is instead transformed into the much easier problem of finding an Eulerian path (going along all edges exactly once) [110]. There are many algorithms for finding Eulerian paths, running in linear time. An example of a DBG, with its path outlined, is provided in figure 14.

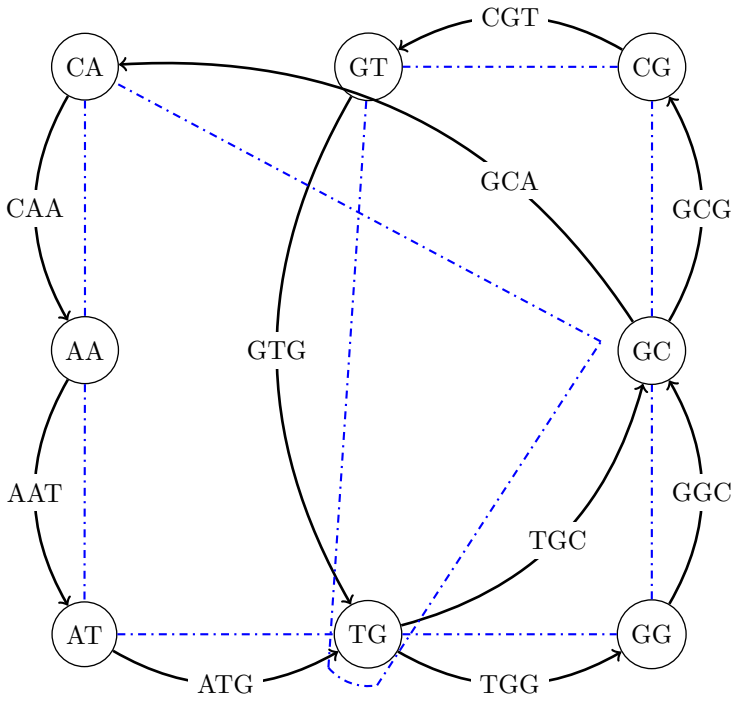


Figure 14: A de Bruijn graph and its Eulerian path.

It can be shown [118] that under certain conditions, an Eulerian path always exists, is unique and corresponds to the initial genome. Unfortunately, real world datasets often don't benefit from these conditions. DBG algorithms struggle with various issues [119]:

- Sequencing errors
- Repeated sequences
- Haplotypes
- Pronouncing *de Bruijn* correctly [120]

Apart from the last point, many efforts have been undertaken to address these issues.

Sequencing errors result in spurious reads (and k-mers) that "stick out" from the graph, create *bubbles*, and make it non-Eulerian. Apart from using the sequencer's own quality scores to weed out such issues, they can be avoided by performing read error correction prior to building the graph: since sequencing errors tend to be rare, the unusually low frequency of the resulting k-mers can be noticed, especially when compared to that of their (presumably genuine) closest k-mers (in terms of Hamming distance). Read error correction is thus a customary step in DBG-based assembly. Examples of implementations include ECHO [121], BayesHammer [122], or SHREC [123]. Since the step is so important, many DBG-based assemblers ship with a corrector, whether it be

third-party software, such as SPAdes [124] (using BayesHammer) or their own implementation, such as EULER-USR [125].

Repeated sequences make multiple paths possible, or a repeat collapse, thus creating misassemblies. This can be handled using *multi-graphs*. In these data structures, nodes (overlaps) can be linked by m edges (k-mers), where m is the multiplicity of the k-mer being repeated. Determining it can be difficult, especially given that coverage can be uneven across a genome. Without that information, paired-end reads can still give an estimate of the length of the total repeated stretch (and thus its multiplicity), provided the insert size is long enough [78].

Resolving haplotypes is a long-standing issue in DBG-based approaches, combining both problems about bubbles (heterozygous sequences and polymorphisms falsely marked as errors) and repeated sequences stemming from identical regions. Recent developments have been made to address it, notably with the BWISE assembler [126] making use of *super-reads* as more complex data structures, analogous to the ones used by long-read assemblers (see section 1.2.3).

DBG-based assemblers have taken over the field of genome assembly and most state-of-the-art assembly software are DBG implementations. These include Velvet, one of the earliest such programs [127], SOAPdenovo [128] [129], Abyss, [130], and more recently, SPAdes [124], among the most well-known ones.

1.2.2.5 Scaffolding

The above methods were concerned with creating the most accurate *contigs*. However, contigs are seldom sufficient to reconstruct the entirety of a genome; this is mostly due to repeated sequences that can't be bridged, or unsequenced regions. Scaffolding is thus the process of ordering the contigs despite the presence of such gaps [94].

The advent of new technologies has greatly increased the avenues for scaffolding a genome, and we will treat them below. In the case of short reads, however, the main information comes from paired-end data. Since nearly all modern short-read based sequencers produce paired-end reads, an initial scaffolding can be generated right away and indeed most assembly programs do provide their own scaffolding steps .

The principle is illustrated in figure 15: if both ends of enough read pairs successfully map on different contigs, one may infer that both contigs should be next to each other. The orientation of the reads gives information about which way to orient both contigs.

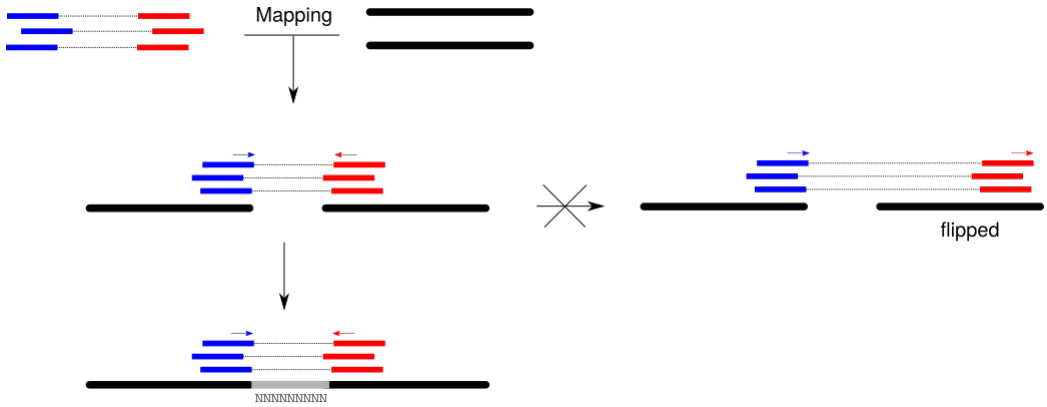


Figure 15: Scaffolding using short paired-end reads. The orientation of both ends help disambiguate that of the contigs within a scaffold.

The sequence in-between is still unresolved and traditionally represented by stretches of ambiguous `NNNN` on a FASTA file, but the length of such gaps can be estimated. It is naturally bound by the insert size of the paired-end reads, and enough alignments at the edge of either contig may shrink the bound further.

Unfortunately, some gaps are longer than the insert size of most sequencers, and are thus impossible to bridge this way. This is particularly true for the genomes of many eukaryotes where repeats may span hundreds of megabases. This is where long reads and other such new technologies come into play.

Apart from assembly software's own implementations, examples of standalone scaffolders include SSPACE [94], GapFiller [131] and ECHO [121].

1.2.3 Long read and hybrid methods

With the advent of long read sequencing technologies, new methods were necessary to take this data into account. Long-read assemblies are crucial for bridge large stretches of repeated sequences [132]. However, due to the still error-prone nature of these sequences, long reads are often coupled with second-generation short reads to correct (or *polish*) them; hence the *hybrid* nature of tools attempting to combine both kinds of data to solve an assembly. An example of a complete pipeline is illustrated in figure 16. Note that not all steps are necessarily present and most programs combine two steps in one (*e.g.* read correction and assembly).

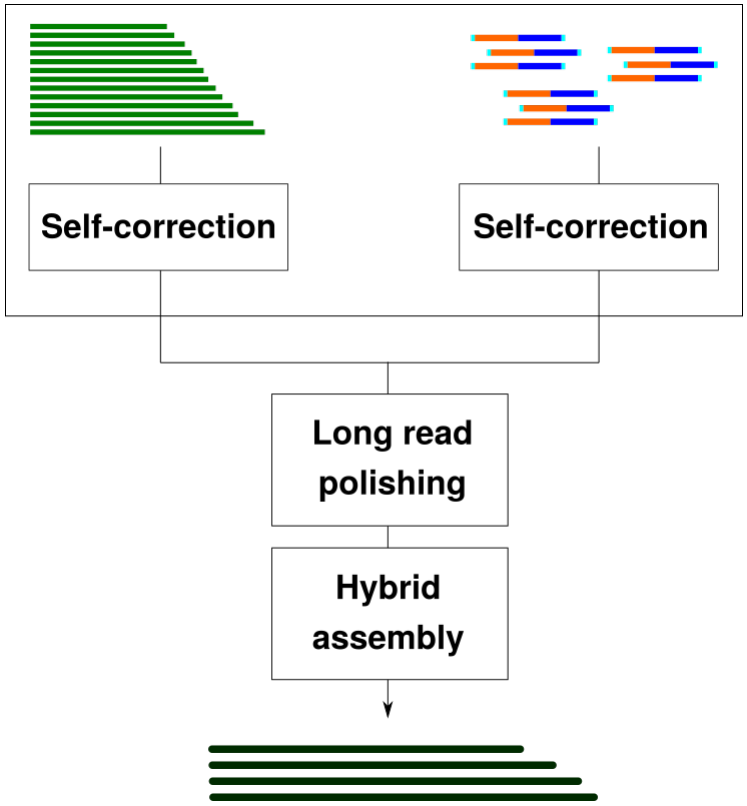


Figure 16: Hybrid, short- and long-read based assembly pipeline. Reads are (optionally) cleaned using error correction algorithms, then the short reads are used to correct the long ones. Both read sets are used for the DBG assembly program.

Read error correction and polishing is a time-consuming task, as it either requires mapping the short reads onto the long ones to find overlaps and potential errors, or building complex DBGs to find correct paths [133], and it is often the bottleneck of such pipelines. Example of long read polishers using short-reads include FMLRC [134], Sprai [135], PBcR [136] or LoRDEC [137]. This step is sometimes avoided in favor of long read self-correction with programs like LoRMA [138] or HGAP [139], which doesn't require a short-read library.

Like modern short-read assemblers, hybrid and long-read assemblers generally use DBGs and try to find Eulerian paths to get contigs. Among the most prominent programs are Canu [140], hybridSPAdes [141] or MaSuRCA [142] which have seen increased and successful use as of late in many assembly projects such as the genome of the clownfish [143] or the tropical teak tree [144].

1.2.4 Metagenome assembly

This section is focused on the special case of metagenomics and field-specific algorithms that have been developed to investigate genome assembly in a complex community and

the underlying chromosome dynamics.

1.2.4.1 Challenges in metagenomics

The "holy grail" of metagenomics is to be able to accurately reconstruct and characterize the genome of every single species from a given sample, as shown in figure 17; however, in setting out to do so, a number of assumptions made in the previous sections have to be discarded, making the task even more challenging.

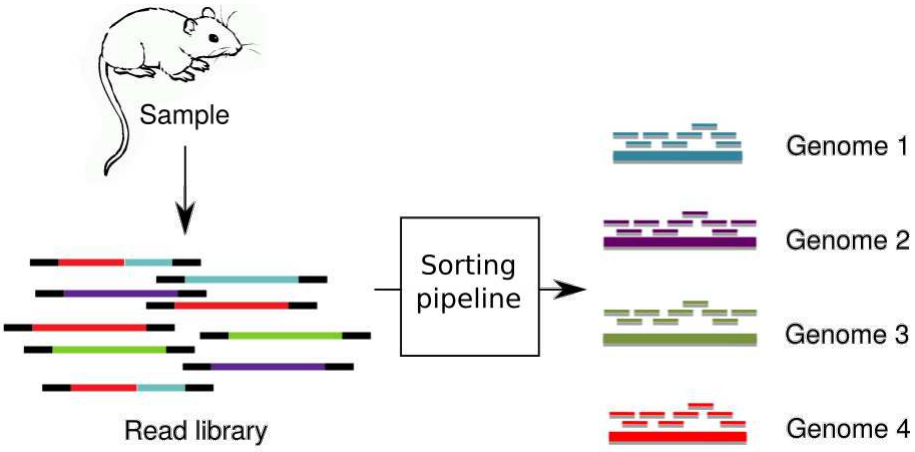


Figure 17: An idealized metagenomics pipeline. DNA is extracted from collected samples, indiscriminately sequenced and accurately sorted into groups so that their constituent genomes can be reconstructed.

First, the diversity found in many metagenomic samples is enormous [145]. Communities of all kinds thrive under very different conditions [146]. It is estimated that the human gut microbiome presents a thousand different species of bacteria [147]; a typical soil sample contains more than ten times that amount [148] [149]. Moreover, the interactions underlying these very complex communities are still to be fully understood, but a number of findings indicate that they hold crucial roles in the maintenance of various ecosystems, from oceans to soil to living hosts; within animals (including humans) and plants, they contribute to their metabolism [150] [151] [152]. They have also been known to alter behavior in humans [153]. Unfortunately, more than 99% of the species found in these communities can't be cultured in lab conditions [154], so very little prior information about the genomes is available.

Second, the coverage distribution across these organisms is very much unequal. In practice, a few species or genuses of bacteria are overabundant and drown out the signal from the remainder. This doesn't mean the rarer bacteria aren't any less essential to the balance of the ecosystem. This heterogeneity is compounded by the presence of multiple similar strains sharing many of the same sequences, much like the presence of haplotypes within an eukaryotic read set complicates traditional genome assembly.

Third, a number of sequences may be shared among species for various reasons:

- Conserved by evolution (*e.g.* essential genes)
- General species similarity leading to sequence homology
- DNA transfers such as conjugation
- Phage infection (either integrated or only within the cell compartment). Bacteria that are part of the same phage infection spectrum will share their viral DNA.

This shows that the characterization of a metagenome (and how it affects an ecosystem) cannot be decoupled from an understanding of the underlying dynamics at play among its constituent species.

1.2.4.2 Deconvolving a metagenome

As we have seen in section 1.2.1.2, two distinct steps are typically undertaken in practice:

- *Metagenome assembly* reproduces the steps described in 1.2.2.4 involving de Bruijn graphs in order to produce high quality contigs that are as contiguous as possible, while taking into account the extra constraints regarding coverage and sparsity. Current state-of-the-art programs include metaSPAdes [155], MEGAHIT [156] [157] or IDBA-UD [158], notable for their use of multiple k values when splitting reads into k-mers.
- *Metagenome binning* refers to the grouping together of sequences according to their species (or genus, etc.). These so-called *bins* are unordered collections of contigs that presumably belong together. In order to do so, a number of assumptions are made about the sequence composition, and each assumption determines a class of binning algorithm [100]. The methods are described as follows.

Coverage-based binning The first kind measures the coverage of each sequence, and reasons that two sequences having the same exact coverage are more likely to belong to the same genome than not (i.e. by chance alone). Sequences are thus sorted and grouped according to their relative abundance. In order to make the method even more robust, and given that sequencing now comes rather cheaply, modern tools draw from multiple sample libraries and track how the coverage of each sequence evolves from one sample to the next. A simplified example is illustrated in figure 18. Therefore, they are often called *coverage based binning* (respectively *differential coverage*) methods [159] [100].

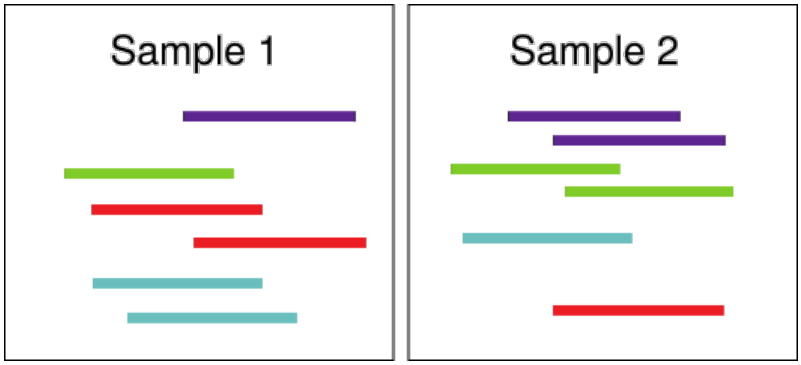


Figure 18: Differential coverage based binning. From sample 1 to sample 2, the green and purple reads see the same variation in abundance (1 to 2), as do the red and teal bins (2 to 1), thus giving hints that they belong together.

These are reasonably robust and have known increasing success [160]. They work well in the case of more abundant bacterial genomes. However, they may struggle when it comes to binning rare sequences, because the signal-to-noise ratio is lower and the above assumption is more difficult to follow when so many sequences only appear a few times. Moreover, the approach also encounters difficulties with edge cases such as repeated sequences or a highly uneven coverage across the genome itself, as is often the case in bacterial genomes where multiple instances of replication take place [161] [100]. These caveats all but make the approach fall short for many rarer genomes of interest.

Examples of coverage-based binning pipelines include GroopM [162] or BinSanity [163].

Composition-based binning The second kind reasons that nucleotide composition is generally uniform across a bacterial genome, and thus any sequence of k bases (k-mers, e.g. *tetranucleotides*, *pentanucleotides* for $k = 4, 5$) is expected to have a more or less constant frequency. For each sequence, a vector of frequencies is extracted (for instance, in the case of $k = 5$, such a vector has $4^5 = 1024$ coordinates representing each possible pentanucleotide) and sequences are then sorted according to their closeness in that feature space. The approach isn't limited to a single type of k-mer and can indeed incorporate any additional feature related to the sequence composition (*e.g.* GC content), such that a general n-dimensional representation of the sequence can be drawn and clusters be formed.

Figure 19: Composition based binning. In a simplified example, the teal and purple reads feature a common tetranucleotide (`GAAT`), as do the red and green reads (`TGCA`), indicating that each pair belongs together.

The approach is powerful but may fail when the assumption of uniform sequence composition isn't verified anymore. Abrupt changes in composition can occur in the event of DNA transfers, such as conjugation (see figure 20). Moreover, phages often drastically differ in sequence composition from their host, so a k-mer based approach would not accurately capture such dynamic events.

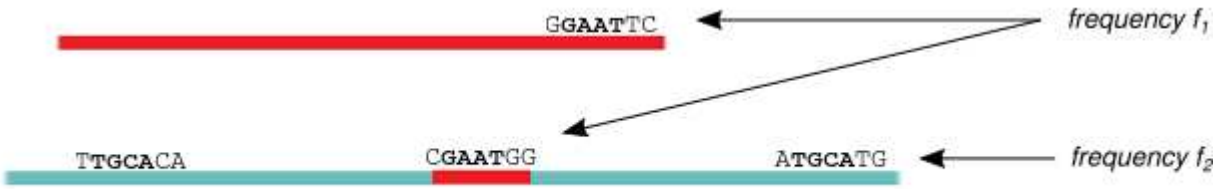


Figure 20: Illustration of sequence composition heterogeneity. Due to the dynamics of DNA transfers, a genome may have abrupt variations in sequence composition across its chromosomal regions.

Exemples of composition-based binning programs include BusyBee [164].

Hybrid and other methods Most recent state-of-the-art binning algorithms use a hybrid approach that makes use of both methods, as well as additional sources such as marker genes. These include Metabat [165], CONCOCT [166], MaxBin [167] [168], CoMet [169] or COCACOLA [170]. Some tools act as a synthesis of the above tools in order to maximize the number of genomes reconstructed this way, such as DasTool [171].

In practice, these tools typically pool all that information into feature vectors, compute pairwise distances between these and perform a (generally unsupervised) clustering algorithm such as affinity propagation (in BinSanity, [163]), DBSCAN (in CoMet, [169]) or spectral clustering (MyCC, [172]). Another approach makes use specific data structures to compute distances between coverage profiles and partition them, called *eigengenomes* [173]. Very recent methods involve the use of deep learning on the signature of the genomes [174]. By combining many different approaches, the goal is to avoid some of the bias inherent in each, but it remains present.

1.3 Genome validation and curation

Once a (meta-)genome is assembled or scaffolded, the question of validating it naturally arises. With no prior knowledge about the reference genome, independent ways of obtaining data are often necessary to ascertain the information about the genome's sequences and order. Nevertheless, many rudimentary or error-prone assemblies are liberally published on public databases, so that they can be improved or otherwise corrected later. Indeed, many modern assembly tools and methods are developed by testing them against existing genomes: the very solid ones act as a benchmark, while a correction of misassembled ones act as added value.

The correctness and validity of assemblies has been the subject of considerable interest this decade and many concerted efforts have been set up to try and benchmark one's tools, among which the most prominent are GAGE [175] and competitions such as the Assemblathons [176] [177].

On the data supplying side, researchers have taken advantage of the recent accessibility of assembling technologies to set out and build a comprehensive database of high-quality genomes, notable consortiums being the 1000 Genome Project (focusing on human genomes) [178] [179] the 10k Genomes Project (focusing on vertebrates) [180] [181] [182] [183], and invertebrate-focusing consortiums such as the Global Invertebrate Genomics Alliance (GIGA) [184] [185]) and the Arthropod Genomics Consortium or i5k project [186]. The stakes and insights to be gained into such a wealth of information make it well worth focusing on the curating and validation of one's genomes.

Genome assemblies can be assessed:

- Absolutely, by ensuring they are correct with external and independent sources.
- Relatively to one another, through the use of carefully tailored metrics for comparison.

1.3.1 Ensuring correctness

It is important to note that there is no such thing as a perfectly assembled genome. Once one ventures outside the realm of specific cell lines of model organisms, genome assemblies almost always feature a number of errors, gaps, unincorporated sequences and other such misassemblies. Nevertheless, a number of verifications make it easy to ensure an assembly isn't inconsistent with available data. These range from basic checks to integrated data validation from independent sources.

1.3.1.1 Basic verifications

Contamination A trivial preliminary way is checking for contamination or otherwise foreign sequences that have somehow gotten through the whole assembly pipeline [187]. A variety of tools have been developed for this purpose and typically act as a first screening upon genome publication [188], an example of which being PhylOligo [189].

Annotation transfer Another intuitive method is to verify that all previous annotations made on a draft can be successfully transferred to the improved assembly. This change of coordinates is sometimes metonymically known as *liftover*, due to the public tool made available by the University of California Santa Cruz (USCS) at `http://genome.ucsc.edu/cgi-bin/hgLiftOver`. Other tools have been designed to that effect, such as CrossMap [190], notably used by Ensembl. One essentially converts one set of genome coordinates to the other, with the possibility of raising red flags when a significant number of features could not be transferred. Annotations and otherwise characterized genome features take many different forms (gene models, expressed sequence tags (ESTs), or even plain introns, exons, known centromeres, etc.) are often short in size, but

relatively widespread. They are especially crucial given that assemblies are often used for comparative or evolutionary purposes, where the analysis of synteny blocks (and thus gene orders) depends on the ability to transfer these features successfully [191].

Detecting misassemblies Some misassemblies can be cleaned up right away with mapping issues. Shotgun reads, or a reference genome if available, act as preliminary filters to detect false breakpoints, translocations or inversions. Early aligners were relatively slow and resource-consuming, especially for genome-wide pairwise mapping, making the process nontrivial; with the advent of modern mappers such as minimap2 [192], the task has been considerably simplified, and it is now part of standard assembly validation pipelines such as QUAST [193] (or its large genome counterpart, LG-QUAST [194]) or Reapr [195].

Miscellaneous There also exists a number of rough indicators that will give a global outline of the analysis and raise warnings if the assembly went wrong; if a karyotype is available, the number of chromosomes should coincide with the main scaffold count. Their relative size should also match, as well as their respective centromeric ratios if applicable (assuming the centromere repeat pattern is known) [191].

1.3.1.2 Independent data integration

Once basic checks are made, ensuring the global integrity of the genome structure is somewhat respected, more data from independent sources is needed to validate an assembly on a finer level. In this section we will summarize current technologies that have been successfully used in assembly projects.

Genetic maps Genetic maps are a convenient way of ensuring sequence order is respected [196]. The technology is mature and it has been at the basis of many assembly projects, including ones in our current thesis work. It is essentially based on linkage desequilibrium (LD) data, whose extrema characterize so-called recombination hotspots and coldspots among DNA strands. Additionally, linkage groups often give a strong indication as to the genome structure, as they almost always map one-to-one to the chromosomes. As such, they are a good source to validate contig ordering for the purpose of genome scaffolding [197]. Examples of scaffolding software using genetic maps include ALLMAPS [198].

Long reads As we have seen, long reads are now the standard for genome validation and polishing. They have shown promising success, especially for notoriously large and complex genomes such as that of plants [199] [200] such as the tomato or wheat genome [201] [202]. When they are not used as a baseline for the assembly process (*i.e.* be used in the assembly graph or data structure for a *de novo* genome) they may fill gaps that are unaccounted for [203]. There is no exact guideline on whether one or the other approach should be preferred, as both have been used successfully. Presumably, long reads with

high coverage and quality are better suited as a baseline material for the scaffolding process. Alternatively, strategies have been developed to merge assemblies, *e.g.* software such as Metassembler [204] will, given two genomes, use the second one to correct the first. Other assembly merging programs such as CAMSA treat all assemblies equally and may be used for merging more than two genomes this way thanks to multi-mapping [205]. With the proliferation of assemblies from different sources and the difficulty of integrating them all, many tools have been developed to attempt to merge large-sized genomes, all with some level of success and with no clear consensus best solution [206].

Lastly, if the coverage and accuracy of long reads are insufficient to yield a high-quality assembly by themselves, the reads themselves can still be used to fill some gaps when applicable with the use of pipelines such as PBJelly [207]. This was done for the assembly project of the honey bee [208] or the black raspberry [209].

RNAseq RNAseq reads are a very robust way of polishing transcribed regions in a genome. This is especially useful since these are usually the regions that are the focus of subsequent analyses. The obvious drawback is that regions that are not transcribed are unaccounted for, and these are usually the ones traditional assembly methods already struggle with (*e.g.* repeated sequences).

Linked reads Linked reads are a novel sequencing method developed by 10X Genomics. Long DNA molecules are partitioned and amplified, and all fragments derived from a single long molecule are tagged with a uniquely identifiable barcode. This way, two distant loci can be rightfully assumed to belong to the same chromosome. Linked reads dispense with increased coverage and focus instead on the breadth of the molecule being sequenced, so the individual long molecules are not fully sequenced. This does not result in a single long read molecule but still lets one reconstruct large-scale haplotypes, call structural variants or improve assemblies [210]. Linked reads have known growing and successful usage in large assembly projects such as that of the sperm whale [211], the Sitka spruce [212] or in metagenomics with deconvolving software such as Minerva [213].

Optical mapping Optical mapping uses a restriction map (called an *optical map*) of a single elongated DNA molecule placed under a microscope. The DNA is digested by a restriction enzyme and the resulting fragments are stained with dye. The intensity of the fluorescence determines the size of each fragment. The main draw to this technique is that it preserves the order of fragments and doesn't need any amplification. As such, it has often proven successful in genome scaffolding projects, especially for repeat-rich genomes such as, yet again, plants [199]. They are also useful when validating assemblies [214].

Hi-C Hi-C is increasingly becoming a crucial technology for genome scaffolding. Since Hi-C based assembly is the basis of our present work, we will detail the framework in its own section (see section 1.4.7.1).

1.3.2 Validation metrics

Validation metrics are typically used to compare several assemblies, or to verify that one's pipeline, whether it involved optical maps, Hi-C, etc., actually improved an assembly from a reference.

1.3.2.1 Size distribution metrics

These inform about how much of an assembly or scaffolding is contained within a few molecules. Ideally, the entirety would be found within a few that correspond to actual chromosomes. In practice, a range of so-called Nx/Lx metrics are in use, illustrated in figure 21:

- N50 is the length of the scaffold below which all greater scaffolds do not make up more than 50% of the total assembly in size. Same goes for N90, N20, etc. Notably, N0 is the largest scaffold and N100 is the smallest scaffold.
- L50 is the index of the scaffold of length N50 (starting from the longest).
- If the size of the original genome is known (or if a putative reference is available), NG50 measures the above 50% ratio with respect to that original size (as opposed to the total size of the assembly). Same goes for NG90, LG50, and so on. This more refined terminology comes from QUAST [193], the *de facto* standard for genome validation software.
- If a reference genome is available, the misassembly count can be paired with an additional set of metrics, called NAx: the genome is broken down into contiguously aligned regions as though they were separate contigs and an N50 is computed from these.

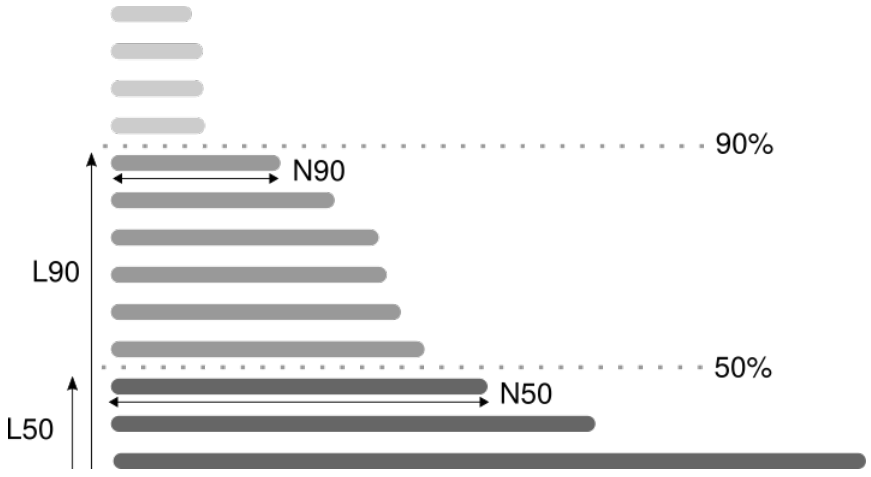


Figure 21: Illustration of assembly metrics.

Generally, the higher one's N50, the better (barring spurious fusions). Recently, the Vertebrate Genome Project (VGP) set new quality standards, whereby released genomes must, among other statistics, have an N50 greater than 1 Mb for contigs and 10 Mb for scaffolds [215].

1.3.2.2 Completeness metrics

Completeness tells us to what extent a DNA sequence looks like an actual genome, based on its genic composition. In practice, one checks for the presence of expected known genetic markers found in the genome. They are used as a benchmarks and depend on the species being examined and its position in the tree of life. Other composition-related information such as k-mer content is also assessed.

Completeness metrics are usually paired with *contamination* metrics: while one indicates how many features are lacking, the other indicates what sequences are extraneous and should be removed. *Contamination* here is to be taken in a relatively loose sense, distinct from the one covered in the previous section, as it also includes duplicated and extra copies of genetic markers.

Gene content The most prominent validation method in the literature is gene completeness. Each distinct lineage features a number of unique, single-copy genes that are conserved across a significant portion of the lineage's species. Going back through the tree of life, a more general (and thus smaller) set of markers can be identified, and so on. Completeness validating tools therefore search the genome for the presence of such very general markers, generally using Hidden Markov Model profiles with gene prediction software like prodigal [216] [217] or Augustus [218] [219] [220]. When given specific lineage information (or if they are able to infer it), the tools refine their search and statistics with more markers. Each lineage has its validation tools and marker databases, and the following are currently considered standard within their own researcher communities:

- Bacterial and archaeal genomes are usually validated with CheckM [221], taking advantage of the conservation of 43 conserved, single-copy marker genes among 97% of all publicly archived genomes (initially pulled from the IMG database [222] known for its trusted reference genomes). They are mostly genes coding for ribosomal proteins and RNA polymerase domains. Additionally, for each bacterial lineage, additional single-copy genes respecting the 97% criterion are also considered markers. CheckM identifies missing (general or lineage-specific) markers, duplicates, and thus yields a completeness and contamination report. As per CheckM's terminology [105], near-complete genomes have a completeness 90% and a contamination 5%; medium-quality genomes have a completeness 70% and a contamination 10%;and partial genomes have a completeness 50% and a contamination 4%. Its rise to prominence has made it the standard for any metagenomic validation.
- Eukaryotes are usually validated by BUSCO [223]. The tool looks for benchmarking universal single-copy orthologs (or BUSCOs, hence the name) specific to a

lineage. Many such sets are publicly available on the Ortholog Database (ODB) [224] and get updated [225]. For instance, there are 303 such orthologs for eukaryotes as of ODBv9. There are also sets for bacteria, although they tend to be less widely used than CheckM. BUSCO scores often don't reach 100%, even for reference genomes; rather, the reference should act as a benchmark for other assemblies. Another completeness tool, CEGMA [226] [227] used to be formerly widespread but it has since then been largely superseded [228].

- There are a number of databases for specific lineages: the PLAZA platform publishes core gene families of its own for plants [229], and the Fungal Genome Mapping Project (FGMP [230]) also provides databases for validating fungal assemblies. BUSCO remains extensively used in both cases as well [228].

k-mer content Estimating a genome's k-mer content distribution can be useful if genomic information is already available for the species or neighboring ones, as it gives an idea of what sequences are missing or superfluous. The distribution should match, or be close to, that of the reference (if one is available), or that of reads from different sources (with respect to coverage). It should also be close to the k-mer distribution of neighboring species in the tree of life [231]. If there are any contaminants, they usually stand out in the distribution. Since computing k-mer statistics is such a common quality-control task, various tools have been developed, such as KMC [232] [233] and KAT [234], and they are integrated as part of larger pipelines such as QUAST.

Repeat content Most eukaryotic genomes have repeated sequences such as transposable elements. The profile, quantity and composition of these repeats has been extensively studied and a consensus sequence can be reconstructed for a family of repeats, leading to the creation of the Repbase Update database [235] [236]. Using repeat detection tools such as RepeatMasker [237] or Red [238], one may verify that the purported species' repeat content and nature matches that of the database.

1.4 Our framework: chromosome conformation capture

In the light of all the aforementioned technologies and their current direction, our approach tackles the genome assembly problem in a complementary angle, rather than superseding any of these methods. In this section, we will articulate the main principle of our technological framework - Chromosome Conformation Capture (3C) - as well as how it has been successfully applied to solve a variety of biological questions related to chromosomal architecture and functional analyses. We will also note how access to complete chromosome-level assemblies is a natural path that fully complements our framework of understanding chromosomes' architecture and evolutionary implications.

1.4.1 Base principle

The idea that spatial information about the physical DNA molecule could be used to infer functional properties is not new. In 1983, Mark Mitchell and Peter Dervan proposed a synthesis of bis-(monoazidomethidium)octaoxahexacosanediamine (BAMO) which they used as a cross-linking agent in order to bind DNA fragments from the λ bacteriophage genome and probe its spatial structure. This resulted in a *nearest-neighbors map*, featuring five contacts, that enabled the authors to speculate about a possible 'solenoid' conformation of the bacteriophage's [239].

However, the base principles behind cross-linking and how they relate to the genome conformation were laid out much later. In 2001, Rippe drew from general hydrodynamic principles and polymer physics to establish a first practical working model of DNA coiling (also referred to as *random looping*) [240]. The model (and subsequent work on it) is detailed below. The biological protocol proper was first established by Job Dekker (in Nancy Klekner's laboratory) in 2002 [241] and is illustrated in figure 22.

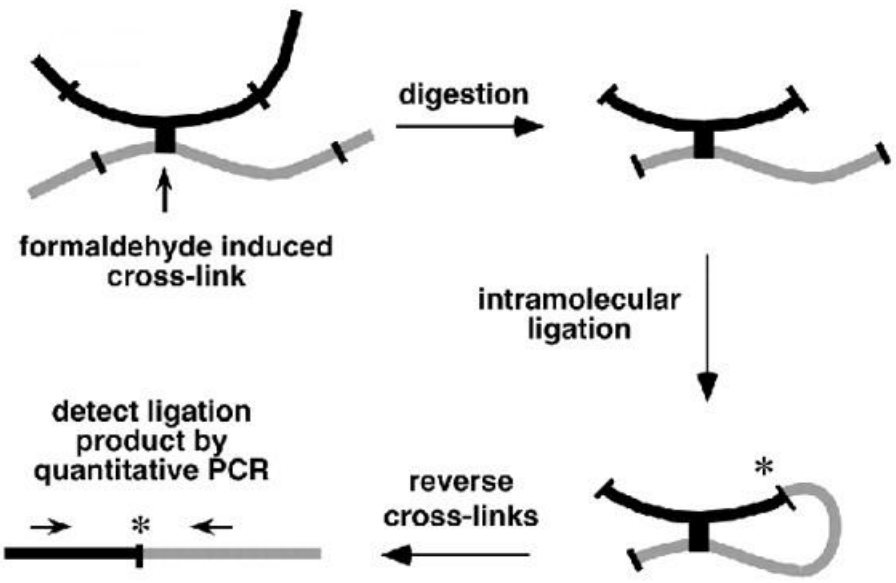


Figure 22: The chromosome conformation capture protocol. DNA is crosslinked, digested, ligated to form a 3C library. The contacts are reported on a contact map that represents interactions between DNA fragments.

Source: Taken from Dekker *et al.*, 2002 [241].

The idea is one may obtain a picture of the chromosomal architecture by measuring the contact DNA collisions between each locus pair in the genome. To do so, the chromatin and its surrounding proteins are cross-linked by formaldehyde, a very small and reactive

molecule that covalently bridges DNA with proteins, as well as proteins with proteins. The DNA is then cut with a restriction enzyme, then ligated again. The end result is that sequences that are close to each other in tridimensional space will be trapped in the crosslinked DNA-protein complex and eventually ligated together. A set of chimeric sequences is thus formed, called a *3C library*. It is essentially like a regular DNA library that can be PCR-amplified and sequenced in *paired-end*; by counting the occurrences of each locus pair being part of the same paired-end read, one gains access to the global contact frequencies across the whole genome.

1.4.2 3C-derived protocols in practice

The 2000's saw an evolution of 3C-based protocols as they were adapted and made to work on other species and underwent the next-generation sequencing wave. The articulation of the initial idea into different protocols (and subsequent applications) is detailed below [242].

1.4.2.1 Base 3C

The original 3C study was performed on *Saccharomyces cerevisiae* and showed the peculiar ring-shaped conformation of its chromosome III [241]. This seminal paper not only described the basis of the experimental protocol, but was also proposing a modeling approach to represent data that is still included in articles today.

The protocol was then adapted to other species such as mammals. For instance, 3C also showed that inter-chromosomal contacts may occur under specific circumstances underlying functional mechanisms; these range from immune response regulation to the homologous pairing of X chromosomes before X-inactivation [243]. It was also used in yeast to propose the existence of gene loops, as well as to show that following a double strand break chromatin becomes more insulated locally around the break.

1.4.2.2 Circularized Chromosome Conformation Capture (4C)

In 4C, or *one vs. all*, only one site is the focus of interest, and one examines its interactions with all other loci. A second ligation step is added to circularize that site. Circular DNA molecules containing the corresponding sequence act as "bait" and the unidentified interacting sites are inverse PCR-amplified and sequenced. 4C has been notable for showing that chromatin spatially segregates into active euchromatin and inactive heterochromatin domains [244]. It was also used to confirm the presence of chromatin looping at the β-globin locus in mammals and its importance in gene expression [245].

1.4.2.3 Carbon-Copy Chromosome Conformation Capture (5C)

5C corresponds to *many vs. many*; in other words, many loci are selected for interaction detection, typically spanning a relatively small region. To do so, every fragment within that region is ligated to universal primers. Fragments that are found to be annealing at a restriction site during the ligation-based amplification are ligated and presumed to

be interacting [246]. However, 5C has relatively low coverage and is ill-suited for indiscriminate screening of a whole genome. 5C was notably used to characterize interaction profiles at the X-inactivation center locus [247].

1.4.2.4 Whole genome Hi-C

With the advent of cheap second-generation sequencing technologies, Hi-C or *all vs. all* has become the standard for genome interaction studies [248]. Every single interaction between every fragment pair is thus accounted for. This is possible by an additional step whereby a biotinylated nucleotide is introduced at the edges of digested DNA molecules, before the ligation step. This enriches the library in molecules that have been digested then ligated together, diminishing dramatically the cost of sequencing. The protocol is illustrated in figure 23. Hi-C was showcased in the first genome-wide contact map of the human genome in 2009 [249]. In 2010, the genome-wide contact map of the yeast genome was published using an alternative protocol similar in spirit, i.e. aiming at enriching the sequencing library with informative events. Genome-wide contact maps are now the standard and in subsequent parts of this work, one should always assume that Hi-C was performed to yield our datasets, unless specified otherwise.

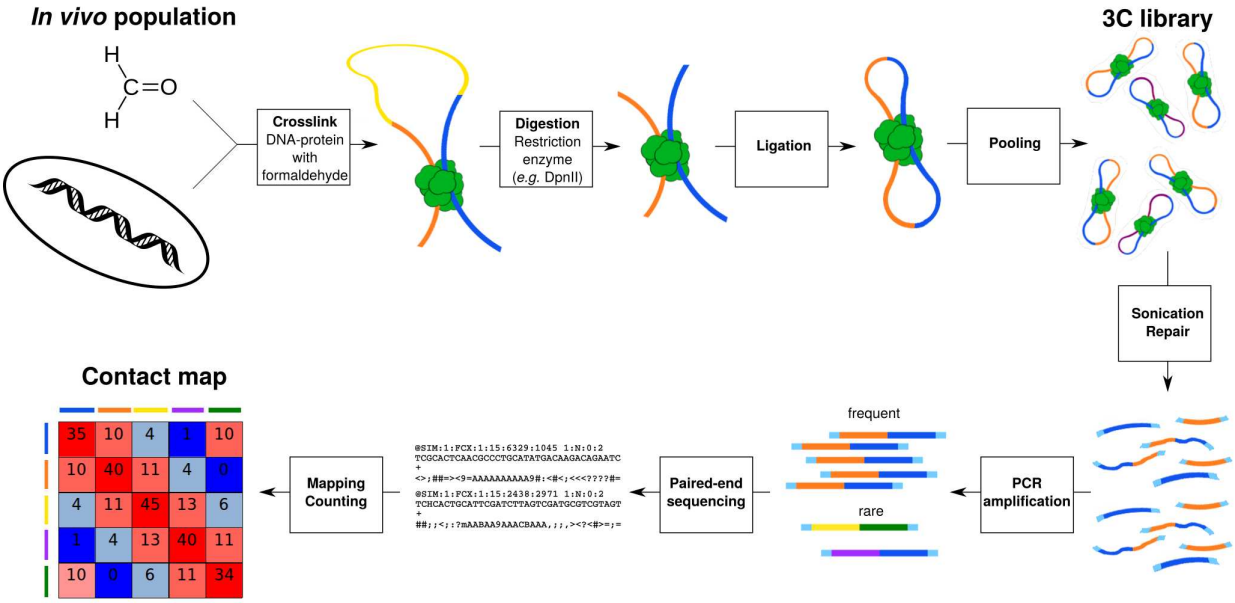


Figure 23: The Hi-C protocol.

1.4.2.5 Single-cell protocols

Single-cell 3C or Hi-C focuses on chromosome conformation within a single cell [250]. Early results confirmed that pooling single cell contacts yields a normal population-wise contact map [251]. It is unclear whether Hi-C data is ergodic, *i.e.* whether the behavior of a single chromosome, when tracked and averaged over time, would yield equivalent

data to that of a population's worth of chromosomes at a single point in time [252]. Because there is so little material, the signal generated by this protocol tends to be very sparse and relatively few interactions can be successfully recovered [253]. Nevertheless, single cell Hi-C has got increased interest due to its ability to differentiate homologous chromosomes within a cell, hence sometimes being called *Dip-C* for diploid single cell Hi-C [254].

1.4.3 Theoretical model

The model described by Rippe treats the DNA polymer as a sequence of N freely jointed monomers [240], as illustrated in figure 24. The length l of these monomeric segments is called the *Kuhn length* of the polymer.

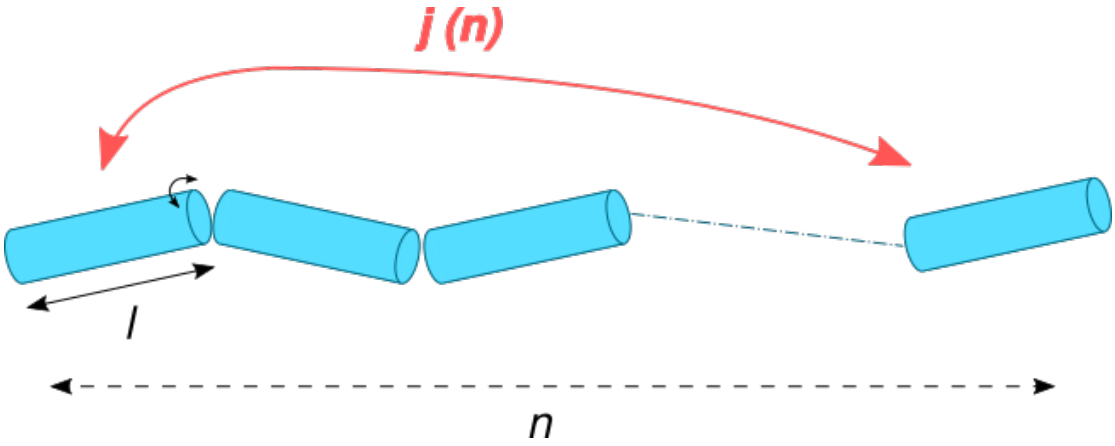


Figure 24: A freely-jointed polymer chain model. In this model, $j(n)$ represents, for each monomer, the local concentration of a neighbor separated by n monomers.

We want to assess, at any point in the polymer, the local concentration $j(n)$ of one segment located n Kuhn lengths away from it.

For a circular DNA molecule, and assuming Rippe's hypothesis, it is given by:

$$j(n) = 0.53 \cdot l^{-3} \cdot (n - \frac{n^2}{N})^{-\frac{3}{2}} \cdot e^{\frac{d-2}{n-\frac{n^2}{N}+d}} \tag{2}$$

and for a linear molecule, the equation can be deduced from above by setting $N \to \infty$:

$$j(n) = 0.53 \cdot n^{-\frac{3}{2}} \cdot e^{\frac{d-2}{n+d}} \cdot l^{-3} \tag{3}$$

The standard interpretation is that initial resistance (due to the polymer's rigidity, as given by its persistence length) decreases the frequency of close contacts, whereas far-off contacts also naturally decrease with distance. Since $j(0) = 0$ and $j(n) \to 0$ as $n \to \infty$, there exists a contact peak, dependent on d and typically equal to a few Kuhn lengths (see figure 25).

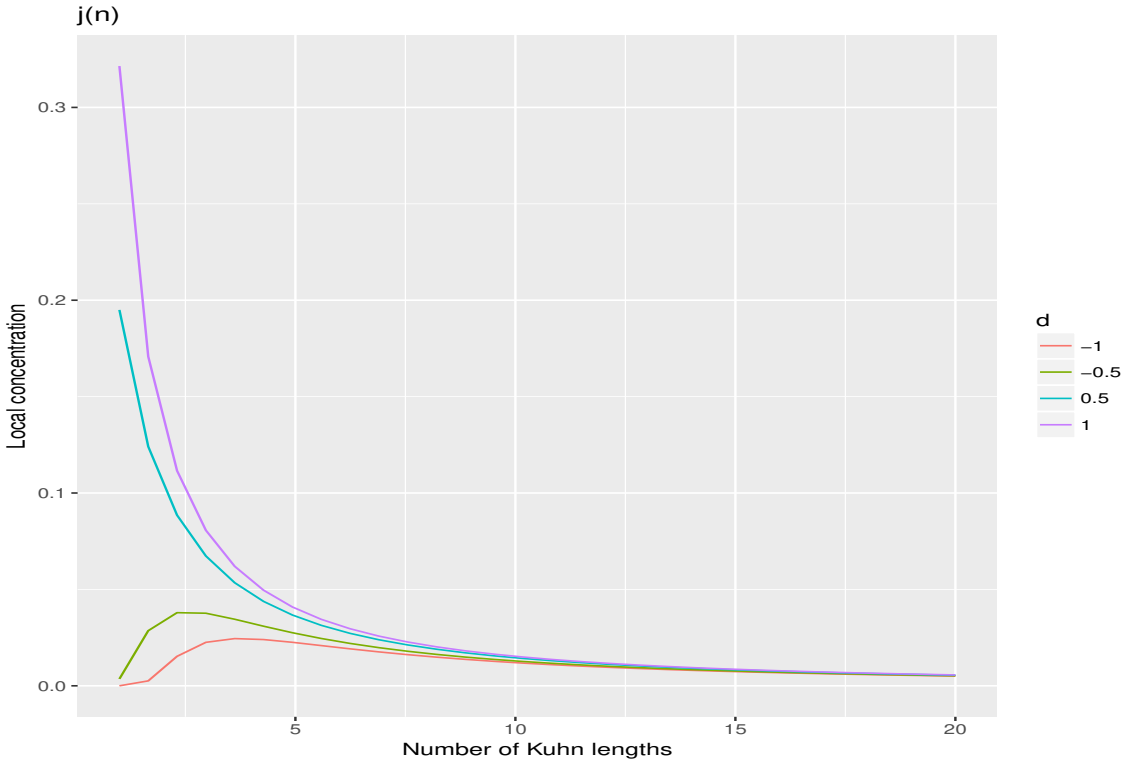


Figure 25: Evolution of local concentration as a function of distance in Kuhn segments (linear polymer, $l = 1$, $A = 1$).

Kuhn lengths can be impractical when reasoning with base pairs as a unit. This is because the flexibility of the chromatin may change across organisms, cell cycle conditions, etc. Let L be the length of a base pair, which we will assume to be constant along the DNA molecule. We therefore define the *genomic coordinate* s with the simple following variable substitution:

$$s = \frac{n \cdot l}{L} \tag{4}$$

If we reason that the above model is generally applicable to chromatin, and that base pairs are an accurate representation of monomers in the polymer chain, we obtain a probability of contacts $P(s)$ that directly depends on the genomic distance s between two loci, as shown in figure 26.

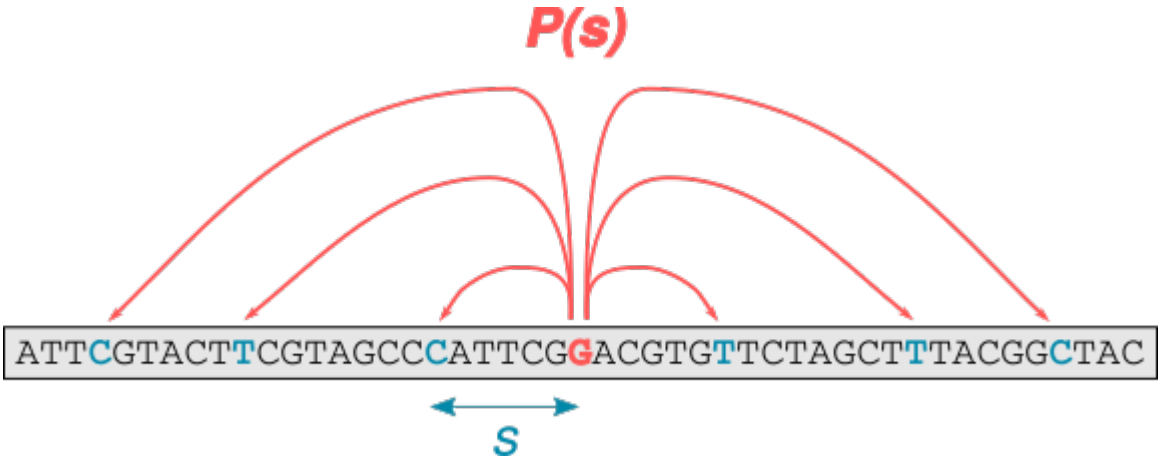


Figure 26: A freely-jointed polymer model adapted to chromatin. The frequency of contacts $P(s)$ is linked to the genomic distance s.

In practice, the quantity $P(s)$ is directly proportional to the concentration $j(n)$. Let A be such a constant pre-factor accounting for everything, including the variable substitution. In the case of a linear polymer, one obtains:

$$P(s) = A \cdot s^{-\frac{3}{2}} \cdot e^{\frac{d-2}{s \cdot \frac{L}{l}+d}} \tag{5}$$

The $-\frac{3}{2}$ exponent corresponds to concentration decay for ideal polymers at equilibrium state. However, current literature has also described DNA polymer in a fractal globule state [255] [249], where the exponent value is -1. This, combined with empirical observation, suggests that a whole range of intermediary states is possible [256] as illustrated in figure 27, *i.e.* the exponent must be an independent variable γ:

$$P(s) = A \cdot s^{-\gamma} \cdot e^{\frac{d-2}{s \cdot \frac{L}{l}+d}} \tag{6}$$

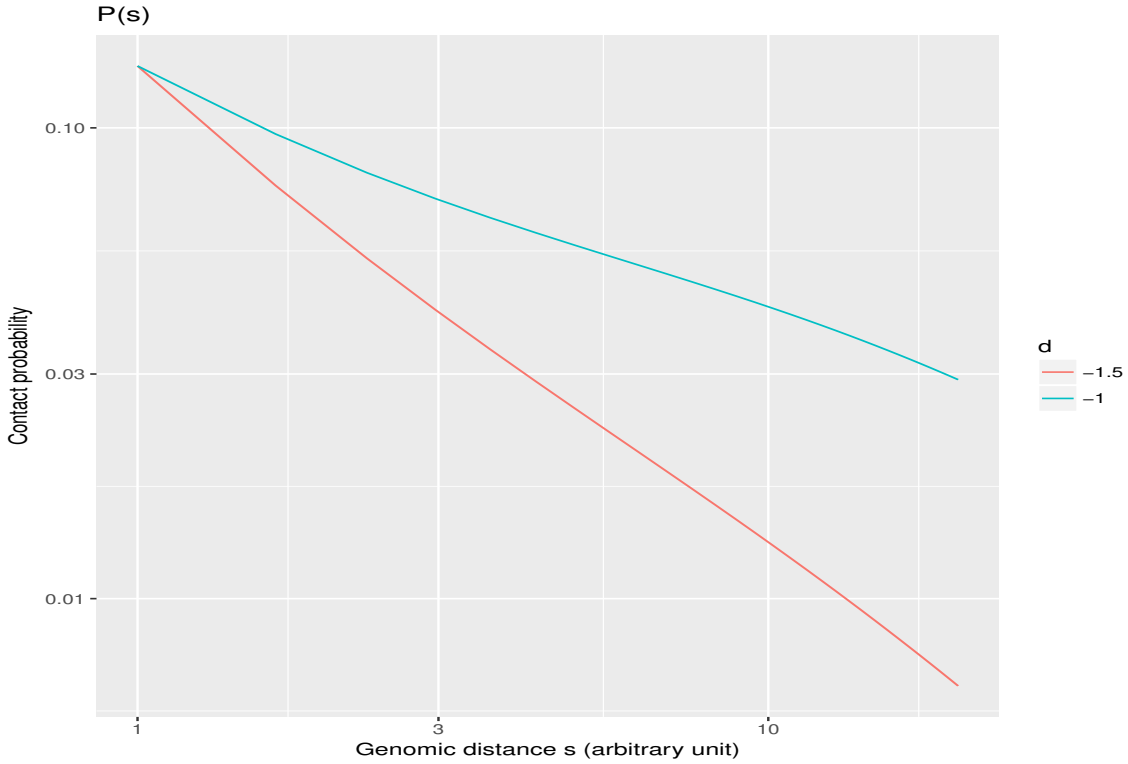


Figure 27: Evolution of the contact probability $P(s)$ under different polymer conditions: $\gamma = -1$ and $\gamma = -\frac{3}{2}$ (linear polymer, $l = 1$, $L = 1$, $A = 1$, $d = 0.6$). All the area between the two lines is a possible state.

However, the d parameter is hard to ascertain in physical terms, and could be dispensed with. Moreover, in the context of Hi-C experiments, very short range contacts (*i.e.* preceding the peak) are very hard to observe in practice, and s has sufficiently large values so that the exponential part negligibly affects the power law.

Recent work suggested that the frequency decay follows a (roughly) piece-wise power law: γ roughly takes two (positive) values at short scales (smaller s values) and large scales (larger s values), with a transitioning state at a given threshold [257] [258]. When taking into account the fact that γ is a function of s, and absorbing all pre-factors into either $\gamma(s)$ or A, one obtains a simplified equation:

$$P(s) = A \cdot s^{-\gamma(s)} \tag{7}$$

An example of such a function with a simplified sigmoid-like $\gamma(s)$ function is plotted in figure 28.

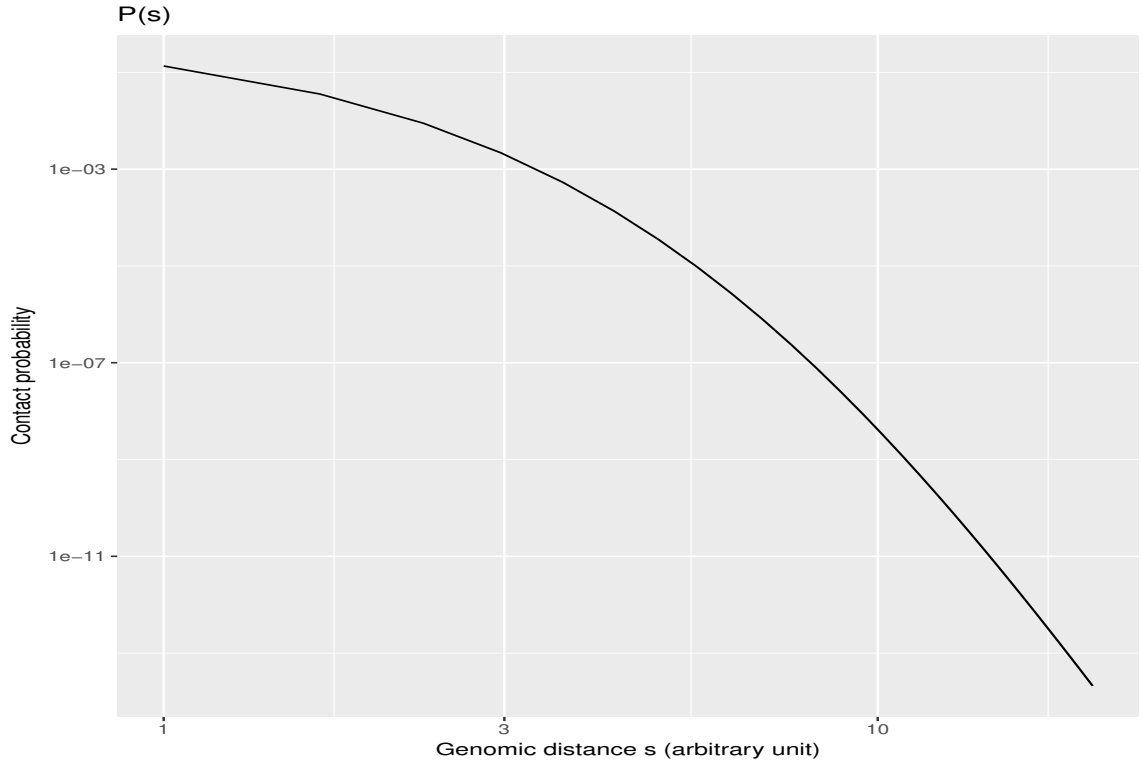


Figure 28: Evolution of the contact probability $P(s)$, $\gamma \in \left[-\frac{3}{2}; -1\right]$.

The study of the $\gamma(s)$ function remains an open question: in practice, it is known to depend on the formaldehyde concentration used in the protocol [256], and has been shown to change in various conditions, such as *e.g.* across the *Saccaromyces cerevisiae* cell cycle [259] [260]. However, we will see in section 1.4.7.1 that for scaffolding purposes its expression can be further simplified.

The question of inter-chromosomal contacts is muddier. Both theory and experimental results confirm that they are markedly lower than intra-chromosomal ones, due to the looping described above and because chromosomes tend to occupy *territories* of their own [261] [262], although there are exceptions. In mammals, the $\frac{intra}{inter}$ chromosomal contact ratio ranges between 40 and 60 [252], but it tends to decrease with coverage. However, increased inter-chromosomal contacts can be observed in special circumstances, such as *e.g.* the clustering of centromeres and/or telomeres in Rabl conformations [263]. In fact, this signal pattern is unique enough that it can be used to accurately identify the position of centromeres in *S. cerevisiae* [264] [265].

The above model gives a baseline for how many contacts one can expect in a Hi-C dataset. However, it is not clear whether there exists an analytical formulation of this function in a way that reflects all the biases, and the same goes for quantifications of the biases themselves. Several sources of signal distortion have been clearly identified, such as local GC content (which seems to follow an unimodal distribution [266]) or fragment length.

In the following section we will see how to process and correct actual biological data from Hi-C-based protocols.

1.4.4 Processing Hi-C reads

As Hi-C protocols grow more popular, a wealth of computational tools have spawned in an attempt to process their resulting libraries [267], [268], [269] [270]. All of them share a few things in common:

- They first *align* the reads against the reference genome of interest (or, in the case of metagenomics, a preliminary assembly).
- Resulting alignments are assessed, filtered, conditionally transformed into actual contacts and counted.
- A *matrix*, or *contact map*, or heat map of all final contacts is drawn.

The whole pipeline is illustrated in figure 29, using a contact map of the two-chromosome *Vibrio cholerae* genome as an example.

A number of distinct features shared across contact maps appear:

- The map is separated into "squares", each corresponding to its own chromosome-wide sub-matrix. Contacts are notably more important within a chromosome than between them. This is the natural interpretation of the polymer model, whereby each molecule interacts more with itself than other molecules. Barring some unusual biological mechanisms [271], this property is always respected and will be exploited in the following parts of our present work.
- The diagonal is notably more enriched in contact than other regions of the map. This is a direct consequence of the $P(s)$ power law (or piece-wise power law). The distance to the diagonal is the genomic distance s itself, and $P(s)$ decreases quickly.
- The corners of each intra-chromosomal submatrix are also enriched. This property is unique to circular chromosomes and simply due to all corners of a circular chromosome contact map representing in fact the same locus.

In the following sections we will go over the pipeline steps and and the various issues each of them may raise.

1.4.4.1 Mapping

The alignment step already surfaces a number of problems:

- First, as we are concerned with capturing DNA collisions beyond the usual immediate neighbors, it is important that each end of the read pair be mapped independently of one another. Many state-of-the-art aligners, such as Bowtie 2 [272] or minimap2 [192] are not designed for this, hence the need to independently map either read of a pair.

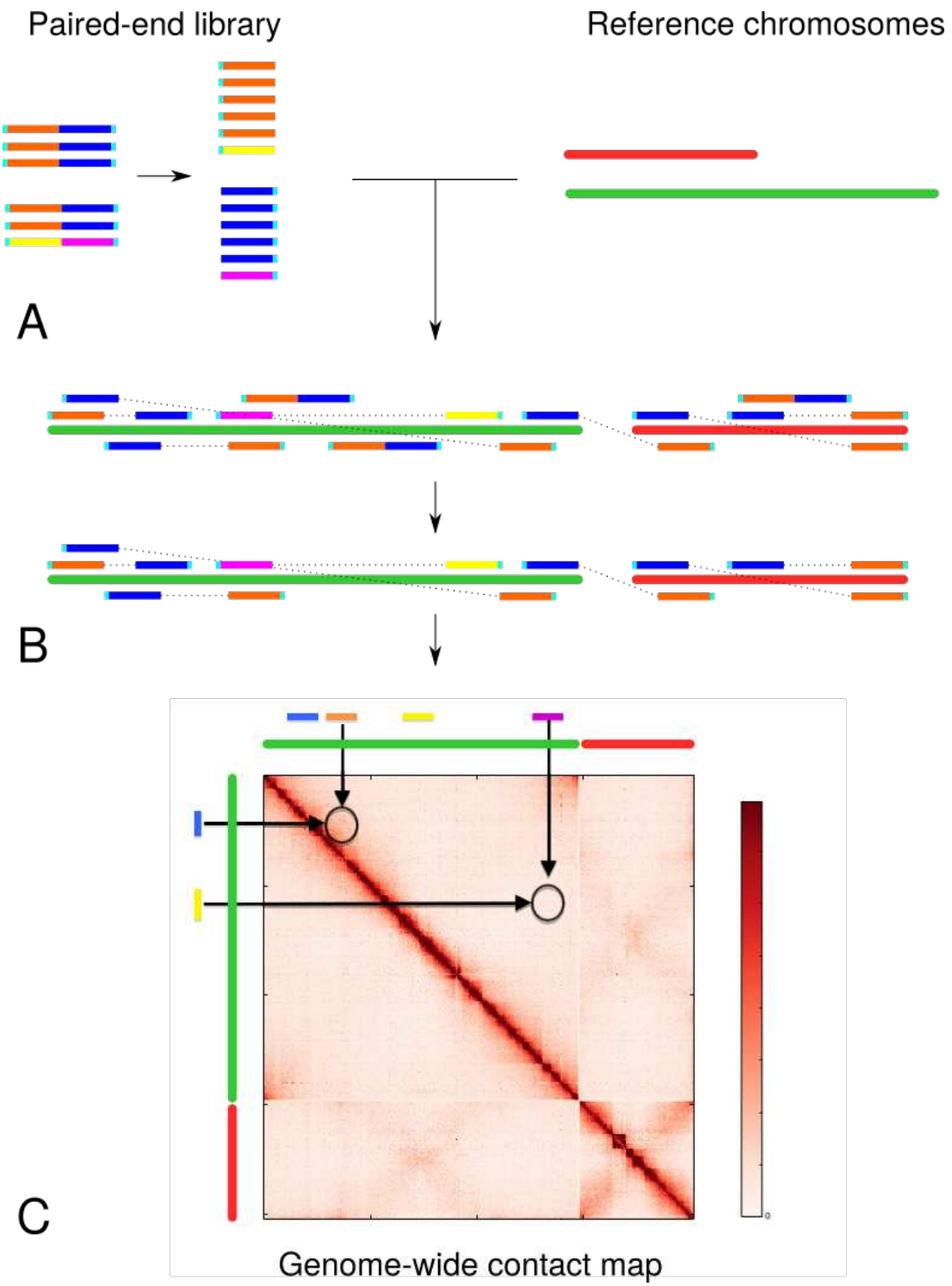


Figure 29: A standard Hi-C pipeline. Each end of the read pairs is (A) mapped independently onto the reference genome, (B) filtered so as to only retain "true" Hi-C contacts, then (C) reported onto a genome-wide contact map.

- Second, repeated sequences, or sufficiently similar sequences are completely ignored by standard mappers, and multi-mappers are not very suited for Hi-C processing. This results in artifacts and data loss. Solutions to this problem are scant, although recent attempts have been made to try and re-assign discarded contacts due to multi-mapping [285], essentially through interpolation.
- Moreover, due to the relatively random distribution of restriction fragments, the corresponding site could show up anywhere in the read pair, leading to some reads only partially mapping or being rejected due to chimeric bits. In order to alleviate that, it is often customary to iteratively truncate each read pair by regular intervals (*e.g. 10 bp*) and independently map each truncated read set onto the reference genome with the hopes of maximizing captured contacts. This step, however, can be time- and resource-consuming for diminishing returns.

1.4.4.2 Filtering

It is important to consider whether an alignment can be considered a true contact. The first criterion is the mapping quality Q, which relates to the probability E that an alignment is wrong by $E = 10^{\frac{Q}{10}}$; in the literature, a threshold of 30 is often chosen, meaning that one alignment out of a thousand will be wrong on average, and is usually satisfactory in practice. The second factor concerns the nature of the alignment itself; indeed many artifacts, illustrated in figure 30, arise due to the nature of the protocol:

- *Loops* occur whenever a DNA fragment wraps around itself instead of forming contacts with far-off neighbors.
- *Uncut* sequences occur whenever the restriction enzyme fails to actually digest a site between two DNA fragments, and both remain bound together and don't form other contacts.
- Other unexplained events (*weirds*) that are presumed to be artifacts due to the ends having the wrong orientation in the alignment. These only make up a small portion of all events.

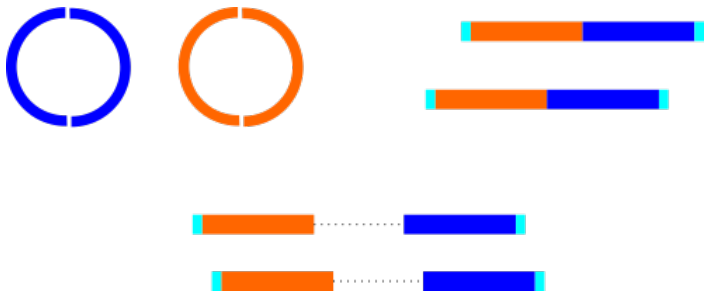


Figure 30: Artifact and real contacts. Artifact contacts (top) are divided into *loops* (left), *uncuts* (right) and *weirds* (not shown). Contrast with "true" contacts (bottom) between separated loci, as a result of an actual digestion-ligation.

These artifacts tend to drop off as the distance between two fragments increases; a threshold is typically set below which all alignments are considered artifacts and discarded. The remaining ones are then used to generate the contact map.

1.4.4.3 Contact map generation

A contact map is essentially a heatmap of all contact counts between all loci in the genome. The choice of contact map and its representation is not neutral; due to restriction fragments being heterogeneous in size and too numerous to handle in large genomes, contacts are typically regrouped, or sum-pooled, or *binned* into larger regions:

- Fixed length bins (*kb-based binning*), typically 2 kb to 100 kb for eucaryotes, depending on the genome size and the library sequencing depth.
- A whole number of restriction fragments (*fragment-based binning*).

The effect of binning is shown in figure 31.

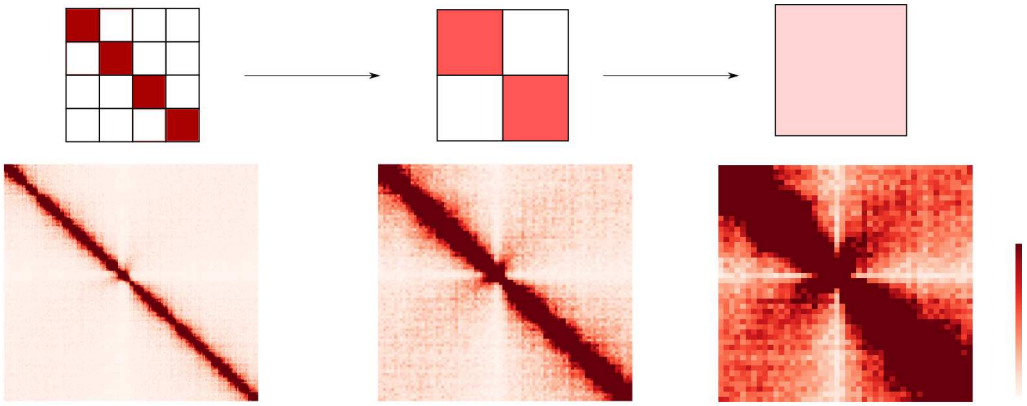


Figure 31: The effect of matrix binning. A simplified example (top) and a curated *S. cerevisiae* single-chromosome dataset (bottom) are recursively sum-pooled twice by groups of two fragments, yielding lower-resolution maps with stronger signal.

Binning at different resolutions allows a multi-scale analysis and determining the correct binning for one's interpretation is crucial, as some contact patterns are only visible at short (respectively large) scales. Fixed length bins alleviate somewhat contact biases related to fragment size, with the caveat that bins (and contacts thereof) do not represent the reality of physical DNA molecules in the protocol anymore. Fragment-based bins are more faithful to the experiment but still retain biases due to size heterogeneity, although the variation levels off at lower resolutions.

1.4.5 Handling contact maps and bias

As with any kind of data, Hi-C contact maps are subject to random errors (noise) and systematic errors (bias). Noise is often due to poor sequencing depth or poor-quality libraries and the signal-to-noise ratio can be alleviated by binning further, although at a cost of decreased resolution. On the other hand, bias is often inherent to the Hi-C protocol and organism in questions and takes many sources:

- Repeated sequences are simply unmappable and an unknown amount of signal can be ”lost” among these regions. In practice, matrices are often riddled of empty columns and rows that represent these repeat ”gaps”. A showcase example is shown in figure 32 (left).
- Sequences that are not quite repeated, but strongly homologous, may fool aligners into finding interaction signal when it only represents the occasional alignment error due to sequence similarity. This results in a very recognizable *homologous pattern* between the two regions in question. The pattern in question is shown in figure 32 (right).
- The fragment size distribution can be highly heterogeneous and this may adversely affect the contact distribution: larger fragments naturally receive more contacts due to their increased ”surface area” of potential interactions and not because of extra affinity in 3D. A practical example, shown in figure 33, shows that restriction fragments can show a lot of variation in length.
- The GC distribution is typically neither perfectly uniform nor balanced across a genome and this may bias the restriction site distribution. This can be remedied somewhat by using a GC-neutral enzyme (such as DpnII, whose restriction site is GATC) or several restriction enzymes and combining the libraries.
- Some chromosomes are naturally more covered in some regions. This is especially true among bacteria where several rounds of replications take place at the same time, and the origin may be up to eight times more covered than the *ter* region [161].

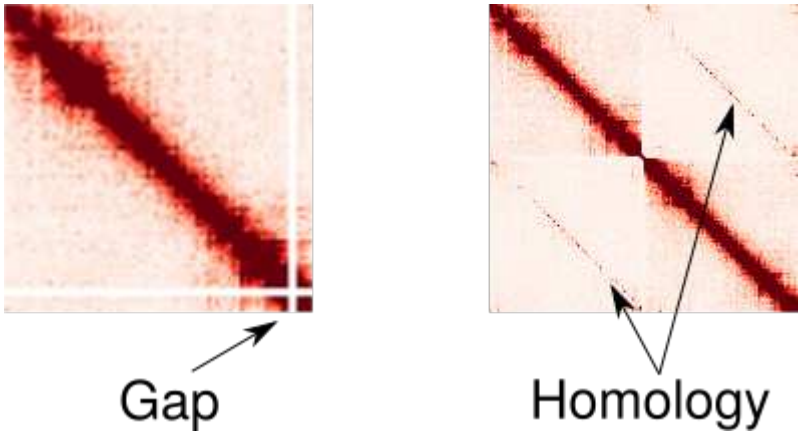


Figure 32: Two examples of mapping issues on a modified dataset: a gap created by the presence of repeated sequences (left) and extraneous signal between two homologous sequences (right).

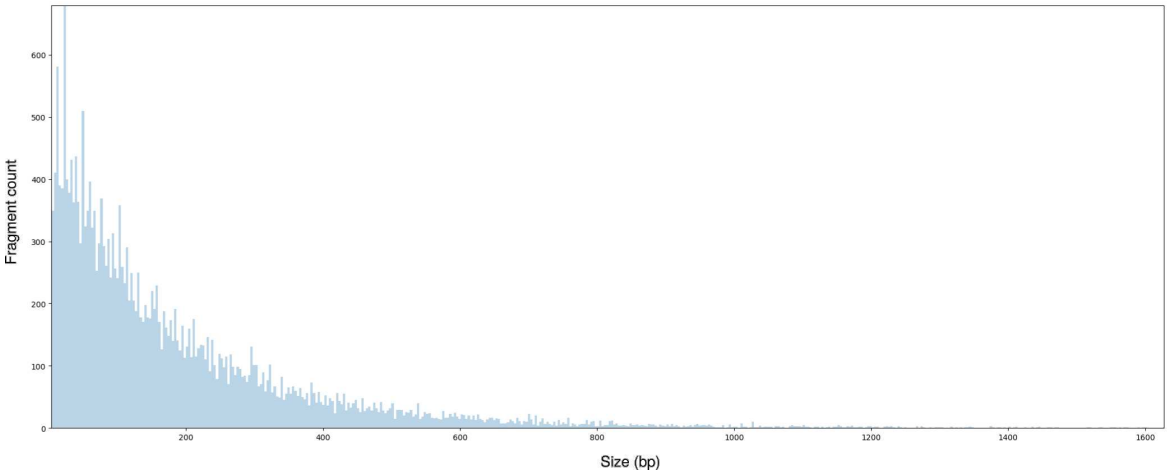


Figure 33: The (truncated) size distribution of the HpaII (site CCGG) restriction fragments on the *E. coli* genome.

Therefore, the need for correcting contact matrices naturally arises. In these sections we will cover different techniques for matrix normalization, noise reduction and comparison.

1.4.5.1 Normalization and correction

In the context of this work, and Hi-C data in general, a *normalization* is an attempt to transform the raw contact counts from the sequencer into actual interaction frequencies or probabilities by removing biases brought by extra-biological factors. Many normalization procedures have been documented regarding Hi-C data. They can be roughly

sorted into two different categories:

- Bottom-up: in this framework, one attempts to enumerate most error sources, quantify them, and adjust the signal accordingly. This requires one to accurately model the 3C experiment that yields the matrix either *ab initio* or by assuming that all bias source have been accounted for, a rather strong hypothesis. Nevertheless, probabilistic frameworks have been developed to that effect [273] [274] [275] [276] [277].
- Top-down: one makes no strong assumption about the bias sources and simply attempts to "regularize" the vectors individually with empirical procedures. They often tend to be inspired from linear algebra. Two notable examples of such normalizations are the iterative correction and eigenvector decomposition (ICE) [278] and sequential component normalization (SCN) [279] that have proven successful in subsequent analyses. Others include the Knight-Ruiz balancing algorithm [280], or simply natural norms ($\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$, etc.).

Throughout this work, one should assume that we have used the SCN when applicable. It is an iterative process, and can be defined by the following: let $M_0 = (m_{ij})_{ij}$ be a contact map of n fragments, where m_{ij} is the raw contact count between fragment i and fragment j. We define:

$$M_1 = \left(\frac{m_{ij}}{\sum_k m_{ik} \sum_k m_{kj}}\right)_{ij} \tag{8}$$

and define M_2 recursively with respect to M_1, etc. The $(M_n)_n = (M_0, M_1, M_2, ...)$ sequence can be shown to converge toward a matrix M where all of its vectors (columns or rows) sum to one, *i.e.* are probability vectors. Empirically, only a few iterations are sufficient to obtain matrices that are suitable for further analyses and interpretations.

Other procedures also coined normalizations have been recently published, attempting to address specific, structural sources of biases, such as copy number variants (CNVs) [281] [282].

1.4.5.2 Signal enhancing

Signal enhancing refers to methods and procedures that increase the signal-to-noise ratio. It is essential to facilitate the interpretation of dynamic events, such as rearrangement calling.

A very common source of noise is the low coverage of one's genome. At very high resolutions, on the restriction fragment level, relatively few contacts occur, and most matrices take up a binary aspect. Even using one the aforementioned binning methods, some regions in a genome may remain insufficiently covered to detect any pattern or confidently interpret any signal. As such, a number of tools have been developed that attempt at inferring the more obscure regions from the informative ones, following two broad approaches:

- *Matrix-based* approaches treat the contact map as an image. Since the contact distribution is generally continuous, and interaction counts of neighboring sequences are likely to be close, one may infer the value of a missing pixel from the the pixels around it, much like in imaging. Methods may range from very naive (convolution with a Gaussian kernel) to very elaborate: HiCPlus predicts high-resolution matrices from low-resolution ones by training a convolutional neural network on similar datasets [283]. A chapter of our work is devoted to a manuscript on matrix-based signal enhancement.
- *Graph-based* approaches reason that a contact map can be seen as the adjacency matrix of a weighted, undirected graph whose nodes are the bins and whose edges are the interactions. A simplified example is shown in figure 34. This approach is very common in Hi-C analysis, as the following sections will show. In the context of signal enhancement, instead of drawing inference from matrix neighbors, one takes *graph neighbors* into account. A neighborhood becomes a short path through the graph, instead of a pixel window. Examples of graph-based enhancement procedures include Boost-Hi-C [284].

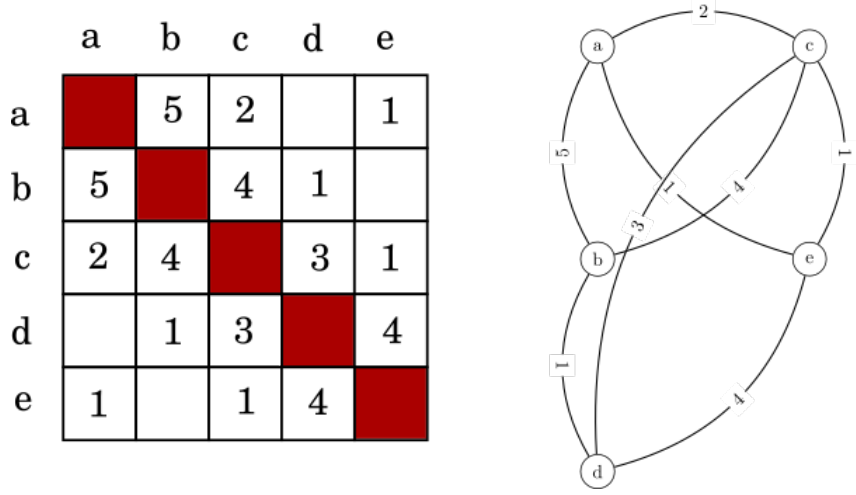


Figure 34: A contact map and its corresponding network. The map is the adjacency matrix of the graph. Zero values are omitted.

Other methods attempt to address specific caveats, such as gaps left by repeated sequences. Since multi-mapped reads do exist, and only have a limited set of potential alignments, one may attempt to re-assign them according to a probabilistic model, as is done by mHi-C [285].

1.4.5.3 Reproducibility and control

As Hi-C tools spawned and the field gained more prominence, these issues of data and pipeline reproducibility arose. Combining all the steps described above to produce,

process and handle Hi-C data can be time-consuming and error-prone. Subtle and hard-to-detect modifications in the final Hi-C dataset can be incurred by the most mundane sources such as one's choice of aligner or binning scheme. One must therefore categorize the source of variation:

- Hi-C protocols are not unified, and variations in how the experiment was carried out may induce changes among datasets on the same population, cell line, etc. This may cause issues among the community when one tries to reproduce published results. The advent of commercial Hi-C kits such as that of Arima Genomics may alleviate that. Two datasets that stem from independent experiments are called *biological replicates*, while datasets that stem from the same populations are called *technical replicates*.
- Most Hi-C pipeline implementations that have been made available typically differ in the way they process the data, what back-end software they use (*e.g.* the aligner could be Bowtie 2, BWA, minimap, etc.) which format they accept, etc.

Computational issues are thus compounded by biological ones. To that effect, many methods have been designed, mostly borrowed from linear algebra and graph theory since contact maps fit both fields pretty well. As is often the case, there is no consensus solution:

- Most of them compute correlation coefficients (Spearman or Pearson) between the matrices [286], with the drawback that these measures are particularly outlier-sensitive. This can be alleviated by instead correlating some more robust proxy measures related to the matrix [287].
- Some methods, like OneD, are specifically geared for structural and copy number variants [288], by relying on the *contact profile* (summed bins) of the matrix.
- Other tools rely on multidimensional scale reduction (MDS): for instance, HiC-spector [289] computes the Euclidian distance between the first twenty eigenvectors of each matrix's Laplacian. These vectors are presumed to contain the bulk of the structuring signal. The final score is shown to separate pseudo-replicates and biological replicates on one hand, from different cell line datasets on the other hand.
- Another tool exploiting a graph-based approach GenomeDISCO [290], performing random walks on the corresponding graph to smooth the matrix and thus ignore outliers.

These reproducibility issues have been subject to increased scrutiny as researchers attempted to detect structural variants, karyotypic aberrations and other such differences between cell lines, as is common in cancer detection [286]. These are prime examples of chromosome dynamics where the use of these tools grows crucial.

1.4.5.4 Matrix comparison

It is common to have to compare two or more different Hi-C datasets to draw an interpretation of interest. Just like matrix reproducibility boils down to quantifying the "sameness" between two datasets, matrix comparison attempts to quantify their "essential difference". As multiple matrices get compared, the question of defining a *contact map distance* arises.

A very simple and common practice is to qualitatively evaluate log-ratios between two matrices, pixel by pixel. Unfortunately, it may fall short when differences one wishes to see get drowned out by the noise: as figure 35 shows, despite the matrices A and B being noticeably different when compared side to side, it is unclear whether the trend could be detectable based solely on the resulting ratio C. Nevertheless, with enough coverage, qualitative differences, and methods for "averaging out" the noise, the approach still has had success [291].

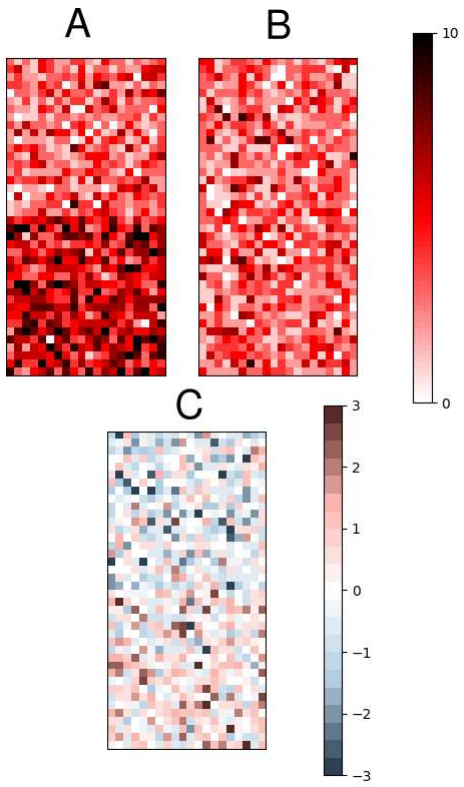


Figure 35: Two simulated contact maps (A) and (B) and their log-ratio (C). B and the upper half of A are generated according to a Poisson random variable of parameter λ and the lower half of A is generated according to a Poisson random variable of parameter 2λ.

Other traditional methods such as spectral analysis or PCA have also met results. For instance, a PCA on the pairwise Euclidian distances between contact maps along the

cell cycle of *S. cerevisiae* showed that 1) points taken under the same stage tended to cluster together, and 2) taken together, the clusters were scattered across a cycle that respected the order of cell cycle stages (figure 36).

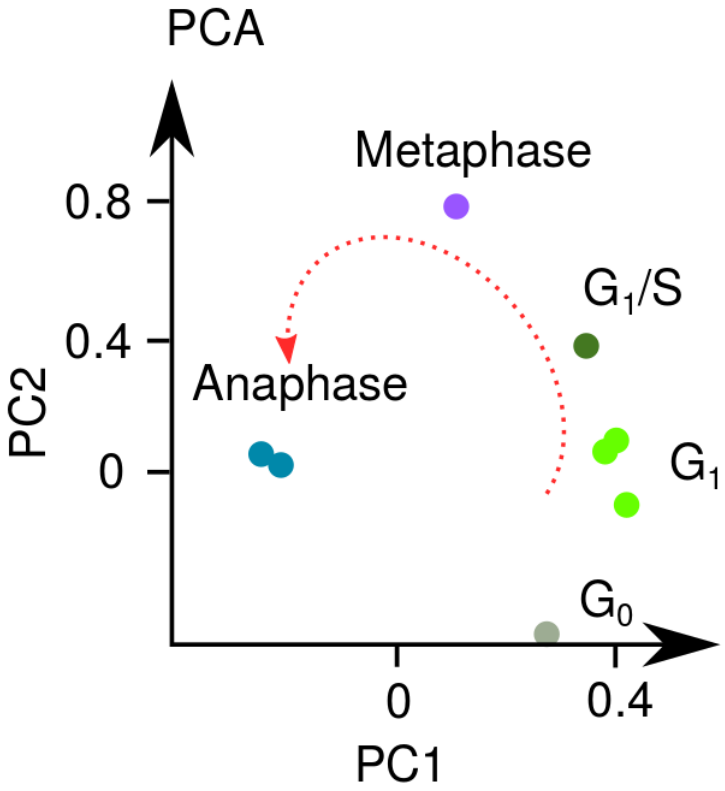


Figure 36: A PCA of contact maps taken at different points during the cell cycle of *S. cerevisiae*. The coordinates of the points in principal component space are consistent the orientation of the cell cycle itself.

Source: Adapted from Lazar-Stefanita *et al.*, 2016, [259].

However, with a limited number of datasets or limited coverage, more refined methods are necessary:

- diffHiC [292] implements a sophisticated model to account for both sources of variability (technical and biological) using a quasi-negative-binomial distribution for contact counts, which generalizes previous models based on binomial distributions [293].
- HiCCompare [294] performs a joint non-parametric regression (a LOESS) to eliminate the noise present in both datasets and computes Z-scores between regions of interest. The joint-normalization design makes the method sidestep most sources of bias between replicates.

- FIND [295] and SELFISH [296] reason that any local difference between two regions of interaction but also be reflected by their surroundings, and compute differences in distribution in windows around loci pair to draw a distinction between noise among replicates and actual differences of interest.

Many of these methods struggle with copy number variants, as they alter the contact distribution and probability ($P(s)$) however it is modeled. Note that some of the reproducibility tools described in section 1.4.5.3 are also commonly used as comparison software.

1.4.6 Dynamics implications

After reviewing how Hi-C data works, we will give some of its principal results on chromosome architecture and dynamics.

1.4.6.1 Compartments

In eukaryotes, notably mammals, chromatin folds into *compartments*. This was shown by the first Hi-C experiment and the first genome-wide contact map of the human genome [249]. It features interlaced megabase-sized stretches that alternate between *active*, euchromatic and *inactive*, heterochromatic regions, also called *A/B compartments*. Genes in A-compartments are transcriptionally active whereas those in B-compartments are inactive [297]. Notably, the X chromosome structure changes considerably depending on whether it's active or inactive [298] [299]. Moreover, regions in A- (respectively B) compartments tend to cluster together, at the exclusion of B- (respectively A) compartments. This gives the corresponding contact maps a characteristic "checkerboard" look 37, as contact-rich and contact-poor regions alternate in both directions.

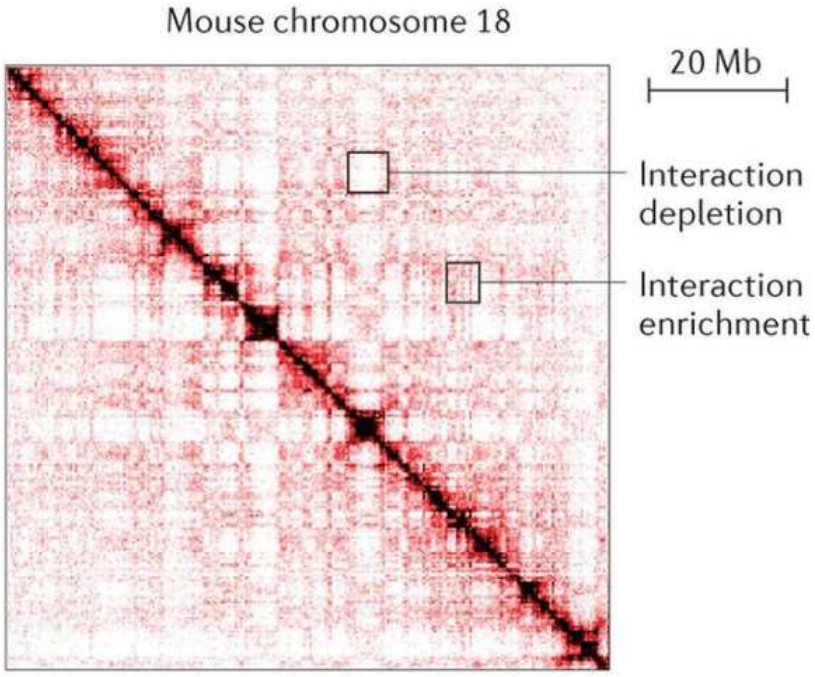


Figure 37: A/B compartments in a mouse chromosome contact map. The "checkerboard" pattern stems from contact-depleted regions alternating with contact-enriched regions.

Source: Adapted from Dekker *et al.,* 2013, [297].

A/B compartment membership is also highly correlated to replication timing profiles: regions in A- (resp. B) compartments are replicated early (resp. late) [300]. A- (resp. B) compartments are also correlated with high (resp. low) GC content and DNA accessibility, as well as active histone marks (resp. repressive histone marks and lamina association [301]). Membership is also highly predictive of cell type. However, this A/B classification is not static: the chromatin architecture has been shown to reorganize during cell differentiation. Up to 36% of the genome switches compartment at some point during the process. This shows that the compartment organization can be plastic and partly contribute to some cell-type specific patterns of gene expression [302]. Compartment membership is also known to be correlated, and in fact can be predictively reconstructed, with epigenetic data [303].

A/B compartment classification can be determined by computing the first eigenvector of the normalized contact map correlation matrix: its components alternate signs as one switches from an A- to a B-compartment. Other statistical methods have been suggested to compute it, such as CScoreTools [304].

1.4.6.2 Topologically associating domains

Each compartment is made up of smaller-scale structures whose sequences preferentially interact with each other, called *topologically associating domains*, or TADs, that are a few hundreds of kilobases in size. They act as building blocks in many animal species (but not in yeast): in humans and mice, more than 90% of the genome is structured along a series of over 2,000 TADs [297].

In mammals, TADs are primarily defined by their borders, whose loci are enriched in CCCTC-binding factor (CTCF) and condensin [297] [305], as shown in figure 38. Both are known for orchestrating their formation and for their role in gene regulation. Genes present within a TAD are expressed at the same time during cell differentiation [297], and sequences within a TAD are also replicated at the same time during S phase [306]. They share the same set of regulatory elements; disrupting a TAD also disrupts gene regulation, potentially pathogenically [307].

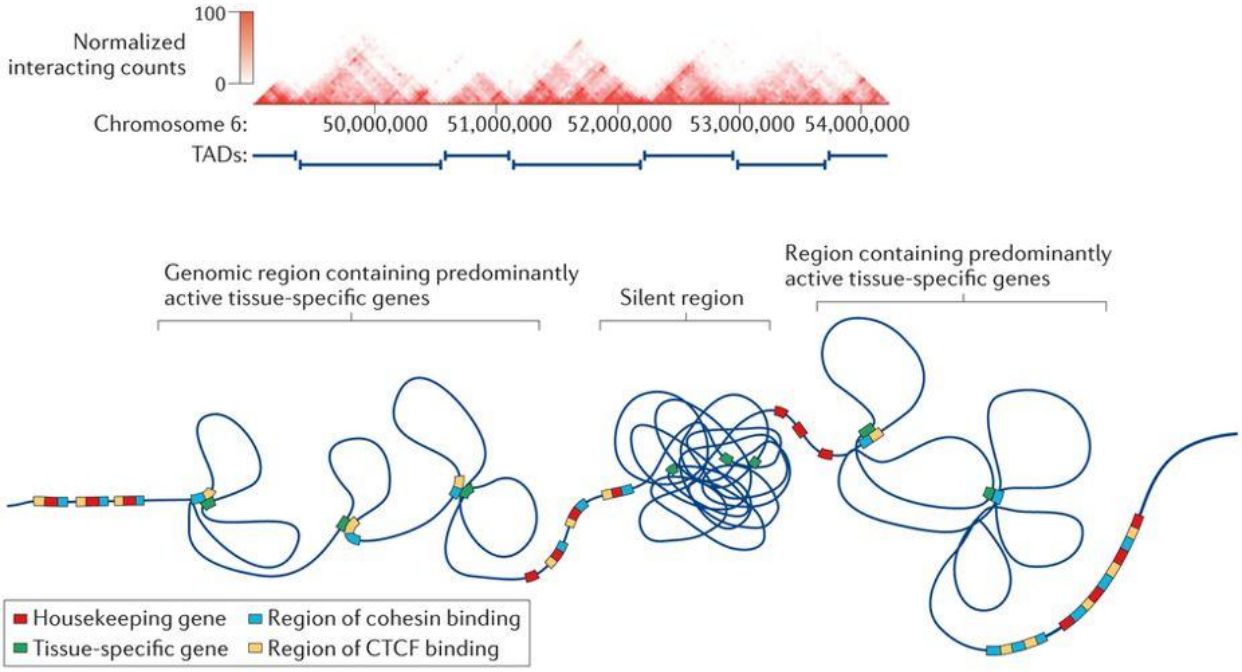


Figure 38: TADs in a mouse chromosome. The contact map (top) shows how they are delimited by differential interactions on either side of the borders, and the schematic (bottom) illustrates the role of condensin and CTCF binding at the border loci to maintain the architecture.

Source: Taken from Pombo *et al.,* 2015, [305].

On the other hand, in *Drosophila melanogaster*, TADs are not bound by CTCF sites or cohesin. Instead, they tend to arise as a statistical property of the DNA polymer when many instances are merged and pooled. This suggests that the chromosome dynamics

within those TADs are different and markedly less constrained as what is found in mammals [308] [309] [310] [311].

TADs can be computationally identified with the so-called *directionality index* (DI) [312], or the *insulation index* [313]. A variety of implementations exist in the literature [314] [315] [316]. The DI is computed with a statistical test between two vectors of opposite orientations along a locus: since a TAD border is characterized by a stark contrast between its left- and its right-hand side, the DI will peak at each border. The insulation index compares contact counts along each direction. More sophisticated methods of TAD-calling have been suggested, notably using Gaussian mixture models [317] or network modularity approaches [318].

1.4.6.3 Chromatin loops

Loops are the 3D juxtaposition of two distant (in 1D) loci across a chromosome. The distance can range from 10 kb to 200 kb [319]. They have been long known for their role in gene regulation, notably through early nuclear ligation assays evidencing a promoter-enhancing proximity for the rat prolactin gene [320]. Later, in of the first 3C studies linking conformation to function, it was shown that the β-globuline gene promoter and its corresponding target, located 50 kb away, were physically linked by a chromatin loop [321]. The advent of Hi-C technologies later confirmed initial insights that transcription could be activated or repressed with the juxtaposition of far-off gene loci, as well as the role of CTCF in their structuring [322]. However, not all CTCF-enriched sites are necessarily involved in loop formation [319].

Chromatin loops show a distinct pattern on a Hi-C map, a 2D signal peak far-off from the diagonal, as shown in figure 39. In 2014, Rao and colleagues performed an extensive Hi-C analysis of many cell human and mouse cell lines, confirming the presence of more than 10,000 such loops, many of which were conserved among cell lines [299].

Figure 39: Hi-C profile of a chromatin loop from a human chromosome. The peak represents the junction between two far-off loci.

Source: Adapted from Rao *et al.,* 2014, [299].

The majority of these loops were bound by CTCF and cohesin. Later, it was shown that cohesin loss induced a total loss of all loop structuring [323], causing superenhancers to cluster together.

Neither loops nor the TADs they are forming are static structures; both the initial simulations and single-molecule imaging evidence show that both dynamically break and reform across the cell cycle depending on cohesin degradation and replenishment [324].

Loops are sometimes found at the border of TADs, but not always. A potential mechanism by which chromatin looping results in TAD formation could be *loop extrusion* as first suggested by an initial model [325] that was progressively refined [326] and followed by computer simulations [327] [328] [329] [330]. It relies on a more advanced modeling of the $P(s)$ function than the one we have described. An illustration is shown in figure 40. The chromatin is translocated by a *loop extruding factor* (LEF) (presumably cohesin) until it the molecule meets an obstacle, which would be CTCF [331]. Simulations show that the superposition of many different extrusion states as would be expected in a population result in TAD-like structures.

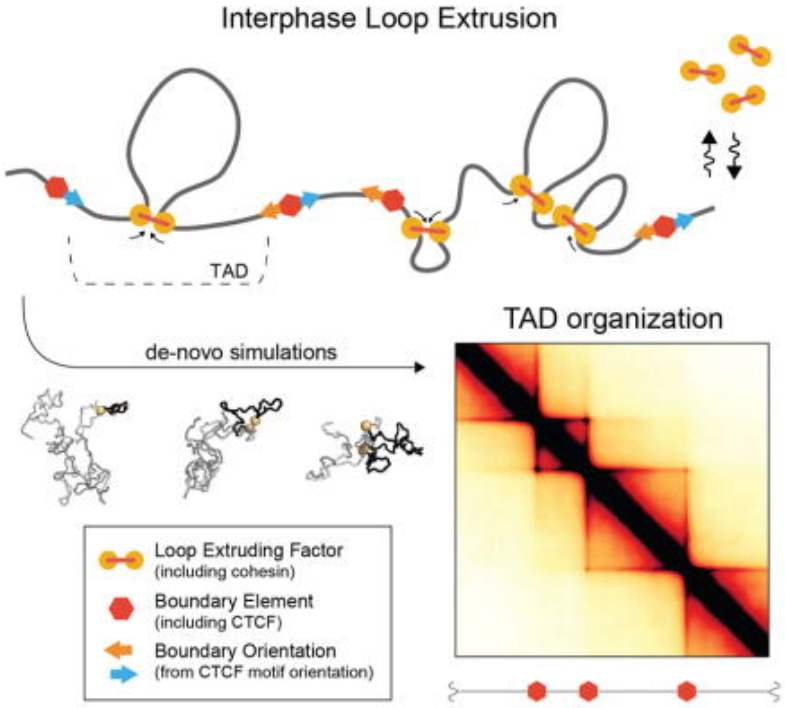


Figure 40: The loop extrusion model in TAD formation. The molecule is extruded between boundary sites and the combination of many different states of extrusion yields TAD-like preferential interaction domains.

Source: Taken from Fudenberg *et al.,* 2016, [330].

Although a growing body of evidence has emerged to point at cohesin being the main LEF [332] [333], it has not been directly confirmed. On the other hand, live imaging of DNA loop extrusion has unambiguously shown the mechanism to be mediated by condensin [334], another member of the SMC family. With no consensus whether both

are involved, or condensin only, the discussion is still ongoing.

1.4.7 Application for genome and metagenome assembly

In the previous sections we have covered a basic overview of sequencing and assembly methods as well as their lingering challenges, on one hand; on the other hand, we have laid out the basics of processing and interpreting Hi-C data and its relevance to chromosome dynamics. In this section we will combine the two and explain how Hi-C can be used to solve assembly problems and unveil the underlying dynamics.

1.4.7.1 3C-based genome scaffolding

In section 1.4.3, we have seen that, absent very short scales that don't concern us in the context of Hi-C experiments, and whichever equation is used, the contact frequency function $P(s)$ is strictly decreasing. It is therefore bijective, *i.e.* there exists a one-to-one mapping between the contact frequency (3D distance) and the genomic distance (1D distance). Scaffolding a genome based on contact data is therefore equivalent to finding an appropriate $P(s)$ function, and rearranging distances between sequences such that the genome and its underlying contact map best fits that $P(s)$ function.

Existing software There are few available Hi-C based scaffolding programs: Lachesis [335] was one of the earliest attempt but had a number of drawbacks, the most notable of which being the requirement to specify a number of scaffolds or chromosomes in advance, resulting sometimes in aberrant chromosomes with large number of improper rearrangements. It is now deprecated. Another more recent tool, 3D-DNA checks first for "misjoins", partitions misjointed scaffolds, removes problematic sequences, and merges the remainder with overlaps. This method was hallmarked with the chromosome-level scaffolding of the *Aedes aegypti* genome [336]. More recently, SALSA2 uses a promising approach, directly integrating the weights of the contacts into the assembly graph [337].

Our present work is based on GRAAL (Genome (Re-)Assembly Assessing Likelihood from 3D) [338], a pioneering program developed by Hervé Marie-Nelly, a joint PhD student between the groups of Romain Koszul and Christophe Zimmer. Notably, GRAAL was the first program able to scaffold a truly incomplete eukaryotic genome in 2014. Most of our present work is based on the continuation of GRAAL.

Naive method In order to understand why an elaborated method is necessary, consider the following naive, greedy algorithm 2.

It simply finds the two most interacting fragments and extends in either direction depending on the strongest neighbor of each extremity. Unfortunately, this invariably encounters caveats:

- The approach is only guaranteed to work if the Hi-C data were perfect, *i.e.* each fragment's neighbors in 1D scrupulously respects the strictly decreasing $P(s)$ condition stated above. In practice, that function is subject to noise as any stochastic

Algorithm 2 Greedy Hi-C assembly algorithm

Require: $\mathcal{F} = \{f_1, \ldots, f_n\}$, a set of fragments
Require: $M = (m_{ij})_{ij}$, a contact map of the framgents
  $(u, v) \leftarrow \arg\max_{i,j} m_{ij}$
  $C \leftarrow [u, v]$
  repeat
    $w_1 \leftarrow \arg\max_k m_{uk}$
    $w_2 \leftarrow \arg\max_k m_{kv}$
    if $m_{uw_1} > m_{w_2v}$ **then**
      $C \leftarrow [w_1, C]$
    else
      $C \leftarrow [C, w_2]$
    end if
  until an incompatibility arises

variable, in addition to the biases that we have covered. This means that loops are created and the contact chain never extends to the full genome.

- Even if the data *were* perfect, Hi-C data remains a measurement over a population of cells and we have seen the DNA polymer is known to be very dynamic: its conformation constantly changes over time, sometimes drastically, with, as we have seen, no indication as to whether its changing behavior is ergodic [252].

Stochastic model Ultimately the difficulties arise from the fact that Hi-C data is inherently noisy and the errors make such greedy methods fail at conveying the stochastic aspect of contact counts. While the expected number of contacts $P(s)$ is well-studied, the nature of the random variable driving the actual contact counts between two loci pairs remains elusive:

- One may treat each pixel (locus pair) as a counting procedure where each individual contact is a rare event, thus yielding a Poisson distribution. This is what some of the normalization procedures we have covered attempt to regress on.
- Many tools reason that each fragment or bin has a number M of contacts to be distributed across the rest of the genome. The exact value of M depends on the local coverage. From there, reasoning about the probability a certain amount of the M contacts being made at a certain locus naturally leads to a binomial distribution [339] [340].
- One may simply reason that the combination of many independent and identically distributed conformations of chromosomes as reflected by a full cell population leads to a Gaussian distribution for each pixel.

And as we have seen in section 1.4.5.4, other models such as a quasi negative binomial one have emerged. In an analysis of a large aggregation of datasets [263] [259], plotting

the evolution of the variance with the mean, it was strongly suggested that contacts follow a Poisson distribution when they are scarce, and a Gaussian one when contacts increase, with a transitory state between both regimes [256] (see figure 41).

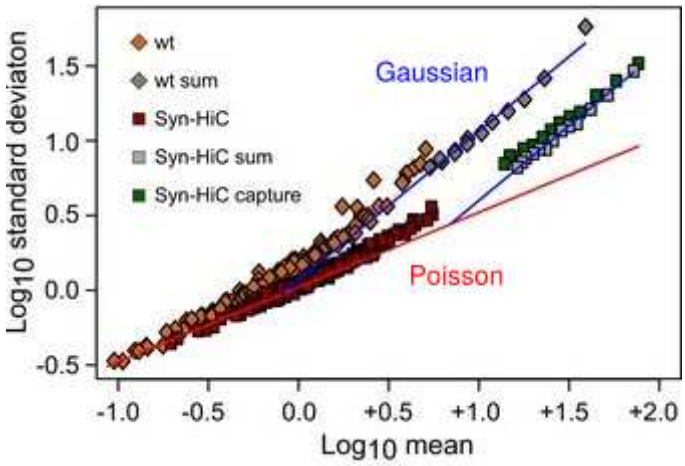


Figure 41: Dual regime of the contact distribution, as illustrated by the relation between the variance and the mean transitioning from Poisson to Gaussian. Datasets were obtained from Mercy *et al.*, 2017 [263] and Lazar-Stefanita *et al.*, 2017 [259].

Source: Adapted from Muller *et al.,* 2018, [256].

Moreover, the regime is resolution-dependent, as more binning will lead to more contacts per bin and thus favor the switch to a Gaussian distribution. Presumably the exact scale at which the transition occurs is the optimal binning scale for a given dataset.

In the context of this work, we will rely on a Poisson distribution only, on the basis that we want our algorithm to be relatively robust at low coverage and to work regardless of any normalization.

The GRAAL algorithm The basic principles of the GRAAL algorithm have been laid out in its original framework [341]. It is inspired from an Markov Chain Monte Carlo (MCMC) method known as Gibbs sampling.

Modified polymer model First, the equation $P(s)$ is simplified so as to fit a simple three-variable model. In the context of this work we will consider $\gamma(s) = \gamma$ to be constant, which we have seen is true at short scales:

$$P(s) = A \cdot s^{-\gamma} \tag{9}$$

However, this function decreases quickly to zero as $s \to \infty$, leading to it having lesser values than the base interchromosomal noise. This is a direct contradiction with empirical evidence or any kind of accepted modeling, so one needs to introduce a third

parameter δ representing that noise level (assumed to be constant) and stipulate that $P(s) \geq \delta, \forall s > 0$. Therefore, our modified model becomes:

$$P(s) = \begin{cases} \max\left(A \cdot s^{-\gamma}, \delta\right) & \text{intra-chromosomal} \\ \delta & \text{inter-chromosomal} \end{cases} \tag{10}$$

The vector of *nuisance parameters* $\xi = (A, \gamma, \delta)$ must therefore be initialized to best fit the contact data. In practice, GRAAL uses the Broyden algorithm (a quasi-Newton method).

Assessing likelihood Let $M = (m_{ij})_{ij}$ be a contact map. Assuming the value of each pixel m_{ij} obeys an independent Poisson counting process, and the expected value (equal to the parameter) of that process is given by our polymer model $P(s) = P(s_{ij})$ (where the genomic distance can be expressed as $s_{ij} = b \cdot |j - i|$ and b is the binning scale), the likelihood of observing a contact count m_{ij} at the pixel (i, j) is given by:

$$L(m_{ij}) = e^{-P(s_{ij})} \cdot \frac{P(s_{ij})^{m_{ij}}}{m_{ij}!} \tag{11}$$

The likelihood of the whole matrix is the product of all likelihoods for each pixel, since they are assumed to be independent:

$$L(M) = \prod_{i>j} e^{-P(s_{ij})} \cdot \frac{P(s_{ij})^{m_{ij}}}{m_{ij}!} \tag{12}$$

It is the ratio of two such quantities that is examined when considering a genome or parameter modification.

Genome mutations In order to perform the assembly, a number of mutations, shown in figure 42 are predefined so they can be tested:

- `Split`: split a contig at the location of a given fragment.
- `Paste`: merge two contigs at the location of a given fragment.
- `Duplicate`: add a fragment to the current fragment set.
- `Delete`: remove a fragment from the current fragment set.
- `Flip`: invert a fragment's orientation.

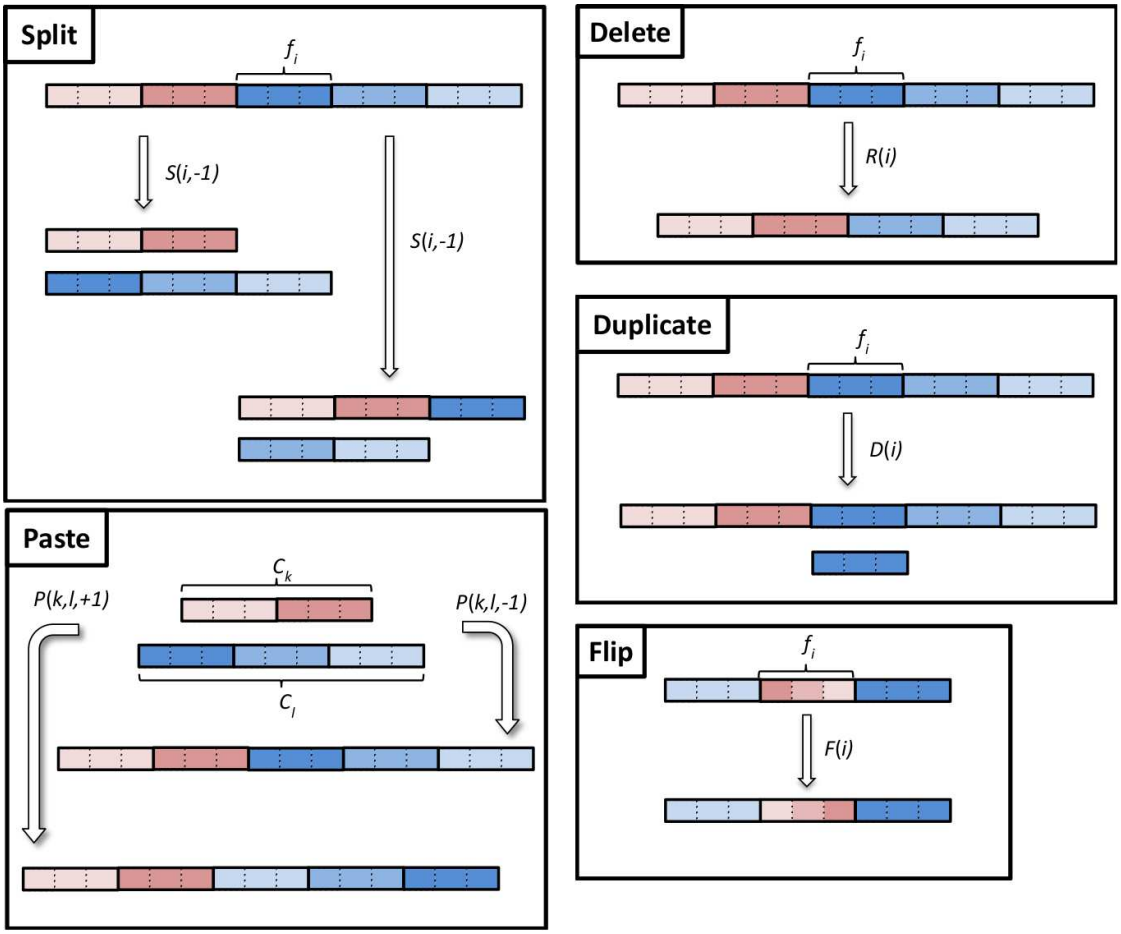


Figure 42: The elementary genome mutations: `split`, `paste`, `duplicate`, `delete`, `flip`.

Source: Adapted from Marie-Nelly *et al.*, 2014, [338].

These operations are elementary but testing each one of them may take a long time. In practice, more advanced operations are defined on top of these (shown in figure 43):

- `Eject`: remove a fragment from a contig, merging the junctions. It is a combination of two `split` and a `paste`.
- `Insert`: inserts a fragment at a given contig's location. It is a combination of a `split` and two `paste`.
- `Translocate`: swap two fragments' respective locations. It is a combination of two `split` and two `paste`.
- `Jump`: remove a fragment and directly place it next to another. It is a combination of an `eject` and a `insert`.

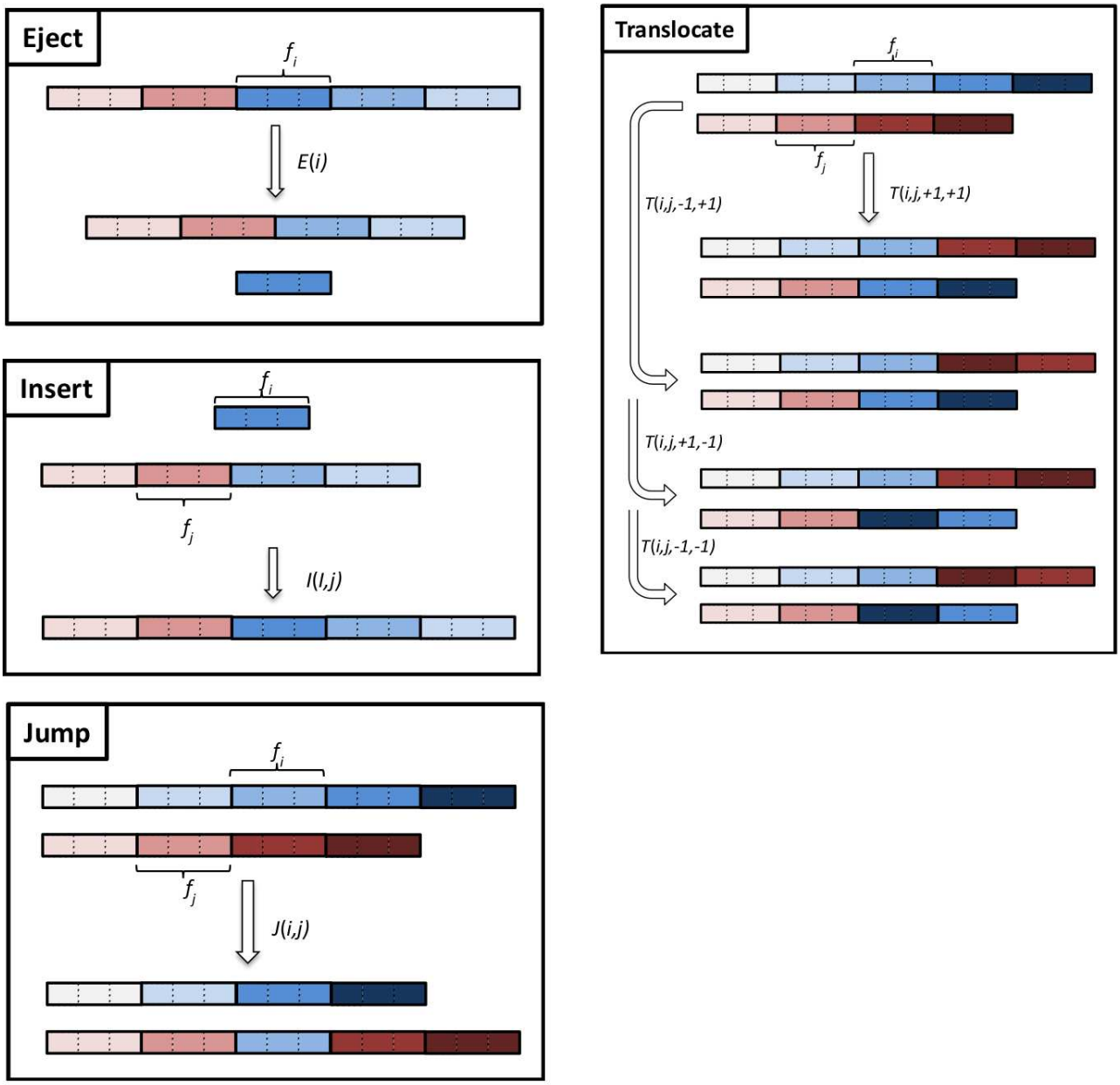


Figure 43: The advanced genome mutations that are compositions of elementary ones: `eject`, `insert`, `translocate`, `jump`.

Source: Adapted from Marie-Nelly *et al.*, 2014, [338].

These operations allow accelerated changes and lighten the computational load.

The GRAAL workflow Now that we have a way to change the genome in discrete units and evaluate the likelihood of any change in genome or parameters, we can proceed

with the workflow, as illustrated in figure 44. Each step is a combination of an update in parameters and a genome mutation.

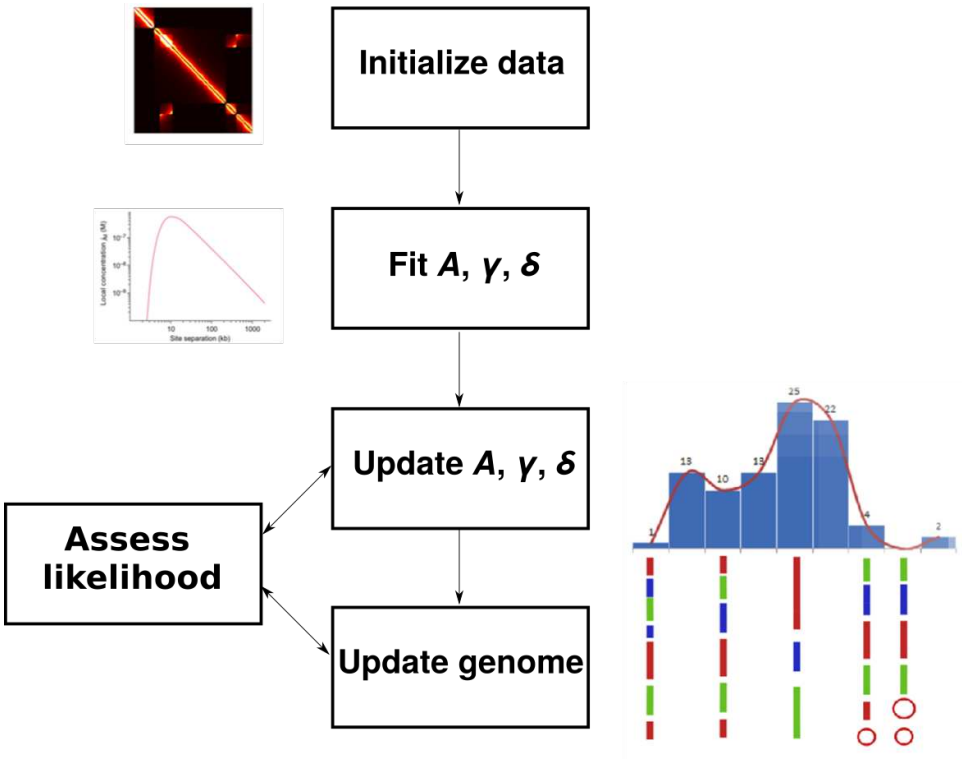


Figure 44: The GRAAL workflow. Data is initialized, parameters are first set to fit the data, then the parameters and genome are iteratively updated.

Source: Inspired from Marie-Nelly *et al.*, 2014, [338].

An update in parameters ξ proceeds as follows:

- Pick a parameter $\theta \in \{A, \gamma, \delta\}$ at random.
- Take $\epsilon_\theta \sim \mathcal{N}(0, \sigma_\theta)$ from a normal distribution with a parameter-specific standard deviation σ_θ, and set $\theta^* = \theta + \epsilon_\theta$. One obtains a new set of candidate parameters ξ^*.
- Accept ξ^* with the probability $r = \min(1, \frac{L_{\xi^*}(M)}{L_\xi(M)})$ where $\frac{L_{\xi^*}(M)}{L_\xi(M)}$ represents the likelihood ratio between the previous and modified parameters.

An update in genome follows a modified version of a multiple-try Metropolis algorithm (MTM):

- Pick a fragment m_i at random with uniform probability.

- Pick a number of k different *neighbors* $(m_1, \ldots, m_k)$. They are drawn with probability $N_i(j) = \frac{m_{ij}}{\sum_{l \neq i} m_{il}}$, in order to be biased towards fragments with close 3D proximity, since these are the ones to perform operations on.
- Consider the set of all new genomes $\mathcal{G}$ that would be obtained by performing each of the nine above mutations on each of the neighbors, separately. It is in practice higher than $9k$ because some mutations (such as `translocate`) may yield different genomes depending on the orientation chosen. Compute the likelihood $L(G), G \in \mathcal{G}$ for each of the corresponding matrices.
- Pick one, G with probability $L(G)N_i(j)$ (where $N_i(j)$ is the neighborhood weight function defined above for each neighbor j).
- Accept G and update the new genome.

This is an accelerated version of the traditional MTM algorithm because G is accepted right away. It is not a time homogeneous Markov chain anymore, but it requires less computations and remains highly efficient.

GRAAL in practice The program operates by *cycles*: each fragment m_i is assessed for a mutation (and the corresponding parameters ξ updated accordingly) and once all fragments have been iterated this way, a new cycle begins. After a number of cycles, convergence usually becomes clear and the genome is considered reassembled. The program was initially tested on *S. cerevisiae*, *Trichoderma reesei* and several human chromosomes. An example is shown in figure 45.

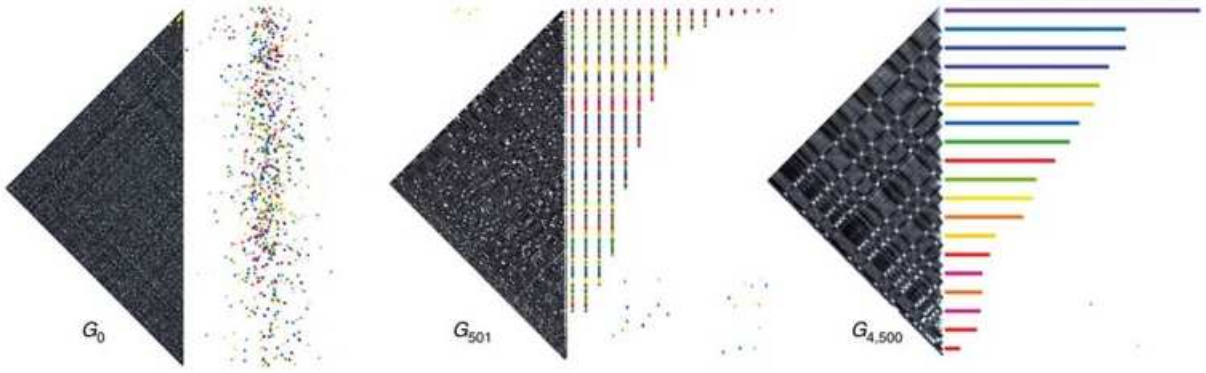


Figure 45: An example of GRAAL scaffolding on the sixteen chromosomes of *S. cerevisiae*.

Source: Inspired from Marie-Nelly *et al.,* 2014, [338].

There are several strengths to the approach:

- The algorithm requires little prior information about the final scaffolding and is in fact unbiased by the starting genome thanks to its Markov Chain nature. The

implementation offers the option of completely splitting every single fragment prior to re-assembly.

- The base design is relatively flexible and can be extended. For instance, the three-parameter model $\xi = (A, \gamma, \delta)$ could be refined into a four- or five-parameter one if more sophisticated models become known. Likewise, designing and implementing new candidate mutations is relatively simple.
- Relatively little coverage is required. As long as the basic assumptions about the contact distributions are respected (more intra- than inter-chromosomal contacts and a strictly decreasing function) a scaffolding is possible.
- Contrary to other methods such as gradient descent, the algorithm can't be "trapped" in local minima. Assuming $P(s)$ conditions are respected, convergence is guaranteed.

However, there are also caveats to be mindful of:

- The contact map can't be normalized. Since only raw contact counts (as generated by the Poisson process) are assessed, the matrix can't be reduced to frequency vectors. This means some of the biases mentioned previously still carry over to the algorithm.
- Coverage heterogeneity can heavily disrupt any attempt to fit a proper $P(s)$ function. Some can be very little covered, some will heavily bias contacts toward them. Since the matrix can't be normalized, some of the most egregious fragments have to be filtered beforehand.
- As discussed before, GC content and fragment size also bias the contact distribution.
- The program is nondeterministic. Any GRAAL run on a given reference will yield different assemblies every time, although after convergence they will be very similar.
- By design, the program tends to disregard any prior information about the genome with *burn-in* cycles. This may or may not be desirable depending on how much one trusts the initial reference, but in most cases one does not actually want to rebuild everything from scratch, lest local artefacts such as spurious inversions or small translocations appear within a contig.

Therefore, the approach needs some improvements and fine-tuning to be exploited to its full extent.

Refining GRAAL into instaGRAAL - PhD project In this section we have laid out and summarized the basic principles behind the original GRAAL scaffolder. The principle was proven to be effective, but only tested on relatively small and low-complexity genomes and without the global picture of a complete integrated genome assembly solution in mind. This is crucial if one is to reliably interpret dynamic events such as rearrangements. Our work thus begins from there: we have successfully implemented an updated version of the scaffolder, dubbed *instaGRAAL*. Among other things, it addresses several key points:

- The *scalability* of the program, which needs to work on genomes that are hundreds of megabases or gigabases in order to be useful in practice.
- The necessity for a form of *polishing* after scaffolding the genome. Since GRAAL introduces artifacts, and the initial reference is (in most cases) a useful source of prior information about the genome structure, one needs to re-inject that information into the new scaffolding so as to correct any spurious mutations introduced by the program.
- The *modularity* of the program needs to be emphasized so as to adapt to all possible case studies. Our work will demonstrate that we have had to tackle very different species and unveil a wide range of dynamic events. Adjusting various hyper-parameters of the program such as the distribution coverage, the binning factor b (constant or not), the number of neighbors k, the range of possible mutations, etc. has proven necessary to yield the best results.
- A possibility to *integrate* independent data sources such as long reads or genetic maps. A sound strategy to infer the proper order of contigs when given several information streams is necessary to obtain a properly complete assembly.

Over the course of this work, we successfully demonstrate its results on three case studies:

- The brown alga *Ectocarpus sp.*, whose 27-chromosome, 200 Mb genome serves as a showcase species to demonstrate the efficiency of our program. Armed with the completeness and misassembly metrics that we have mentioned before, we show that our new program yielded the highest quality genome ever for that strain.
- The joint re-assmebly of two *Trichoderma reesei* strains (QM6A and RutC30) shows a rearrangement, which was confirmed by the literature and fits its evolutionary history. GRAAL has had a successful precedent with that species and our case study is its natural continuation.
- The joint reassembly of two different lineages of the desert ant *Cataglyphis hispanica* shows a chromosome fusion between one lineage and the other, thus going from 27 to 26 chromosomes. Not only were these the first ever high quality assemblies for this species (in either lineage), but the dramatic rearrangements could

give some hints as to the peculiar reproductive (hybridogenetic) strategy of the *C. hispanica* queens.

We also integrate this scaffolding process into a global assembly and validation approach aiming at producing reference-grade quality assemblies and leaving no ambiguity to the dynamic events that we have uncovered. The results are shown in chapter 3.

1.4.7.2 3C-based metagenome binning

The next natural step to genome assembly is metagenome assembly, and the dynamics implications are crucial if one is to understand the interplay among bacteria and between phages and bacteria. Indeed, there is growing evidence that *e.g.* the human phageome dramatically affects the gut microbiome, but the mechanisms themselves are still in their early steps [342] [343]. In the previous sections we have outlined the difficulties at stake if one is to preserve interactions of interest while assembling a metagenome. Here we will demonstrate how our approach alleviates these challenges and allows a bias-free metagenome reconstruction.

Naive reconstruction In order to know whether a 3C-based approach could help reconstruct the genomes found in a complex sample, a simple question would be to test it on a controlled community of relatively few bacteria. In 2014, a proof of concept was achieved with the following mix [344], illustrated in figure 46.

- *Bacillus subtilis*
- *Escherichia coli* with its F plasmid
- *Vibrio cholerae*

A 3C experiment is thus performed on the mix, the resulting library sequenced and assembled *de novo* as though one did not have access to the reference, in order to mimic future conditions in metagenomic experiments. The scaffolds that do make up the majority stretch of each genome act as references for the mapping.

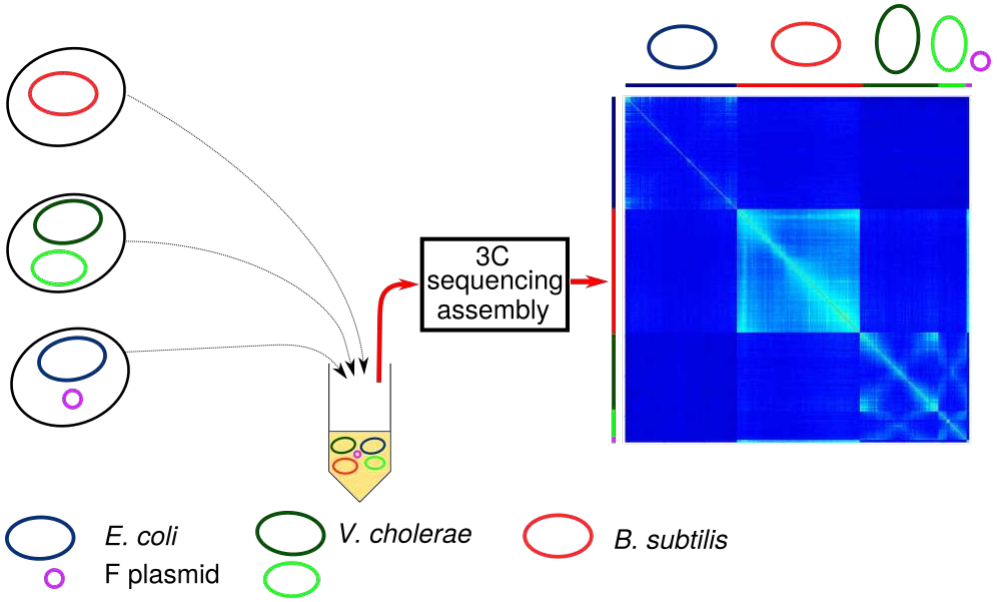


Figure 46: A meta3C experiment on a controlled mix. Three bacteria are indiscriminately sequenced, assembled *de novo* and mapped.

The choice of a two-chromosome bacterium (*V. cholerae*) as well as a bacterium featuring a plasmid unveils an interesting consequence of the experiment: not only a 3C library contains the relevant information to successfully sort the mix into its original three bacteria, but one observes different *levels of noise* between the two chromosomes of *V. cholerae* or between the *E. coli* chromosome and its plasmid. These, although relatively low, are still noticeably higher than the standard inter-species noise.

This lets us envision a hierarchy of compartmental divisions that would allow a full deconvolution on multiple scales at once, as seen in figure 47.

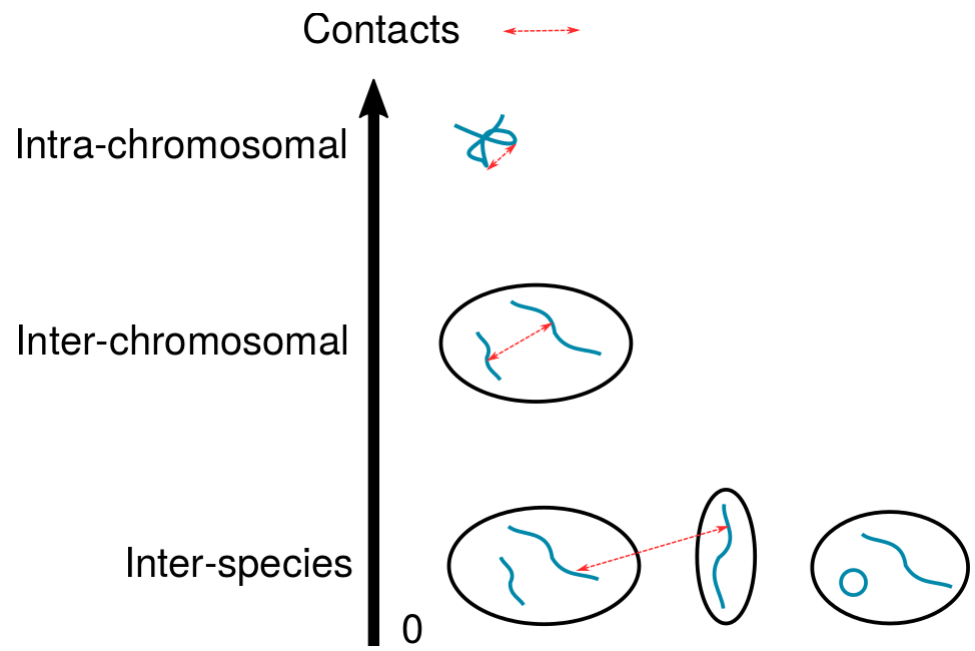


Figure 47: Multi-scale view of contact levels. The order of magnitude of contacts among chromosomes increases the more compartments are shared.

Complex networks and the Louvain algorithm In order to test the hypothesis further, a mix with eleven different yeast strains was tested. This means 11x16 = 176 chromosomes were to be reconstructed, each within their own cell compartment. This, combined with sequence homology issues, would make a naive *de novo* assembly too complex, even with GRAAL.

In order to obtain a preliminary layout of the network, we use the *Louvain algorithm*, an original method borrowed from social networks analysis and used for community detection [345]. We use the same graph-based approach that we have mentioned in figure 34 and partition the *network* of contacts into *communities*. The idea is that sequences (nodes) within a community should see (interact with) each other more than they see sequences outside their community, or what would be expected by chance. This condition is formally known as the *Newman-Girvan criterion*.

In order to quantify this intra/inter relationship among communities, the algorithm makes use of a metric called the *modularity* on the partitioned network. It is defined as follows: let $M = (m_{ij})_{ij}$ be a contact map representing the entire network (or the adjacency matrix of the network). Let C be a set communities that completely partition of the network, such that $c_i \in C$ represents the community to which the node i belongs. Let δ be a simple delta function representing the intra- or inter-community status of a node pair (i, j):

$$\delta_{ij} = \delta(c_i, c_j) = \begin{cases} 1 & c_i = c_j \\ 0 & c_i \neq c_j \end{cases} \tag{13}$$

Let $k_i = \sum_l m_{il}$ the sum of all (possibly weighted) edges attached to the node i (or alternatively, the sum of all contacts in the bin vector i), and $m = \sum_{i \neq j} m_{ij}$ the sum of all edges in the network (or non-diagonal elements in the matrix). The modularity Q is given by:

$$Q = \frac{1}{2m} \sum_{ij} (m_{ij} - \frac{k_i \cdot k_j}{2m}) \cdot \delta_{ij} \tag{14}$$

By definition, the modularity of a partition lies between -1 and 1. At -1, there are only ever inter-communities contacts. It is 0 if nodes are spread in communities as though the partition was random, and it is positive if there are more nodes within communities than an hypothetical random rewiring of the network.

The Louvain algorithm seeks to maximize a network's modularity. Since finding an absolute maximum is a computationally hard problem, it instead focuses on a relatively quick heuristic. It proceeds in a multiple pass approach. The first pass (detailed in algorithm 3 computes, for each node i, the global modularity shift ΔQ incurred by moving the node from its own community to one of its neighbors, and joins the one maximizing the increase if it is positive. The process repeats until no more increase can be found.

Algorithm 3 Louvain single-pass algorithm

Require: $\mathcal{G}, \mathcal{E}$, a network of nodes and edges
Require: $\mathcal{C}$, a partition mapping a node i to its community c_i
Require: $V_{\mathcal{G},\mathcal{E}}(i)$, a function returning the neighbors of i
Require: Q, a modularity function and ΔQ, computing the modularity shift $\Delta Q(i \to j)$ for moving i towards c_j
  repeat
    for i in $\mathcal{G}$ **do**
      $k \leftarrow \arg\max_{j \in V_{\mathcal{G},\mathcal{E}}(i)} \Delta Q(i \to j)$
      if $\Delta Q(i \to k) > 0$ **then**
        $c_i \leftarrow c_j$
      end if
    end for
  until no more increase is possible

In a second pass, all nodes within each community merge to become a single node representing it, and the above algorithm is run once again on the new graph. The process can be iterated repeatedly and is illustrated in figure 48. A remarkable feature of the method is that it runs in $O(n \log(n))$ time: as the nodes get merged, the global data structure adopts a hierarchical, tree-like layout and the algorithm gets faster as it runs through it.

The first application of Louvain algorithm on Hi-C data was successfully done to the aforementioned mix of elven yeast species. All sequences successfully clustered to a community matching a species. Not only that, but figure 49 shows that within each

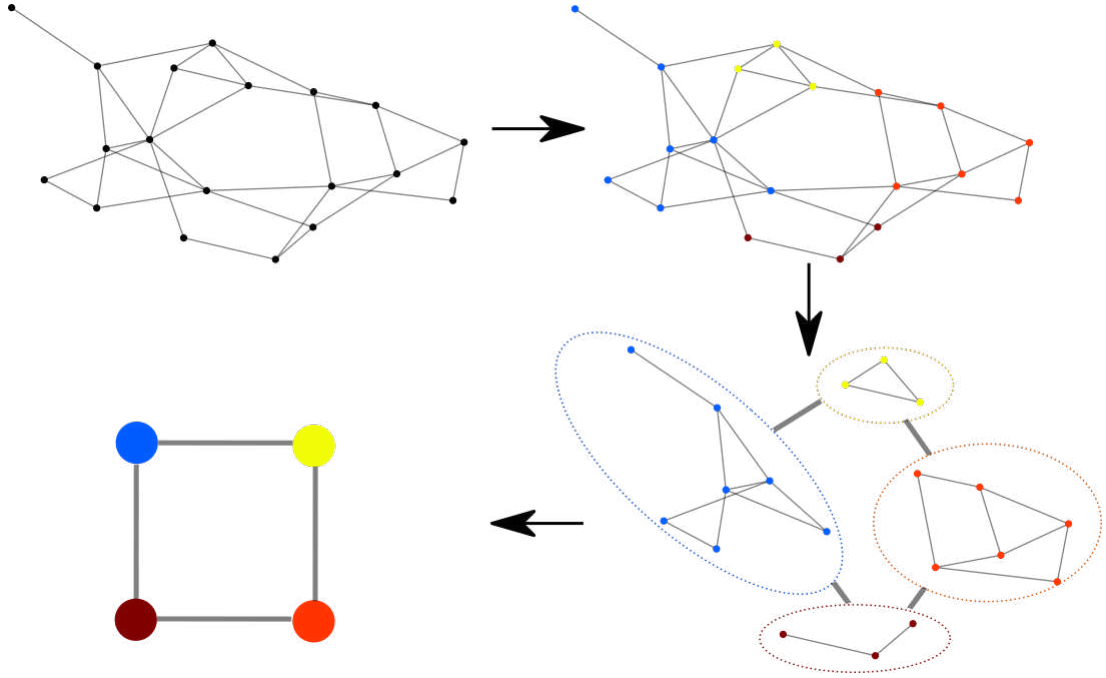


Figure 48: Partitioning of a non-weighted network (top) into four color-coded interaction communities (bottom).

cluster, GRAAL was able to scaffold the corresponding sequences into their respective sixteen chromosomes.

This proof-of-concept work lets us draw an outline of the workflow underlying the basis of our present work and illustrated in figure 50: given a complex sample, perform a preliminary metagenomic assembly on the reads, use a form of binning with a clustering algorithm drawing from the contact information, and optionally reassemble each bin with our (insta)GRAAL scaffolder.

The approach has several strengths:

- Any metagenome binning based on 3C reads is not hampered by the biases we have described in section 1.2.4.2. Sequences that are completely different in coverage and composition will nevertheless cluster together if the 3C data indicates that they belong to the same compartment. This opens new perspectives for the isolation of phages or identification of phage-host relationships and DNA transfers.

- The methods used are scalable and proven. Other criteria than the Newman-Girvan one exist [346], and the Louvain algorithm isn't the only way to partition a network according to this criterion, but it is relatively fast and known to work on a variety of data. Notably, it has been shown to be the most suitable algorithm for clustering simulated 3C data [347].

- The multi-scale aspect means that many different hierarchies can be unveiled all at

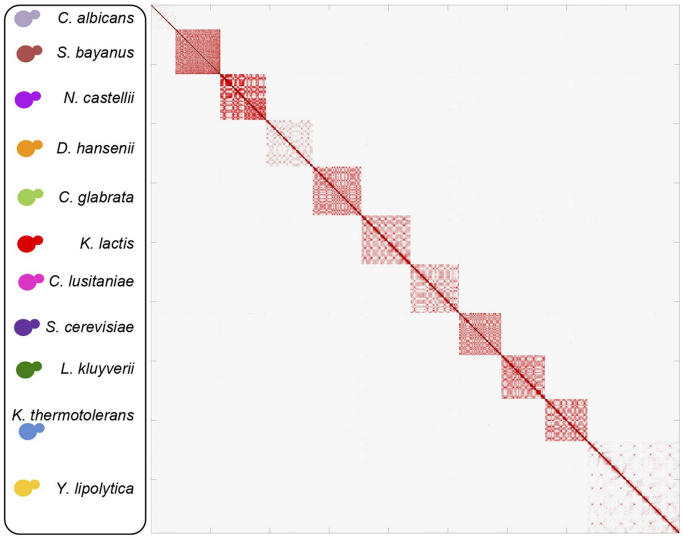


Figure 49: Global contact map of eleven yeast strains with their respective reconstructed chromosomes.

Source: Adapted from Marbouty *et al.*, 2014, [344]

.

once, from broad dynamics within families of species all the way to 3D information about each DNA molecule for each bacterial species.

However, there are still a number of limitations:

- Modularity-based approaches are known for having a resolution limit [348]. When a network grows very large, nodes will randomly merge within the same community, as any neighborhood will appear as a cluster when compared to the large portion of the network with which it doesn't interact.
- The Louvain algorithm is non-deterministic. The output partition largely depends on the order nodes are iterated, etc. This may prove problematic when attempting to cluster shared sequences or sequences that "hop" among genomes due to DNA transfers or other dynamic events.
- Sometimes the contacts can be few and far between, in which case falling back to traditional binning methods may be desired to reconstruct more genomes.

These need to be addressed in order to fully deconvolve a metagenome and understand its dynamics.

Expanding and implementing the meta3C design with metaTOR In this section we have explained the basics of meta3C. The approach had not been tested *in vivo*

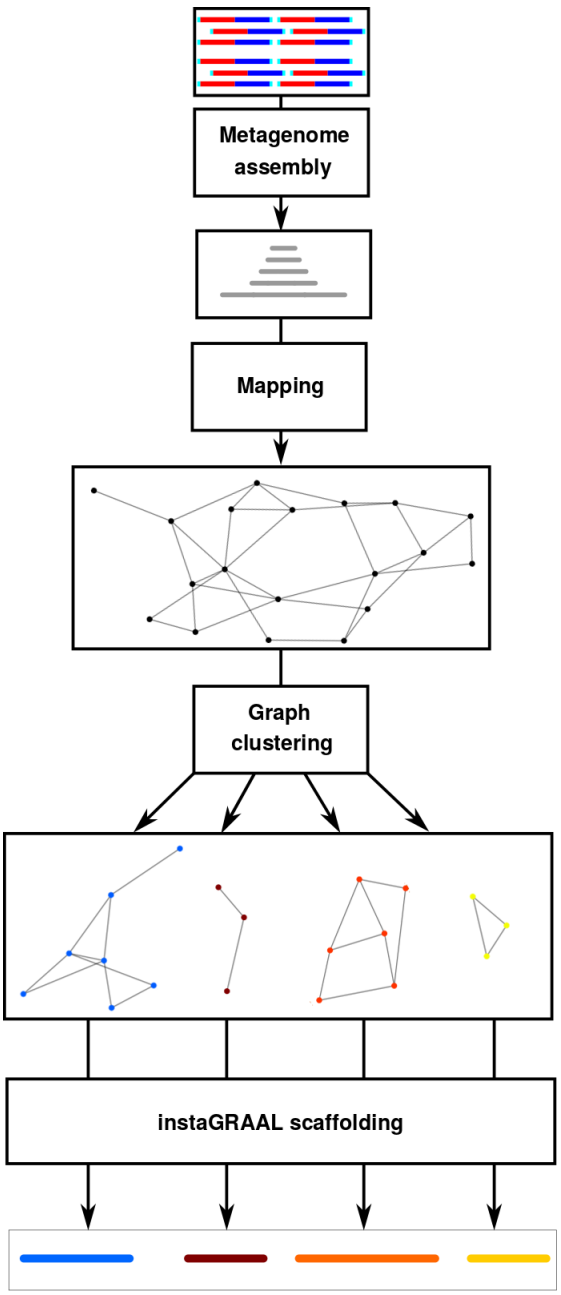


Figure 50: A meta3C workflow. Binning is performed on a preliminary assembly based on 3C contacts before the bins can be scaffolded separately.

prior to starting our work. Moreover, knowledge about the human gut phageome and implications of the relationship between the phageome and the microbiome were only emerging [349] [342] [343], and DNA transfers hard to characterize. Our work provides a further step to help investigate this complex interplay, notably by addressing the above issues:

- The non-deterministic nature of the algorithm can in fact be exploited by running it many times and thus computing a *clustering score* between two sequences in the contact network. This lets us refine the previously binary community classification into a more quantifiable scale. It also lets us identify so-called *core communities* composed of sequences that exclusively cluster together
- The resolution limit can be addressed by recursively running the algorithm onto the subnetworks it has identified. If any artifact arises, it will end up deconvoluted into smaller sub-communities. This also gives us an additional, hierarchical angle of view into the nature of the relationships between genomes.
- Armed with this new design, we thus establish a benchmark between our method and traditional ones. However, it is worth noting that both approaches are in fact complementary and can be combined to yield even more reconstructed genomes.

These improvements upon the original meta3C design prove successful at reconstructing hundreds of genomes, most of them mostly or quasi-complete. We also reconstruct phage genomes and isolate phage-host relationships, including the exact coordinates of some prophages. Lastly, we show the latest implementation of our approach (Metagenome Tridimensional Organisation-based Reconstruction or *metaTOR*) is able to 1) reconstruct rare and/or previously unknown genomes, and 2) outperform existing tools by binning more and better-quality genomnes. The results are shown in chapter 4

1.5 Our thesis work on 3C assembly: increasing layers of complexity

In this chapter we have laid out the general principles of genome sequencing, assembly, the implications for studying the dynamics of chromosomes and how Hi-C fits into this global picture. Over the course of this work we will present our main results from successful applications of Hi-C technology to genome assembly.

Chapter 2 focuses on *serpentine binning*, a basic computational tool that we have developed and implemented, using Hi-C data and simple mathematical methods; although it is not directly concerned with genome assembly, it has proven useful for analyzing, interpreting contact data in any form, and fits into the global picture of contact data enhancers described in 1.4.5.2.

Chapter 3 focuses on large eukaryote genome scaffolding and investigating rearrangements: we will first present our tool, instaGRAAL, and how it was successfully applied

to the reassembly of the brown algae *Ectocarpus sp.*. Then we will cover the joint reassembly of two *Trichoderma reesei* strains, and preliminary results concerning two *C. hispanica* lineages.

Chapter 4 takes the approach to another level of complexity with the challenge of metagenome binning. We first demonstrate a successful and pioneering use of meta3C on a mouse sample resulting in the reconstruction of many genomes and the isolation of phage-host relationships, going as far as peering into the infection spectrum of some phages. We then detail the formalized implementation of this design (dubbed metaTOR) and demonstrate its capabilities on multiple mice samples and test it against other tools.

This work can therefore be seen as a progression as we apply our technological and computational framework to increasingly complex subject matters and draw biological insights into the underlying mechanics of chromosome interaction and evolution. From simple Hi-C data to smaller genomes to larger genomes to metagenomes, each subject matter offers a different set of challenges to tackle, methods to design and tools to implement, all with the common threading line of chromosome dynamics.

2 Hi-C methods for enhancing interaction signal

Genome analysis

Serpentine: a flexible 2D binning method for differential Hi-C analysis

Lyam Baudry[1,2], Romain Koszul[1,*] and Vittore Scolari[1,*]

[1]Institut Pasteur, Unité Régulation Spatiale des Génomes, UMR3525, CNRS, 75015 Paris, France,

[2]Sorbonne Université, Collège Doctoral, F-75005, Paris, France

*To whom correspondence should be addressed.

Abstract

Motivation: Hi-C contact maps reflect the relative contact frequencies between pairs of genomic loci, quantified through deep-sequencing. Differential analyses of these maps facilitate downstream biological interpretations. However, the multi-fractal nature of the DNA polymer inside the cellular envelope results in frequency values spanning several orders of magnitude: contacts involving loci pairs at large genomic distance are much sparser compared to closer pairs. The same is true for poorly covered regions such as telomeres and repeated sequences. Poor coverage translates into low signal-to-noise ratios. There is no clear consensus to address this limitation.

Results: We present a fast, flexible procedure operating on simple data that takes into account the contacts in each region of a contact map. Binning is performed only when necessary on noisy regions, preserving informative ones. This results in high-quality, low-noise contact maps that can be conveniently visualized for rigorous comparative analyses.

Availability: The software is available on the PyPI repository and https://github.com/koszullab/serpentine. Documentation and tutorials are provided at https://serpentine.readthedocs.io/en/latest/.

Supplementary information: Supplementary data are available at Bioinformatics online.

Contact: vscolari@pasteur.fr or rkoszul@pasteur.fr

1. Introduction

Chromosomal conformation capture experiments provide a quantitative way to infer the spatial proximity of DNA segments (Hi-C contact maps, (Lieberman-Aiden *et al.*, 2009). Downstream analyses include normalization, contact quantification, 3D pattern recognition, etc. However, experimental variability can influence data analysis in slight yet irreproducible ways. While this does not affect the analysis where robust trends are not altered by noise, the quality of small-scale and local comparisons suffers in poorly-covered regions, limiting comparisons. A variety of sophisticated models taking into account the contact distribution (Lun and Smyth, 2015; Stansfield *et al.*, 2018) have been developed to tackle these limitations, but these packages don't deal with very sparse information. Other approaches bin pixels (i.e. sum-pooling) to increase the signal-to-noise ratio in regions with few or no contacts at the expense of resolution, but do so over the entire map.

We have shown that if binning is uniform over the entire matrix, the distribution will be dominated by sampling noise (Poisson distribution) at large genomic distances, while at smaller genomic distances the resolution will be limited by the bin size, losing the opportunity to observe high-resolution features. More information on the rationale is provided in supplementary material (section 1).

To tackle the lack of a suitable uniform resolution over the entire map, we developed a normalization-free, flexible method that only bins low-covered regions. The procedure makes no assumption about the contact distribution and only alters it locally. It requires two contact thresholds beyond which local binning stops. Parameters can be chosen according to the data or automatically inferred from the contact distribution. When applied to low-resolution positions, it unveils hidden patterns and improves the quality of log-ratio maps.

2. Method

Joint binning and comparison: Serpentine-shaped bins randomly aggregate with nearby pixels depending on their values (Figure 1A). Given two input thresholds t and $m < t$, a serpentine stops aggregating if the sum of the values of its pixels is lower than t in *any* contact maps, or lower than m in *all* of them. The log-ratio of the serpentine-binned matrices is then computed to visualize pattern differences (Figure 1C, right matrix, figure 1D and 1E, top-left half matrices). Full details on the algorithm are described in supplementary materials (section 3 and 4).

Application: we applied serpentine binning to compare two published *Saccharomyces. cerevisiae* Hi-C datasets binned at 2.5 kb during meiosis (Muller *et al.*, 2018; Figure 1B; Supplementary information). The data is not normalized. The log-ratio contact map (Figure 1C, leftmost matrix and plot) shows significant noise and local variance across poorly covered regions. The MD-plot distribution showing the log-ratio vs. log-average values displays a large divergence at small average values, corresponding to sampling noise.

Comparison with other methods: We compared serpentine binning with classic re-binning and Gaussian kernel convolution, as both methods are normally used to improve map visualization (Figure 1C). Although the signal-to-noise ratio improves globally when using either method, short-scale events are dwarfed by the rest of the signal. Moreover, noisy regions are not improved. On the other hand, serpentine binning smoothens low-covered regions, confirming that differences between datasets in these areas are not significant, while strong short-scale patterns emerge. This remains true in down-sampled matrices (Figure 1D; Supplementary materials, section 6).

Observation of hidden patterns: we applied our binning algorithm to detect increases in trans- homologous interactions after 4h into meiosis (Figure 1E, top-right matrix). The pattern correlates with *cis* loops bridging Rec8 binding sites, identifiable from ChIP-chip (Ito

et al., 2014). It is not visible in the raw ratio map (bottom-left matrix); overall, it points at potential inter-chromosomal contacts to be further investigated.

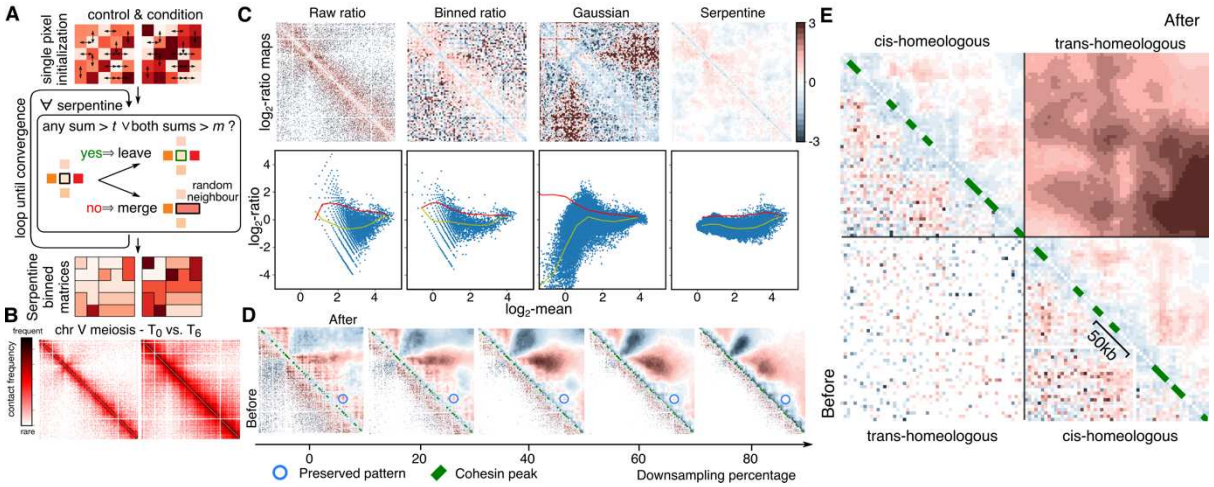


Figure 1. (**A**) Algorithm flowchart. (**B**) Input matrices (chromosome V). Left: T0 (interphase, control). Right: 6h into meiosis. (**C**) Log-ratio of (from left to right) raw maps, after re-binning, after gaussian convolution, after serpentine binning (chr. V, T0 vs. 4h). (**D**) Down-sampling effects (chr. V, T0 vs. 6h). Top-right: serpentine binning. Bottom-left: raw ratio. Loops (blue circle) form in meiosis between cohesin-enriched positions (green dots in the diagonal). In serpentine-binned matrices a strong loop can still be identified after down-sampling. (**E**) cis- and trans- homeologous contacts (SynHiC region, T0 vs 4h).

Acknowledgements

The authors thank Julien Mozziconacci and Axel Cournac for feedback. Conflict of Interest: none declared.

Funding

This research was supported by funding to R.K. from the European Research Council under the Horizon 2020 Program (ERC grant agreement 260822).

References

Ito,M. *et al.* (2014) Meiotic recombination cold spots in chromosomal cohesion sites. *Genes to Cells*, **19**, 359–373.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289–293.

Lun,A.T.L. and Smyth,G.K. (2015) diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, **16**, 258.

Muller,H. *et al.* (2018) Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for Hi-C. *Molecular Systems Biology*, **14**, e8293.

Stansfield,J.C. *et al.* (2018) HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics*, **19**, 279.

Supplementary materials

1. Principles

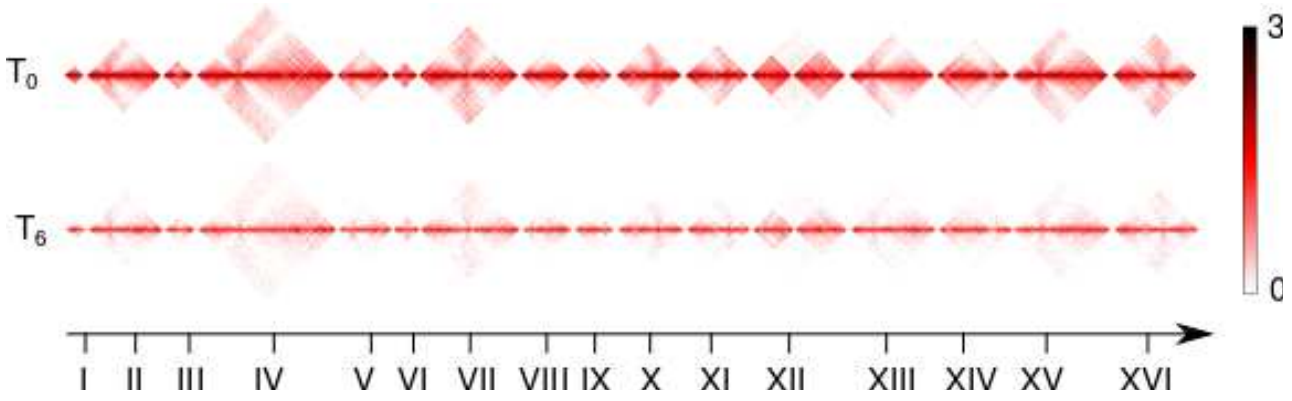
In Muller et al., 2018 we have shown that at a fixed genomic distance *s*, for high enough coverages *C*, the distribution of contacts modelling the observed pixel values can be characterized by 1) a mean $\mu(s, C)$, measuring the polymeric signature $P(s)$, and 2) a standard deviation $\sigma(s, C)$, reflecting an estimation of the biological variability. Both functions, in this interpretation, are strictly proportional to *C*. However, the experimental generation of a contact map is altered by random sampling, happening at the PCR-amplification and sequencing level. This process results in Poisson distributed values.

The Poisson distribution is characterized by all its cumulants being equal to μ. For this reason, when PCR-driven random sampling is the dominant process, $\sigma = \mu^{1/2}$, and thus the biological variability contained in $\sigma(s, C)$ is lost. This happens for all values of $C < t$, where *t* is a threshold such that $\sigma(C) < \mu(C)^{1/2}$.

To overcome this effect, we locally bin the matrices using serpentine binning and make sure that the values of coverage for each bin are all above such a threshold.

2. Contact map generation

Data was taken from (Muller *et al.*, 2018, BioProject PRJNA464299). Reads were mapped, each end independently, using the Bowtie 2 aligner with the –very-sensitive-local option against the *S. cerevisiae* reference genome taken from (Yue *et al.*, 2017). An iterative alignment procedure was used: for each read, the length of the sequence being aligned was gradually increased by 20 bp steps until the mapping became unambiguous (mapping quality > 30). Paired reads were aligned independently, and each mapped read was assigned to a restriction fragment. Alignments were filtered for artifacts as described in (Cournac *et al.*, 2012) and binned along 2.5 kb sequences.



Supplementary Figure 1: Contact maps of all sixteen chromosomes of S. cerevisiae at T0 and six hours into meiosis

3. Joint binning and comparison

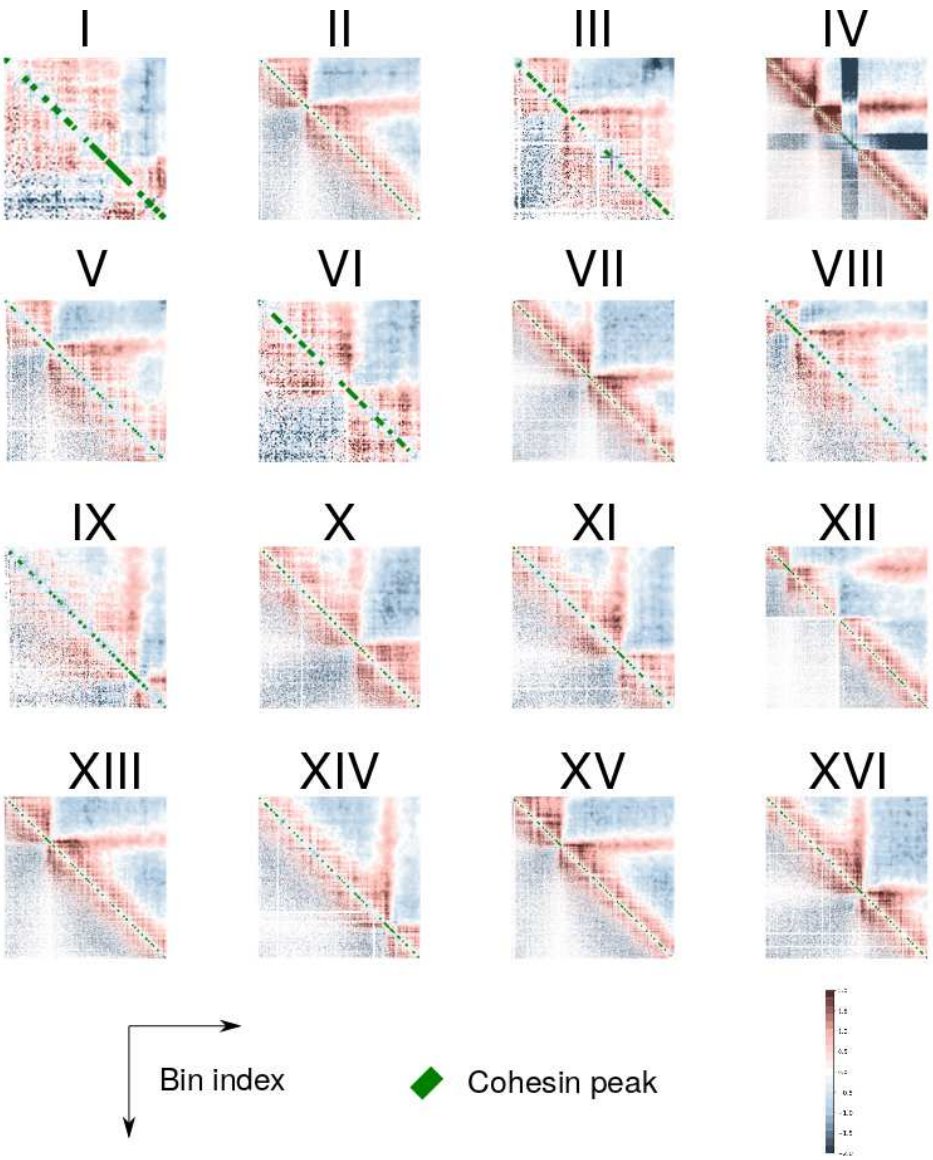
The binning algorithm requires two input thresholds t, $m < t$ and at least two input contact maps; typically, a contact map in a given experimental condition is compared to a control map. A serpentine is a subset of pixels identifying a single connected region (along the four spatial directions) in the pixels set M. Each pixel in M is initialized as a serpentine singleton. Then, a serpentine is drawn randomly. If the sum of the coverage of its pixels is lower than t in *all* contact maps, or lower than m in *any* of them, then it is suitable to merge with another serpentine randomly chosen among its neighbours. Two serpentines are neighbours if they have at least one pair of adjacent pixels. The flowchart for merging is illustrated in figure 1A. Once all serpentines have been iterated over, the process begins anew until the total number of serpentines remains constant across two iterations, indicating that the serpentine structure cannot evolve further. The resulting contact maps are then binned serpentine-wise, i.e. each pixel value is replaced with the average value of its final serpentine. The algorithm is run independently N ($N > 4$) times to ensure serpentines are not biased toward any specific 2D direction. The final binned matrix is the average of all binning runs. The log-ratio of the serpentine-binned matrices is then computed to visualize pattern differences (figure 1C, right matrix, figure 1D top-left half matrices).

4. Parameters optimization

In the absence of biological replicates, users can use MDplots (figure 1C, bottom plots) to determine an appropriate value for the threshold t. The goal is to have a uniform noise-to-signal ratio that does not depend on signal intensity. This choice is justified by the fact that the MDplot divergence at low coverages is largely driven by random sampling, rather than biological variability. When biological replicates are available, the coverage threshold under which the effects of sampling becomes dominant over technical and biological variance can be otherwise estimated.

5. Chromosome-level differential analysis

We generated a contact map for each chromosome of *S. cerevisiae* as described above. They are displayed in Supplementary Figure 2. We computed the corresponding log-ratios between contact maps (excluding infinite and undefined values) and performed our serpentine binning procedure on each log-ratio. As the algorithm is not normalization-dependent and log-ratios cancel out most biases, we didn’t perform any normalization. The contact maps of all chromosomes (before and after serpentine binning) are displayed in supplementary figure 2, complete with the coordinates of cohesin peaks.



Supplementary Figure 2: Log-ratios of contact maps before and after 6 hours taken from Muller et al., 2018. The raw log-ratio data points are on the lower left corner and the serpentine-binned contact map is on the upper right corner of each map. An artifact due to sequence capture on the Syn-Hi-C region can be observed in chromosome IV.

6. Down-sampling

To benchmark our method, we applied the algorithm on two matrices that we down-sampled in decreasing proportions. Serpentine binning highlights the strongest patterns in the full contact map even at high rates of down-sampling, whereas the raw log-ratio fails to do so (Figure 1D).

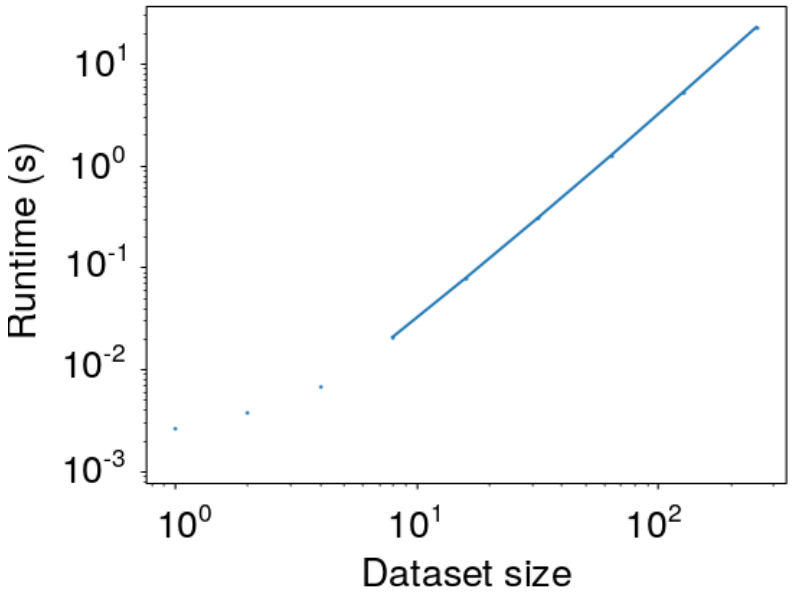
7. Comparison of cis- and trans- contacts

The comparison of cis- and trans- contact is obtained by performing serpentine binning independently on the two cis- and the single trans- sub-matrices, in order to avoid bins that span

different regions. The matrices are finally merged and the log-ratios are normalized on the mean contact values taking over the sole cis- region of the matrix. This assumes that the general polymeric compaction is less altered than the cis- to trans- contact ratio.

8. Performance

The algorithm's runtime has worst-case complexity $O(n^2 \log(n))$, which roughly corresponds to a serpentine structure necessitating the maximum amount of merging at every iteration. Supplementary Figure 3 shows experimental benchmarks confirming it. In practice, the runtime is dominated by the initial iteration (involving n (n + 1) / 2 serpentine singletons) acting as the bottleneck of the total algorithm runtime, and subsequent iterations run much faster.



Supplementary figure 3: The runtime scales as n^2 where n defines the dimensions of the input matrices (n x n).

Supplementary references

Cournac,A. *et al.* (2012) Normalization of a chromosomal contact map. *BMC Genomics*, **13**, 436.

Muller,H. *et al.* (2018) Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for Hi-C. *Molecular Systems Biology*, **14**, e8293.

Yue,J.-X. *et al.* (2017) Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics*, **49**, 913–924.

3 Genome assembly and uncovering intra-species genome dynamics

3.1 The instaGRAAL scaffolder

In the introduction we have presented the main motivation for building upon GRAAL to improve genome assemblies. This section concerns our submitted work, presenting the following main results:

- The implementation of instaGRAAL as an improved version of GRAAL, notably featuring post-scaffolding polishing as well as genetic map integration
- The immediate application of our program on the genome of the model brown alga *Ectocarpus sp.*, complete with extensive validation, thus yielding the highest-quality genome for this species.

This work acts as a preliminary step to investigate genome dynamics once they are fully scaffolded and validated.

Chromosome-level quality scaffolding of the *Ectocarpus* sp. *g*enome with instaGRAAL, a proximity ligation-based scaffolder

Lyam Baudry[1,2,3], Martial Marbouty[1,2], Hervé Marie-Nelly[1,2,3], Alexandre Cormier[4], Nadège Guiglielmoni[1,2], Komlan Avia[4], Lieven Sterck[4], J. Mark Cock[4], Christophe Zimmer[2,5], Susana M. Coelho[4,&], Romain Koszul [1,2&]

[1] Institut Pasteur, Unité Régulation Spatiale des Génomes, UMR3525, CNRS, 75015 Paris, France

[2] Institut Pasteur, Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), USR3756, CNRS

[3] Sorbonne Université, Collège Doctoral, F-75005, Paris, France

[4] Sorbonne Université, Laboratory of Integrative Biology of Marine Models, Algal Genetics, UMR 8227, Roscoff, France

[5] Institut Pasteur, Imaging and Modeling, UMR3691, CNRS, 75015 Paris, France

[&] contacts: romain.koszul@pasteur.fr or coelho@sb-roscoff.fr

ABSTRACT

Hi-C has become a popular technique in recent genome assembly projects. Hi-C exploits contact frequencies between pairs of loci to bridge and order contigs in draft genomes, resulting in chromosome-level assemblies. We developed instaGRAAL, a complete overhaul of the program GRAAL suited for large genomes, which uses a Markov Chain Monte Carlo algorithm to perform Hi-C scaffolding. InstaGRAAL features a number of improvements, including a modular polishing approach that optionally integrates independent data. To validate the program, we used it to generate a chromosome-level assembly for the model brown alga *Ectocarpus* sp., and quantified improvements compared to the initial draft.

Keywords: Ectocarpus, Hi-C scaffolding, Hi-C, genome assembly, MCMC, GPU, parallel computing

Background

Despite continuous and impressive developments in DNA sequencing technologies, technical challenges remain regarding the assembly of sequence data into full length chromosome assemblies, especially for large genomes [1,2]. Conventional assembly programs and pipelines often encounter difficulty closing gaps in draft genome assemblies caused by regions enriched in repeated elements. At the chromosome level, these programs often incorrectly orient DNA sequences or predict incorrect numbers of chromosomes [3]. Conventional assemblers efficiently generate overlapping set of reads (i.e. contiguous sequences, or contigs) but encounter difficulties linking these contigs together into scaffolds. Consequently, many available genomes feature gaps which need to be bridged to reach a chromosome-level structure. These computational limitations are being addressed thanks to active support from the community and competitions such as GAGE [4] or the Assemblathon [5] but there is as yet no systematic, reliable way of producing near-perfect genome assemblies of guaranteed optimal best quality without a considerable amount of empiric parameter adjustment and manual post-processing evaluation and correction [6]. [6]

Recent sequencing projects have typically relied on a combination of independently obtained data such as optical mapping, long read sequencing, and chromosomal conformation capture (3C, Hi-C) to obtain large genome assemblies of high accuracy. The latter procedure derives from techniques aiming at recovering snapshots of the higher-order organization of a genome [7,8]. When applied to genomics, Hi-C-based methods are sometimes referred to as proximity ligation approaches, as they quantify and exploit physical contacts between pairs of DNA segments in a genome to assess their collinearity along a chromosome, and the distance between the segments [9]. Early studies demonstrated that Hi-C scaffolds large eukaryotic DNA regions using control datasets [10–12]. The Hi-C scaffolder GRAAL (Genome Re-Assembly Assessing Likelihood from 3D), a probabilistic

tool that uses a Markov Chain Monte Carlo (MCMC) [12] was able to generate the first chromosome-level assembly of an incomplete eukaryotic genome (see also [13]. Since these proof of concept studies, the assemblies of many genomes of various sizes from eukaryotes [14–16] and procaryotes [17] have been significantly improved using scaffolding approaches exploiting Hi-C data.

Although GRAAL was effective on medium-sized or small (<100 Mb) eukaryotic genomes such as that of the fungi *Trichoderma reesei* [18], scalability limitations were encountered when tackling genomes whose complexity and size required significant computer calculation capacity. Furthermore, as observed also with other Hi-C-based scaffolders, the raw output of GRAAL includes a number of caveats that need to be corrected manually to obtain a finished genome assembly. To tackle these limitations, we developed instaGRAAL, an enhanced, open-source program optimized to reduce the computational load of chromosome scaffolding and that includes polishing steps to automatically complete the assembly process. The polishing, which aims to minimise assembly errors, can exploit available genetic linkage data.

InstaGRAAL was applied to the 214 Mb haploid genome of the model brown alga *Ectocarpus* sp*.,* which is currently only published in draft form [19]. Brown algae are a group of complex multicellular eukaryotes that have been evolving independently from animal and land plants for more than a billion years. *Ectocarpus* was the first species within the brown algal group to be sequenced, as a model organism to investigate multiple aspects of brown algal biology including the acquisition of multicellularity, sex determination, life cycle regulation and adaptation to the intertidal [20–23]. A range of genetic and genomic resources have been established for *Ectocarpus* sp. including a dense genetic map generated with 3,588 SNP markers [24]. Here we used instaGRAAL to generate a high quality, chromosome-level assembly of the *Ectocarpus* genome and used resources generated for

this model organism, in particular the dense genetic map, to comprehensively validate the assembly.

Results

From GRAAL to instaGRAAL

The technical limitations of GRAAL were i) high memory usage when handling Hi-C data for large genomes (*i.e.* over 100 Mb), 2) difficulties when installing the software, and 3) the need to adjust multiple *ad hoc* parameters to adapt to differences in genome size, read coverage, Hi-C contact distribution, specific contact features, etc. We designed instaGRAAL to address all these shortcomings. First, we rewrote the memory-critical parts of the program, such as permutation sampling and likelihood calculation, so that they are computed using sparse contact maps. We reduced the software's dependency footprint and added detailed documentation, deployment scripts and containers to ease its installation. Finally, we opened up multiple hard-coded parameters to give more control for end-users while improving the documentation on each of them, and selecting relevant default parameters that can be implemented for a wide range of applications. These parameters include the size of the neighbourhood to sample for each bin and the relative coverage threshold for retaining bins in the contact distribution (see Discussion). Overall, these upgrades resulted in a program that was lighter in resources, more flexible, and more user-friendly.

Other problems encountered with the original GRAAL program included 1) the presence of potential artefacts introduced by the permutation sampler, such as spurious permutations (e.g. local inversions) or incorrect junctions between bins; 2) difficulties with the correct integration of other types of data such as long reads to resolve conflicts and 3) the need to filter out sequences that were either too short, included repeated motives or had low coverage prior to scaffolding. We addressed these points by implementing correction

strategies that not only identify and remove artefacts but also reinsert problematic sequences, which are initially put aside during the filtering step, into the final scaffolds (see Methods). These steps can exploit linkage data when available. Overall, when compared to GRAAL's raw output, the resulting "polished" assemblies are significantly more complete and more faithful to the actual chromosome structure.

The core principles of GRAAL and instaGRAAL are similar: both exploit a MCMC approach to perform a series of permutations (insertions, deletions, inversions, swapping, etc.) of genome fragments based on an expected contact distribution. The parameters (A, *α* and δ) that describe this contact distribution are first initialized using a model inspired by polymer physics [25]. That model describes the expected contact frequency *P(s)* between two loci separated by a genomic distance *s* (when applicable):

$$P(s) = \begin{cases} max(A \cdot s^{-\alpha}, \delta) : \in tracontacts \\ \delta : intercontacts \end{cases}$$

The parameters are then iteratively updated directly from the real scaffolds once their size increase sufficiently [12]. Each fragment or stretch of adjacent fragments (referred to here as a 'bin', see Methods) is tested in several positions relative to the neighbouring fragments. The likelihood for each arrangement is assessed from the simulated or expected contact distribution, and the arrangement accepted or rejected [12]. This analysis is carried out in cycles. A cycle is completed when all fragments of the genome have been processed in this way. Any number of cycles can be run iteratively and the process is usually continued until the genome structure ceases to evolve, as measured by the evolution of the parameters of the model. The core functions of the program use Python libraries as well as the CUDA programming language, and therefore necessitate a NVIDIA graphics card with at least 1 Gb of memory.

Scaffolding of the *Ectocarpus* sp. chromosomes with instaGRAAL

To test and validate the instaGRAAL program, we generated an improved assembly of the genome of the model brown alga *Ectocarpus* sp.. A reference genome consisting of 1,561 scaffolds generated from Sanger sequence data is available for this species [20]. A Hi-C library was generated from a clonal culture of a haploid partheno-sporophyte carrying the male sex chromosome using a GC-neutral restriction enzyme (*Dpn*II). The library was paired-end sequenced (2x75 bp – the first ten bases were used as a tag and to remove PCR duplicates) on a NextSeq apparatus (Illumina). Of the resulting 80,521,968 paired-end reads, 41,288,678 read pairs were both concordantly and unambiguously mapped onto the reference genome using bowtie2 (quality scores below 30 were discarded), resulting in 2,554,639 contacts bridging 1,806,386 restriction fragments (Fig 1a) (see Methods for details on the experimental and computational steps). The resulting contact map in sparse matrix format was then used to initialize instaGRAAL along with the restriction fragments (RFs) of the reference genome (Fig 1a-b) (see Table S1 for an example of sparse file matrix).

Figure 1

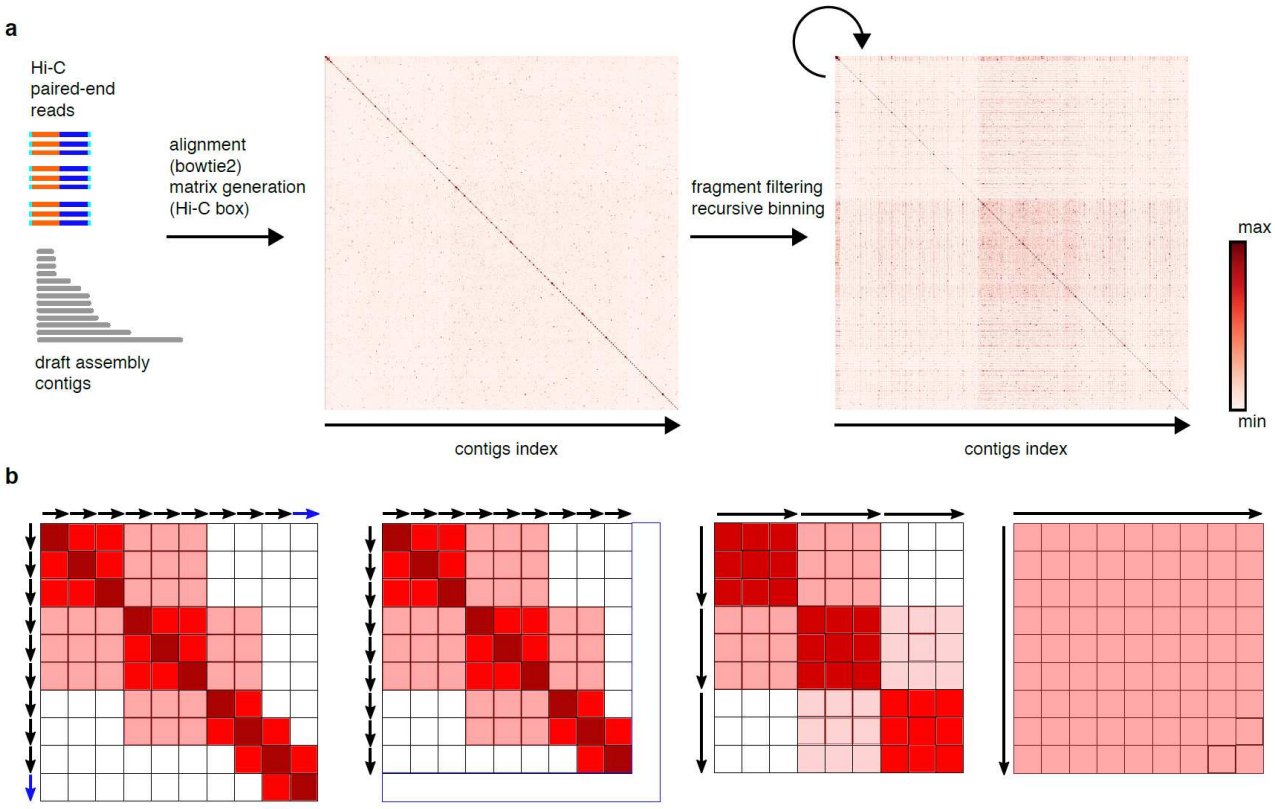


Fig. 1 : Matrix generation and binning process.

(A) (from left to right): i) The input data to be processed, paired-end reads to be mapped onto the *Ectocarpus*. sp. draft assembly; ii) the raw contact map before binning, where each pixel is a contact count between two restriction fragments (RF); iii) the raw contact map after binning, here each pixel is a contact between a determined numbers of RFs (see B). (B) schematic description of one iteration of the binning process (from left to right): i) initialisation of the contact map, where each pixel is a contact count between two RFs; ii) filtering according to coverage, discarding RFs less covered than one standard deviation below the mean and RFs that are too short; iii) sum-pooling along all pixels in a 3x3 square, grouping all RFs by three.

Convergence of the assembly towards 27 major scaffolds

In order to evaluate the program's consistency, given the probabilistic nature of the algorithm, we ran it three times with different resolutions. Briefly, we filtered out RFs shorter than 50 bp and/or whose coverage was one standard deviation below the mean coverage.

Then, we sum-pooled (or binned) the sparse matrix by groups (or bins) of three RFs five times, recursively (Fig 1a-b). Each recursive instance of the sum-pooling is subsequently referred to as a level of the contact map. A level determines the resolution at which permutations are being tested: the higher the level, the lower the resolution, the longer the sequences being permuted and, consequently, the faster the computation. The binning process is shown in Fig. 1b. Regarding *Ectocarpus* sp., we found that level four (bins of 81 RFs) was an acceptable balance between high resolution and fast computation on a desktop computer with a GeForce GTX TITAN Z graphics card. Moreover, whether instaGRAAL was run at level four, five or six (equivalent to bins of 81, 243 and 729 RFs respectively), all assemblies quickly (~6hrs) converged towards similar genome structures (Fig. 2a).

Figure 2

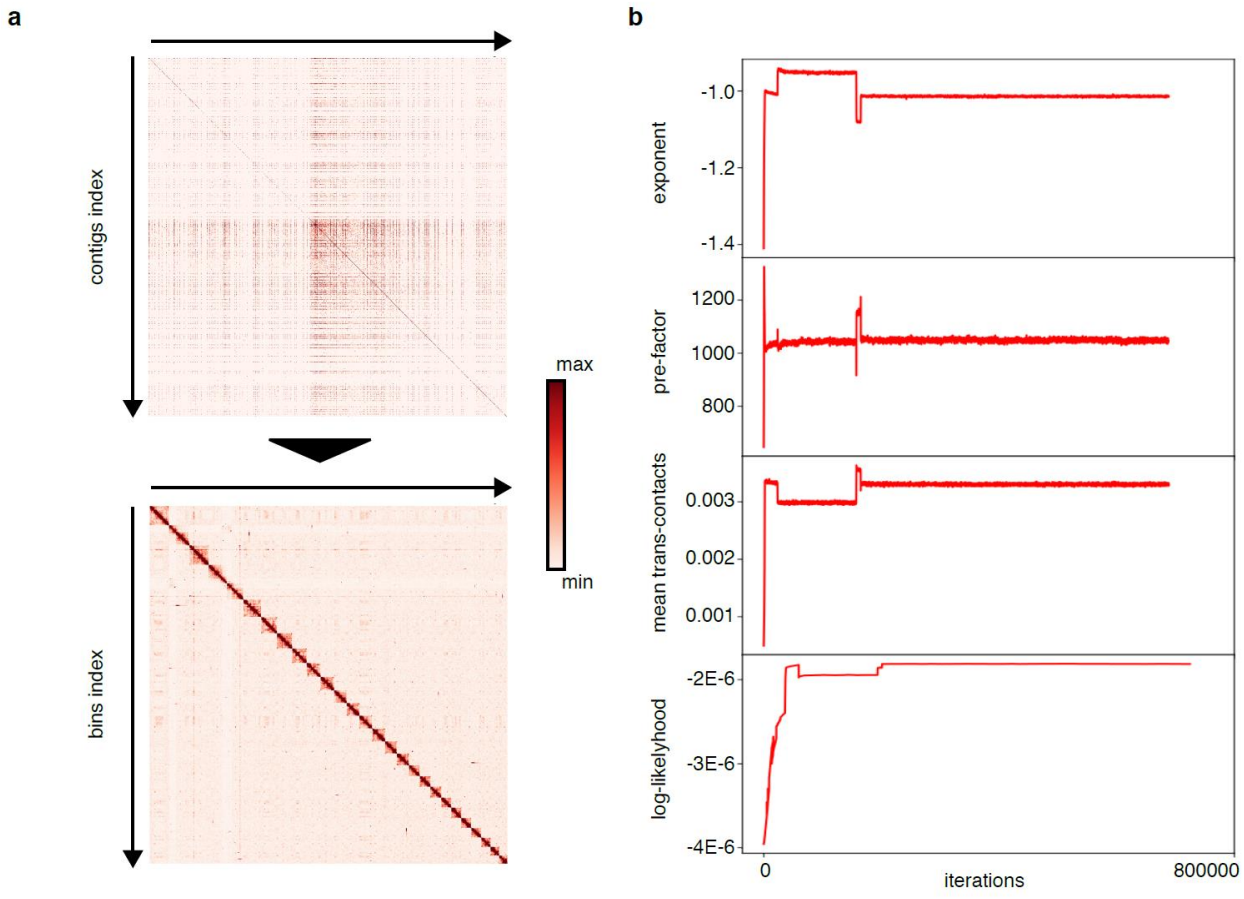


Fig. 2 : Evolution of the contact map, the parameters of the polymer model and the log-likelihood of the contact map.

(A) The raw contact map before (upper part) and after (bottom part) scaffolding using instaGRAAL. (B) The evolution of three parameters of the polymer model (exponent, pre-factor, mean trans-contacts) and the log-likelihood through the iterations.

In order to assess whether an assembly converged, we plotted the evolution of the log-likelihood, as well as the model parameters such as mean trans-scaffold contacts and the exponent of the power law used in the model that ceased to evolve (Fig. 2b). The interquartile ranges (IQR, used to indicate stability in Marie-Nelly et al., 2014) of all parameters decreased to near-zero values at the end of each scaffolding run, again indicating that the convergence was stable and that the final structures oscillated near the final values in negligible ways. More qualitatively, each run led to the formation of 27 main scaffolds (Fig. 2a) with the 27th largest scaffold being more than a hundred times longer than the 28th largest one (Fig. 3) (movie S1). Each of the 27 scaffolds was between four and ten times longer than the combined length of the remaining sequences (Fig. 3). This strongly suggests that these 27 scaffolds correspond to chromosomes, which is consistent with previous estimations based on karyotype analyses [26]. Taken together, these results indicated that instaGRAAL had successfully assembled the *Ectocarpus* genome into chromosome-level scaffolds. Notably, as the supplementary movie suggests, scaffold-level convergence is visible after a few cycles only, indicating that instaGRAAL is able to very quickly determine the global genome structure most likely to fit the contact map. The remainder of the cycles is devoted to intra-chromosome refining.

Figure 3

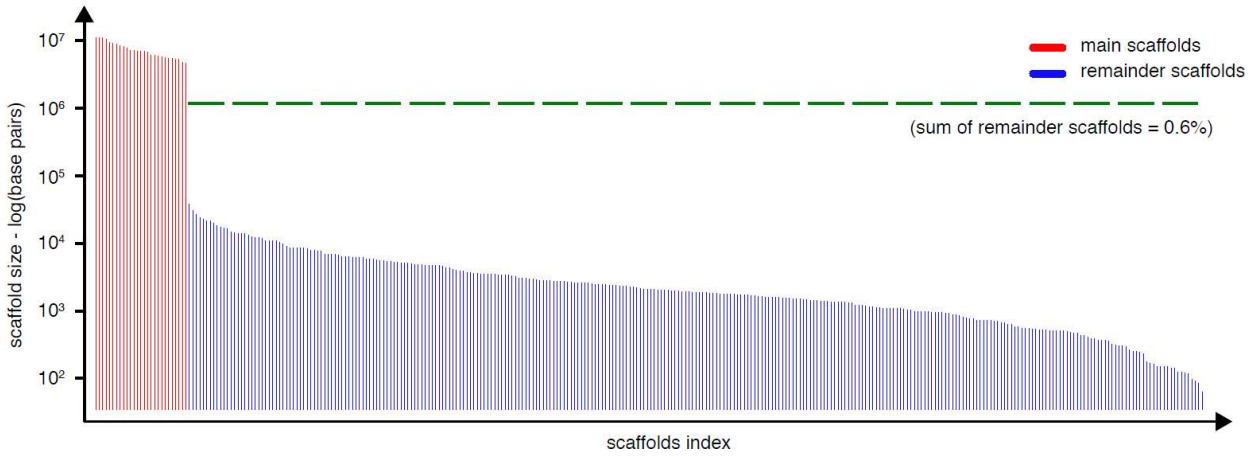


Fig. 3: Size distribution (log scale) of the final scaffolds after 250 instaGRAAL iterations.

After filtering, and prior polishing, 27 main scaffolds (red bars) or putative chromosomes were obtained. The dotted green horizontal line represents the proportion of the filtered genome that was not integrated into the main 27 scaffolds and represent less than 0.6% of the initial assembly. Each scaffold presents, after normalization, a high quality Hi-C profile with features that are typical of eukaryotic genomes (Figure S2c).

Polishing the chromosome-level assembly

As stated above, the instaGRAAL improvements include a number of procedures that aim at correcting modifications of the reference assembly contigs introduced during the Hi-C based scaffolding (Fig. 4).

Figure 4

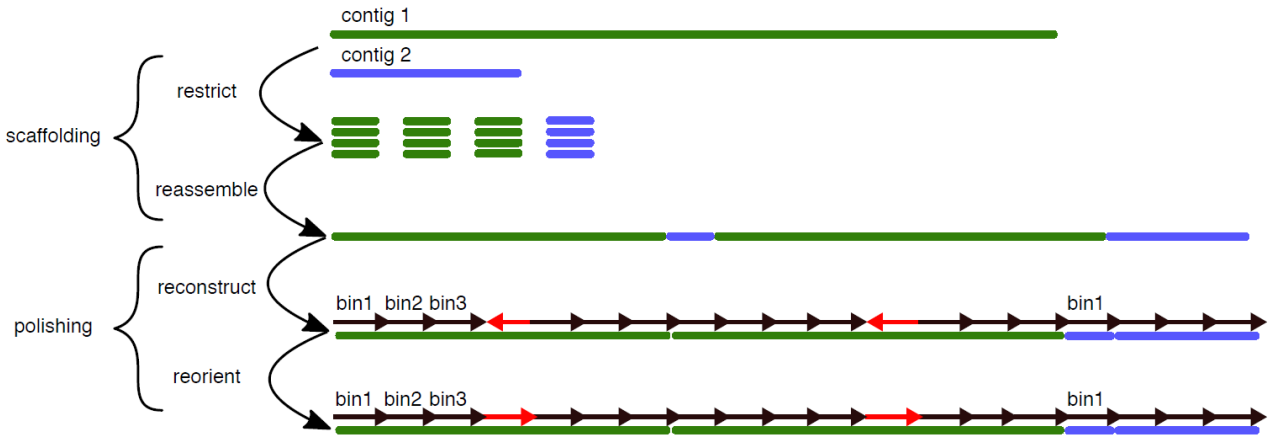


Fig. 4 : Step-by-step correction procedure.

How the polishing procedure keeps track of contig bins (from top to bottom) : i) *in silico* digestion by the restriction enzyme and binning, yielding a set of bins; ii) reassembly of all fragments without reference to their contig of origin; typically, groups of bins from the same contig naturally aggregate, but some bins get scattered among other scaffolds; iii) reconstruction of the original contigs by relocating scattered bins next to the biggest bin group; iv) reorientation of bins within original contigs according to the concensus orientation.

These modifications principally involve discrete inversions or insertions of DNA segments (typically corresponding to single bins or fragments) (see also Marie-Nelly et al. 2014). Such alterations are inherent to the statistical nature of instaGRAAL, which will occasionally improperly permute neighbouring bins because of the high density of contacts between them. These are part of a broader set of assembly errors (subsequently referred to as ‘misassemblies’) that we detected by mapping the reference contigs, generated by instaGRAAL and analysing the mapping results using QUAST. We corrected misassemblies detected in this manner as follows: First, all bins processed by instaGRAAL that belonged to the same contig were constrained to their original orientation in that contig (Fig 4). If a contig was split across multiple scaffolds, the smaller parts of this contig were relocated to the largest one, respecting the original order and orientation of the bins. Then, we reinserted

all the sequences that had been removed by filtering prior to running instaGRAAL (e.g. contig extremities with poor read coverage; see Methods and Marie-Nelly et al., 2014a) into the chromosome level scaffold at their original position within the contig of origin when such a position could be found.

A total of 3,832,980 bp were reinserted into the assembly in this way. These simple steps alleviated the artificial contig truncation problem observed with the original GRAAL program. Some sequences had been filtered out but had no reliable neighbour that they could be associated with, because the entire initial scaffold they belonged to was filtered prior to assembling. These sequences were thus left as-is and appended at the end of the genome. These sequences represent 543 scaffolds spanning 3,141,370 bp, which is less than 2% of the genome. Together, these steps removed all the misassemblies that had been detected by QUAST (Table 1).

Assembly	**Pseudochromosomes (bp)**	**GRAAL**	**instaGRAAL**
N50	6,528,661	6,867,074	6,813,345
NG50	6,528,661	6,725,743	6,813,345
N75	5,613,161	5,693,784	5,686,617
NG75	5,613,161	5,672,622	5,686,617
L50	12	11	11
LG50	12	12	11
L75	19	18	19
LG75	19	19	19
# genomic features	350,497 + 7,261 part	342,253 + 9,766 part	350,555 + 7,261 part
Complete BUSCO (%)	76.9	76.24	77.56
K-mer-based compl. (%)	99.97	98.53	100

Table 1. Comparison of Nx, NGx (Nx with respect to the reference) stats and other genomics

statistics for the different assemblies (Pseudochromosomes, GRAAL and instaGRAAL).

In an effort to further improve and validate the assembly, we exploited genetic linkage data generated for a high density linkage map study to search for potential translocations between the extremities of the scaffolds [24]. This optional analysis, now implemented in instaGRAAL, detected many such events in the unpolished version, but none in the polished assembly. The polished instaGRAAL assembly is therefore fully consistent with the genetic recombination data, confirming the efficiency of the procedure.

Comparisons with previous *Ectocarpus* genome assemblies and validation of the instaGRAAL assembly

To further validate the polished instaGRAAL assembly (subsequently referred to as the polished assembly), a comparison was carried out with three earlier *Ectocarpus* genome assemblies (Table1 and Table S2): 1) the reference assembly, mentioned above, which is highly-accurate but highly fragmented (1,561 scaffolds) generated using Sanger sequencing data [20]; 2) an assembly generated by combining genetic recombination data and the Sanger assembly [19,24] (subsequently referred to as the pseudochromosome assembly) and 3) an assembly generated by running the original GRAAL program on the reference genome data (subsequently referred to as the GRAAL assembly).

We aligned each assembly to the reference assembly to detect misassemblies and determine whether the genome annotations (362,919 features) were conserved. We then validated each assembly using genetic linkage data (see Methods). For each assembly, we assessed a variety of metrics, the most important being the number of misassemblies, the fraction of conserved annotations, ortholog completeness and cumulative length/Nx

distributions (Table 1). These assessments were carried out using BUSCO [27] for ortholog completeness (Figure S1) and QUAST-LG's validation pipeline [28] for the other tests. QUAST-LG is an updated version of the traditional QUAST pipeline specifically designed for large genomes. We follow the terminology used by both programs, including for example the BUSCO definition of ortholog and completeness, as well as QUAST's classification system of contig misassemblies, which correspond to strong discrepancies that necessitate a correction, and scaffold misassemblies, which correspond to breakpoints between contigs that can be presumably ignored since we explicitly want to correct them. The differences between metric values for the GRAAL assembly, generated without any manual correction, and the polished instaGRAAL assembly provided an estimate of the correctness gap and of the improvements made.

The polished instaGRAAL assembly was of better quality than both the pseudochromosome assembly and the GRAAL assembly (Table 1 and Figure S1). The polished assembly incorporated 795 of the reference genome scaffolds (96.8% of the sequence data) into the 27 chromosomes compared to 531 of the reference genome scaffolds (90.5% of the sequence data) for the pseudochromosomal assembly based on the high density genetic map [19]. Moreover, this assembly contained fewer misassemblies, retained more annotations and was more complete (both in terms of k-mers and BUSCO ortholog content). For some metrics the differences were marginal, but always in favour of the instaGRAAL assembly. BUSCO completeness was similar (76.2%, 76.9%, 77.6% for the GRAAL, pseudochromosomal and instaGRAAL assemblies, respectively, Figure S1), and an improvement over the 75.9% of the reference, but the absolute numbers were quite low, although this could be due to the lack of a set of orthologs that are well adapted to brown algae. All values for quantitative metrics such as N50, L50 and cumulative length distribution increased dramatically when compared with the reference genome (Table 1). N50 increased

more than tenfold, from 496,777 bp to 6,867,074 bp following the initial scaffolding, and to 6,942,903 bp after the polishing steps. Similarly, 99.4% of DNA sequence of the 1,018 contigs was integrated into the 27 largest scaffolds after instaGRAAL processing. The GRAAL assembly had a large number of misassemblies compared to the reference genome but these were efficiently removed during the subsequent steps, which corrected all 1,334 misassemblies and reinstated 8,302 genomic features, so that 357,754 of the 362,919 annotations (99.96 %) were transferred from the reference genome to the final assembly (Table 1). It should be noted that scaffold misassemblies, i.e. discrepancies between contigs (as opposed to within contigs) from the reference assembly and those in the polished instaGRAAL assembly, may not necessarily represent errors in the latter for the simple reason that Hi-C scaffolding is usually carried out because the reference genome is thought to contain scaffolding assembly errors. Nevertheless, this analysis indicated that many of the rearrangements found in the pseudochromosome assembly were potentially errors, and that both GRAAL and instaGRAAL were more efficient at placing large regions where they belong in the genome, albeit less accurately in the case of GRAAL in the absence of polishing. These statistics underline the importance of the post-scaffolding polishing steps, and the usefulness of a program that automates these steps.

Comparison between the *Ectocarpus* InstaGRAAL and pseudochromosomal assemblies

Compared to the pseudochromosomal assembly, the instaGRAAL assembly lost 23 scaffolds but gained 287 that the genetic map failed to anchor to the chromosomes (Table S2). We observed a very limited number of conflicts between the two assemblies. One major difference is that instaGRAAL was able to link the 4th and 28th pseudochromosomes which were considered to be separate in the genetic map [24]. The lack of detection of this link in

the genetic map was likely due to the limited number of recombination events observed in the 80 lines used. The fusion in the instaGRAAL assembly is consistent with the fact that the 28th pseudochromosome is the smallest of the linkage groups, with only 54 markers over 41.8 cM and covering 3.8 Mbp. Moreover, the 28th pseudochromosome had a very large gap, which might reflect uncertainty in the ordering of the markers. Interestingly, the gap is located at one end of the linkage group, precisely where instaGRAAL now detects a fusion with the 4th pseudochromosome. Furthermore, the fact that there is no mix between the scaffolds of the 4th and 28th pseudochromosome on the merged instaGRAAL chromosome but rather a simple concatenation, suggests that the genetic mapping process was simply unsuccessful in joining those two linkage groups and that instaGRAAL correctly assembled the two pseudochromosomes (see Table S3 for correspondences between psuedochromosomes and instaGRAAL super scaffolds).

InstaGRAAL was more efficient than the genetic map in orientating scaffolds (Table S2). Among the scaffolds that could be oriented in the pseudochromosomal assembly, about half of the ‘plus’ orientated were actually ‘minus’ and vice versa. The overall limited number of markers detected in the scaffolds anchored to the genetic map was likely the reason for this high level of misorientation.

Comparisons with existing methods

To date, only a limited number of Hi-C based scaffolding programs have been made publicly available. To benchmark our algorithm, we ran SALSA2 [29] on the same Ectocarpus reference genome and Hi-C reads. SALSA2 is a recent program with a promising approach that directly integrates Hi-C weights into the assembly graph. The program ran for nine iterations and yielded 1,042 scaffolds, with an N50 of 6,552,506 (L50 = 11). Its BUSCO-completeness was 77.6%, identical to instaGRAAL’s. Overall the metrics are satisfactory but

still outperformed by instaGRAAL after polishing. Interestingly, the genome is more complete than the raw GRAAL output, underlining yet again the importance of polishing. The contact map of the resulting SALSA2 assembly, however, still showed noticeably unfinished scaffolds (Figure S3). This, coupled with a lower N50 value, indicates that instaGRAAL is more successful at merging scaffolds when appropriate.

Discussion

InstaGRAAL is a Hi-C scaffolding program that provides a solution for genome assembly projects involving Hi-C libraries. Below we discuss the improvements we have made to the program, its remaining limitations and the steps to tackle them when using instaGRAAL in a genome assembly project.

Reference-based polishing

Our main improvement relates to post-scaffolding polishing. A small number of assembly artefacts are expected to be generated initially as a consequence of the algorithm's most erratic random walks. These defects mainly correspond to local inversions, or disruptive insertions of small scaffolds within bigger ones. These caveats are more prevalent than other kinds of noise because they result in only minor disruptions in terms of contact data: bin inversions do not markedly change the relative distance of their constituent fragments relative to their neighbours, and small scaffolds typically carry little signal due to their size and therefore have a greater variance in terms of acceptable positions. The prevalence of such assembly artefacts can be estimated by examining the orientation of bins relative to their neighbours. A single fragment that has been placed in the opposite orientation compared to all neighbouring fragments is likely to represent an error. Depending on the degree to which one trusts the initial reference contigs, one may be less willing to tolerate

“partial translocations” created by instaGRAAL, whereby a contig is split across two different scaffolds creating a false breakpoint at a restriction site. The polishing procedures implemented here tackles these issues. Depending on how much one trusts the initial reference genome, one or more of the procedures may be applied. They aim at reconstructing the initial contig structure and orientation while preserving scaffold junctions when applicable.

In addition to reorganizing the position and orientation of fragments within the assembly, fragments that are removed during the initial filtering process are reintegrated into the assembly using positional information (contig sequences) derived from the reference genome. For example, for a fragment corresponding to the end of a sequence contig, a specific polishing step, which we call tail filtering, reintegrates the fragment into the same contig based on the original structure of that contig in the reference assembly. Removal of small fragments corresponding to contig ends is the most common occurrence of fragment filtering because the size of these end fragments depends arbitrarily on the position of the restriction sites within the contig. Another common occurrence is repeated sequences that failed to be mapped in the first place.

We believe that coupling of a probabilistic algorithm and deterministic polishing is what lends credence and robustness to our program; the MCMC method finds a high-likelihood family of genome structures, making few prior assumptions and allowing it to almost always find the correct global scaffolding. The polishing combines this result with prior assumptions made about the initial contig structure and refines the genome within each scaffold. In order to give the user a fine-grained degree of control over the polishing, the implementation itself is split into modules that each make an assumption about the initial contig structure necessary to perform the correction, *e.g.* the ‘reorient’ module assumes that the initial

contigs do not have inversions, the ‘rearrange’ module assumes that there are no relocations within contigs, etc.

Genetic map based polishing

Genetic maps have been the traditionally go-to method for generating pseudo-chromosomes until new technologies came along over the last decade. Although they provide a simple way of ordering contigs, they do not always achieve a one-to-one mapping between pseudomolecules and actual chromosomes. Moreover, the linkage desequilibrium (LD) data can be disrupted if a chromosome has unusual features such as an abnormally large non-recombining regions; large stretches can be thus unresolved. Hi-C maps are thus more suited for multi-scale scaffolding. Nevertheless, the insight genetic maps provide as to the ordering of different loci makes them a very good candidate for integration with Hi-C data. In line with our previous reasoning about the probabilistic nature of our algorithm and its need for *in fine* polishing, we believe that if conflicts arise between LD-based and contact-based orderings, the genetic map should be given precedence. On the other hand, if no such conflicts are found, our Hi-C based scaffolding is all the more strengthened.

Sparse data handling

The implementation of a sparse data storage method in instaGRAAL allows much more intense computation than with GRAAL. Because the majority of map regions are devoid of contacts, instaGRAAL essentially halves the order of magnitude of both algorithm complexity and memory load, i.e. they increase linearly with the size of the genome instead of geometrically. This improvement potentially allows the assembly of Gb-sized genomes in five to six days using a desktop computer (and faster with a larger computational resource).

Filtering

Coverage and GC distributions have been a long-standing limiting factor in Hi-C based scaffolding methods. Raw Hi-C data is not uniform in %GC content and read coverage across the genome and these variations are a problem when interpreting the data to generate Hi-C contact maps. Correction and attenuation procedures were developed some years ago to alleviate these biases ([30–32], but these are not compatible with instaGRAAL's way of estimating the contact distribution (for more on this distribution, see [33]. Moreover, they do not handle the problem of fragments with no coverage, such as repeated sequences. A filtering step is therefore needed to remove short and/or low-coverage Hi-C RFs that are likely to disrupt the distribution estimation. Maintaining these RFs would not improve the accuracy of the scaffolding and poor or no filtering may lead to incorrect deduction of genome structure and chromosome number. Indeed, their small size or coverage results in a low-information vector with few contacts with the rest of the genome, while nevertheless influencing model parameter estimation. The remaining RFs provide a more robust foundation to compute and fit the contact distribution. In practice, we found that most of the RFs that were removed by the filtering were either entire small scaffolds that are very difficult to link to the rest of the genome, repeated sequences, or the extremities of larger scaffolds. The scaffold end fragments arise due to restriction sites sometimes occurring very near the ends of scaffolds. These disruptive RFs represent a negligible fraction of the total genome, as shown with our present example (< 3% of the total genome size). Importantly, scaffold extremities are incorporated back into the assembly as part of the polishing steps, since their origin is known. Small, isolated scaffolds, on the other hand, cannot be reinserted during the polishing steps as there are no neighbouring sequences in the assembly. Such scaffolds fail to be integrated in most assembly projects, and their integration remains an

problem. Additional analyses, including the use of independent types of data such as long reads or linked reads, could be needed to integrate such scaffolds into the genome.

Resolution

The binning procedure will influence the structure of the final assembly as well as its quality. For example, low level binning (for instance one bin = three RFs) will lead to a large, sparse contact map with a low signal-to-noise ratio in which many of the bins have poor read coverage. This is because, on average, such bins have relatively few contacts beyond their immediate neighbours. As a result of the low signal-to-noise ratio, an invalid prior model will be generated and, when referring to this model, the algorithm will fail to scaffold RFs properly, if at all. Indeed, our attempts to assemble at high resolutions (low level binning) failed to converge in a timely manner. Moreover, due to its probabilistic nature, the algorithm will generate a number of false positive structural modifications such as erroneous local inversions or permutations of bins. These errors occur as a result of the multiple operations performed between each and every bin across the genome. The number of operations increases with the resolution, since the total number of bins increases geometrically when lower levels of the matrix are being used. In addition, the larger the number of bins, the more genome structure spatial dimensions are generated to handle all the potential combinations of bins. Exploring this space therefore takes longer, and converging toward a correct genome in a timely manner becomes difficult with reasonable computational resources. An optimal resolution ensures that the genome structure is consistent with the original contig structure, while allowing for flexibility at higher scales. We conjecture that a sufficiently powerful machine operating on an extremely contact-rich matrix would be more successful at any level. However, it is unclear whether such resources are necessary when our present assembly demonstrated that good quality metrics can be obtained after a day’s worth of

calculations on a standard desktop computer. Moreover, as noted previously, convergence was qualitatively evident after a few cycles. This suggests that more computational power yields diminishing returns, and therefore that appropriate polishing is a more efficient approach to correct any remaining misassemblies.

Lingering missassemblies

We should stress that all these assemblies still contain errors. Thanks to validation tools, we know that our final genome is the most gene-complete and has no discrepancy with the initial reference or the genetic map, making it the highest quality available assembly for *Ectocarpus* sp. Nevertheless, it is still imperfect due to the limitation of our reference material and the technologies used. If the reference contigs themselves are faulty, or the genetic map itself proves inaccurate, polishing may be faulty as well, and this will be reflected in the final assembly. However, renouncing any kind of polishing exposes the assembly to the same problems encountered by the original GRAAL software. This is why the polishing library is broken down into several procedures described in the implementation and documentation. Indeed, suggestions of potential missassemblies can be found in the final contact map, with the presence of extraneous signal (so-called ‘speckles’) outside the typical distribution; it is, however, non-trival to estimate how they should be corrected if they are indeed misassemblies. Consequently, we have chosen to remain entirely faithful to the reference contigs, given that our stats still show the best improvements and the speckles are few and far between.

Fragmentation

The fragmentation of the starting assembly used to generate the initial contact map has obviously a substantial effect on the quality of the final scaffolding. Because binning cannot

be performed beyond the resolution of individual contigs, however small they may be, there is a fixed upper limit to the a scale at which a given matrix can be binned. A highly fragmented genome with many small contigs will necessarily generate a high-noise, high-resolution matrix. Attempts to reassemble a genome based on such a matrix will run into the problems discussed above (resolution). This limitation can be alleviated, to some extent, by discarding the smallest scaffolds, assuming the remainder covers enough of the initial genome. The contigs that are removed can then be then reintegrated into the assembly during the polishing steps. This ensures an improved Nx metric while retaining genome completeness. It should be noted, however, that the size of the contigs is only important insofar as they need to contain sufficient restriction sites, and each of the restriction fragments must have sufficient coverage. The choice of enzyme and the frequency of its corresponding site is thus crucial. For instance, with an average of one restriction site every 600 to 1,000 bp for *Dpn*II, contigs as short as 10 kb may contain enough information to be correctly reassembled. The restriction map therefore strongly influences both the minimum limit on N50 and genome fragmentation.

Integrating information from the Hi-C analysis with other types of data

Aggregating data from multiple sources to construct a high-quality assembly remains a challenging problem with no systematic solution. As long read technologies become more widespread, there is increasing demand to reconcile the scaffolding capabilities of 3C-based methods with the ability of long reads to span regions that are difficult to assemble, such as repeated sequences. The most intuitive approach would be to perform Hi-C scaffolding on an assembly derived from high-coverage and corrected long reads, as was done for several previous assembly projects [14,34]. Alternative approaches also exist, such as generating Hi-C and long-read-based assemblies separately and merging them

using programs such as CAMSA (Aganezov et al., 2017) or Metassembler (Wences et al., 2015). Lastly, pipelines such as PBJelly (English et al., 2012) have proven successful at filling existing gaps in draft genomes, regardless of origin, with the help of long reads. Our scaffolder shows that high quality metrics can still be attained without the help of long reads, but it can nevertheless integrate them when necessary or available.

Long reads are not the only type of data that may be used to improve assemblies. Linkage maps, RNA-seq, optical mapping and 10X technology all provide independent data sources that can help improve genome structure and polish specific regions. The success of future assembly projects will hinge on the ability to process these various types of data in a seamless and efficient manner.

Methods

Preparation of the Hi-C libraries

The Hi-C library construction protocol was adapted from [7,35]. Briefly, partheno-sporophyte material was chemically cross-linked for one hour at RT using formaldehyde (final concentration: 3% in 1X PBS; final volume: 30 ml). The formaldehyde was then quenched for 20 min at RT by adding 10 ml of 2.5 M glycine. The cells were recovered by centrifugation and stored at -80°C until use. The Hi-C library was then prepared as follow. Cells were resuspended in 1.2 mL of 1X *Dpn*II buffer (NEB), transferred to a VK05 tubes (Precellys) and disrupted using the Precellys apparatus and the following program ([20 sec – 6000 rpm, 30 sec – pause] 9x cycles). The lysate was recovered (around 1.2 mL) and transferred to two 1.5 mL tubes. SDS was added to a final concentration of 0.3% and the 2 reactions were incubated at 65°C for 20 minutes followed by an incubation of 30 minutes at 37°C. A volume of 50 µL of 20% triton-X100 was added to each tube and incubation was continued for 30 minutes. *Dpn*II restriction enzyme (150 units) was added to each tube and the reactions

were incubated overnight at 37°C. Next morning, reactions were centrifuged at 16,000 x g for 20 minutes. The supernantants were discarded and the pellets were resuspended in 200 µL of NE2 1X buffer and pooled (final volume = 400 µL). DNA extremities were labelled with biotin using the following mix (50 µL NE2 10X buffer, 37.5 µL 0.4 mM dCTP-14-biotin, 4.5 µL 10mM dATP-dGTP-dTTP mix, 10 µL Klenow 5 U/µL) and an incubation of 45 minutes at 37°C. The labelling reaction was then split in two for the ligation reaction (ligation buffer – 1.6 mL, ATP 100 mM – 160 µL, BSA 10 mg/mL – 160 µL, ligase 5 U/µL – 50 µL, H2O – 13.8 mL). The ligation reactions were incubated for 4 hours at 16°C. After addition of 200 µL of 10%, SDS 200 µL of 500 mM EDTA and 200 µL of proteinase K 20 mg/mL, the tubes were incubated overnight at 65°C. DNA was then extracted, purified and processed for sequencing as previously described (Lazar-Stefanita et al., 2017). Hi-C libraries were sequenced on a NextSeq 550 apparatus (2 × 75 bp, paired-end Illumina NextSeq with the first ten bases acting as barcodes; Marbouty et al., 2014).

Contact map generation

Contact maps were generated from reads using the hicstuff pipeline for processing generic 3C data, available at https://github.com/koszullab/hicstuff. The backend uses the bowtie2 (version 2.2.5) aligner run in paired-end mode (with the following options: --maxins 5 –very-sensitive-local). Alignments with mapping quality lower than 30 were discarded. The output was in the form of a sparse matrix where each fragment of every chromosome was given an unique identifier and every pair of fragments was given a contact count if it was nonzero. Fragments were then filtered based on their size and total coverage. First, fragments shorter than fifty base pairs were discarded. Then, fragments whose coverage was less than one standard deviation below the mean of the global coverage distribution were removed from the initial contact map. A total of 6,974,350 bp of sequence was removed this way. An initial

contact distribution based on a simplified a polymer model [25] with three parameters was first computed for this matrix. Finally, the instaGRAAL algorithm was run using the resulting matrix and distribution.

For the *Ectocarpus* sp. genome, instaGRAAL was run at level 4 (n = 81 RFs), 5 (n = 243 RFs) and 6 (n = 729 RFs). Levels 5 and 6 were only used to check for genome stability and consistency in the final chromosome count. Level 4 was used for all subsequent analyses. All runs were performed for 250 cycles. The starting fragments for the analysis were the reference genome entirely fragmented into restriction fragments. The MCMC was run with 3 burn-in cycles.

Polishing of genome assemblies

The assembled genome generated by instaGRAAL was polished to remove misassemblies using a number of simple procedures that aimed to reinstate the local structure of the initial contigs where possible. Briefly, bins belonging to the same initial contig were juxtaposed in the same relative positions as in the starting assembly contig. Small groups of bins were preferentially moved to the location of larger groups when several such groups were present in the assembly. The orientations of sets of bins that had been regrouped in this manner were modified so that orientation was consistent and matched that of the majority of the group, re-orientating minority bins when necessary. Both steps are illustrated in Fig. 4. Finally, fragments that had been removed during the filtering steps were reincorporated if they had been adjacent to an already integrated bin in the initial assembly. The remaining sequences that could not be reintegrated this way were appended as non-integrated scaffolds.

Validation metrics

Initial and final assembly metrics (Nx, GC distribution) were obtained using QUAST-LG [28]. Misassemblies were quantified using QUAST-LG with the minimap2 aligner in the back-end. Ortholog completeness was computed with BUSCO (v3) [27]. Assembly completeness was also assessed with BUSCO. The evolution of genome metrics between cycles was obtained using instaGRAAL's own implementation.

Validation with the genetic map

The validation procedure with respect to linkage data was implemented as part of instaGRAAL. Briefly, the script considers a set of pseudochromosomes where regions are separated by SNP markers, and a set of Hi-C scaffolds where regions are bins separated by restriction sites. It then finds best-matching pairs of pseudochromosomes/scaffolds by counting how many of these regions overlap from one set to the other. Then, for each pair, the bins in the Hi-C scaffold are rearranged so that their order is consistent with that of the corresponding pseudochromosome. Such rearrangements are parsimonious and try to alter as little as possible. Since there isn't a one-to-one mapping from restriction sites to SNP markers, some regions in the Hi-C scaffolds are not present in the pseudochromosomes, in which case they are left unchanged. When the Hi-C scaffolds are altered this way, as was found in the case of the raw GRAAL assembly, the script acts as a correction. When the scaffolds are unchanged, as was the case with the instaGRAAL assembly, the script acts as a validation.

Software tool requirements

The instaGRAAL software is written in Python 3 and uses CUDA for the computationally intensive parts. It requires a working installation of CUDA with the pycuda library. CUDA is a proprietary parallel computing framework developed by NVIDIA, and as such requires an

NVIDIA graphics card. The scaffolder also requires a number of common scientific Python libraries specified in its documentation.

List of abbreviations

RF: Restriction fragment

MCMC: Markov Chain Monte Carlo

LD: Linkage desiquilibrium

IQR: Inter-quartile range

3C: chromosome conformation capture

GRAAL: genome (re)assembly assessing likelihood from 3D

Availability of data and materials

The datasets generated and analysed during the current study are available in the SRA repository, SRR8550777.

The instaGRAAL software and its documentation are freely available at https://github.com/koszullab/instaGRAAL.

Assemblies, contact maps and relevant materials for the reproduction of the main results and figures are available at https://github.com/koszullab/ectocarpus_scripts.

Competing Interests

InstaGRAAL is owned by the Institut Pasteur. The entire program and its source code are freely available under a free software license.

Funding

This research was supported by funding to R.K. and S.M.C. from the European Research

Council under the Horizon 2020 Program (ERC grant agreements 260822 and 638240, respectively).

Author's contributions

LB rewrote and updated the GRAAL program originally designed by HMN, CZ, and RK. MM and AC performed experiments. LB performed the scaffolding and analyzed the chromosome-level assembly, with contributions from AC, KA, LS, RK, JMC and SMC. LM, MM and RK wrote the manuscript, with contributions from JMC and SMC. LB, MM, SMC and RK conceived the study.

Acknowledgements

We thank our colleagues from the team, as well as Hugo Darras, Heather Marlow, Francois Spitz, Yann Loe Mie, Jitendra Narayan, Jean-François Flot, Jérémy Gauthier, Jean-Michel Drezen, and numerous Github users and contributors for valuable feedback and comments.

References

1. Khan AR, Pervez MT, Babar ME, Naveed N, Shoaib M. A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. Evol Bioinforma Online [Internet]. 2018;14. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5826002/

2. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19:329.

3. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363–76.

4. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 2012;22:557–67.

5. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience [Internet]. 2013 [cited 2018 Nov 2];2. Available from: https://academic.oup.com/gigascience/article/2/1/2047-217X-2-10/2656129

6. Alhakami H, Mirebrahim H, Lonardi S. A comparative evaluation of genome assembly reconciliation tools. Genome Biol. 2017;18:93.

7. Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326:289–93.

8. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing Chromosome Conformation. Science. 2002;295:1306–11.

9. Flot J-F, Marie-Nelly H, Koszul R. Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. FEBS Lett. 2015;

10. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31:1119–25.

11. Kaplan N, Dekker J. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat Biotechnol. 2013;31:1143–7.

12. Marie-Nelly H, Marbouty M, Cournac A, Flot J-F, Liti G, Parodi DP, et al. High-quality genome (re)assembly using chromosomal contact data. Nat Commun. 2014;5:5695.

13. Marbouty M, Cournac A, Flot J-F, Marie-Nelly H, Mozziconacci J, Koszul R. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. eLife. 2014;3:e03318.

14. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50.

15. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.

16. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016;26:342–50.

17. 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. Sci Adv. 2017;3:e1602105.

18. Jourdier E, Baudry L, Poggi-Parodi D, Vicq Y, Koszul R, Margeot A, et al. Proximity ligation scaffolding and comparison of two Trichoderma reesei strains genomes. Biotechnol Biofuels. 2017;10:151.

19. Cormier A, Avia K, Sterck L, Derrien T, Wucher V, Andres G, et al. Re-annotation, improved large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the model brown alga Ectocarpus. New Phytol. 2017;214:219–32.

20. Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, et al. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature. 2010;465:617–21.

21. Coelho SM, Godfroy O, Arun A, Corguillé GL, Peters AF, Cock JM. OUROBOROS is a master regulator of the gametophyte to sporophyte life cycle transition in the brown alga Ectocarpus. Proc Natl Acad Sci. 2011;108:11518–23.

22. Ahmed S, Cock JM, Pessia E, Luthringer R, Cormier A, Robuchon M, et al. A Haploid System of Sex Determination in the Brown Alga Ectocarpus sp. Curr Biol. 2014;24:1945–57.

23. Arun A, Coelho SM, Peters AF, Bourdareau S, Pérès L, Scornet D, et al. Convergent recruitment of TALE homeodomain life cycle regulators to direct sporophyte development in land plants and brown algae. McCormick S, Hardtke CS, editors. eLife. 2019;8:e43101.

24. Avia K, Coelho SM, Montecinos GJ, Cormier A, Lerck F, Mauger S, et al. High-density genetic map and identification of QTLs for responses to temperature and salinity stresses in the model brown alga *Ectocarpus*. Sci Rep. 2017;7:43241.

25. Rippe K. Making contacts on a nucleic acid polymer. Trends Biochem Sci. 2001;26:733–40.

26. Müller DG. Untersuchungen zur Entwicklungsgeschichte der Braunalge Ectocarpus siliculosus Aus Neapel. Planta. 1966;68:57–68.

27. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

28. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34:i142–50.

29. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. bioRxiv. 2019;261149.

30. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. BMC Genomics. 2012;13:436.

31. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9:999–1003.

32. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43:1059–65.

33. Muller H, Scolari VF, Agier N, Piazza A, Thierry A, Mercy G, et al. Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for Hi-C. Mol Syst Biol. 2018;14:e8293.

34. Consortium (IWGSC) TIWGS, Investigators IR principal, Appels R, Eversole K, Feuillet C, Keller B, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018;361:eaar7191.

35. Lazar-Stefanita L, Scolari VF, Mercy G, Muller H, Guérin TM, Thierry A, et al. Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. EMBO J. 2017;e201797342.

Supplemental Information

Chromosome-level quality scaffolding of the *Ectocarpus* sp. *g*enome with instaGRAAL, a proximity ligation-based scaffolder

Lyam Baudry[1,2], Martial Marbouty[1], Hervé Marie-Nelly[1,2], Alexandre Cormier, Komlan Avia, Lieven Sterck, J. Mark Cock[3], Christophe Zimmer, Susana M. Coelho[3,&], Romain Koszul [1,&]

Table of contents

id_frag_a	id_frag_b	n_contact
0	0	1368
0	1	21
0	2	7
0	3	3
0	4	5
0	7	5
0	8	1
0	9	1
0	12	2
0	15	1
0	22	1
0	23	1
0	26	1
0	27	1
0	33	2
0	36	2
0	37	1
0	51	1
0	69	1
0	74	2
0	76	1
0	97	1
0	99	1
0	107	1

Table S1: example of a sparse matrix.

	Reference genome	Pseudochromosomal assembly	instaGRAAL assembly
Scaffolds integrated into pseudochromosomes (out of 1561)	325	531	793
Percent sequence data Integrated into pseudochromosomes	70.10 %	90.50 %	96.80 %
Integrated oriented scafflolds in the (pseudo)chromosomes	12 %	49 %	100 %
Number of (pseudo)chromosomes	34	28	27

Table S2: comparison of the integrated sequences between the different assemblies and the reference genome.

instaGRAAL	Pseudochromosomal assembly
1	1
2	21
3	4 and 28
4	5
5	13
6	6
7	12
8	7
9	27
10	26
11	3
12	2
13	8
14	14
15	10
16	11
17	19
18	16
19	9
20	15
21	18
22	20
23	24
24	23
25	17
26	25
27	22

Table S3: correspondences between instaGRAAL super scaffolds and pseudochromosomes.

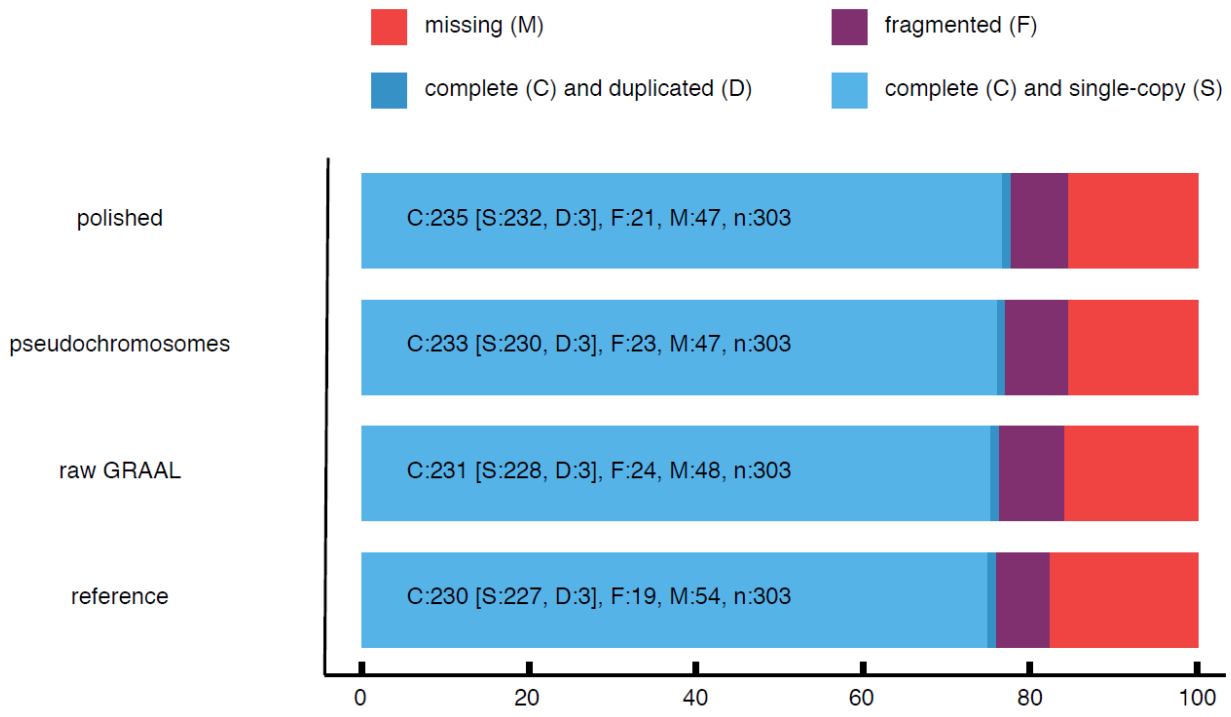


Figure S1: estimates of BUSCO-completeness for the three assemblies and the reference genome.

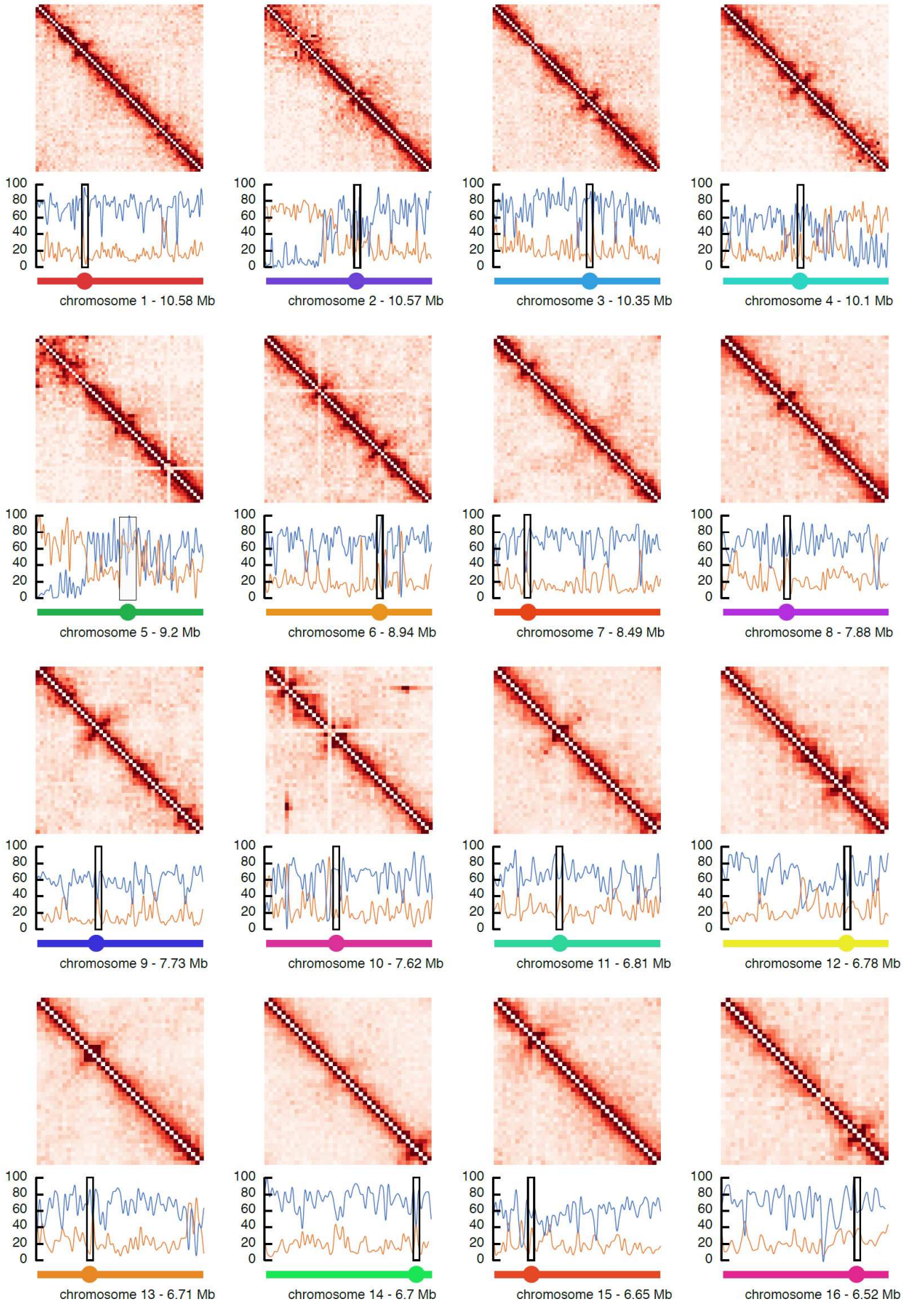

100
80
60
40
20
0
chromosome 1 - 10.58 Mb
chromosome 2 - 10.57 Mb
chromosome 3 - 10.35 Mb
chromosome 4 - 10.1 Mb
chromosome 5 - 9.2 Mb
chromosome 6 - 8.94 Mb
chromosome 7 - 8.49 Mb
chromosome 8 - 7.88 Mb
chromosome 9 - 7.73 Mb
chromosome 10 - 7.62 Mb
chromosome 11 - 6.81 Mb
chromosome 12 - 6.78 Mb
chromosome 13 - 6.71 Mb
chromosome 14 - 6.7 Mb
chromosome 15 - 6.65 Mb
chromosome 16 - 6.52 Mb

Figure S2: Normalized contact map of *Ectocarpus* chromosomes 1 to 16 (left) and 17 to 31 (right).

Contact maps are binned in 200 Kb. The colour scale represents the normalized interaction frequencies as in Fig 5. Percentage of gene sequence (blue) or transposable elements (TE – orange) are indicated under each contact matrix. Putative centromere sequences (rectangle) were called with centroid.

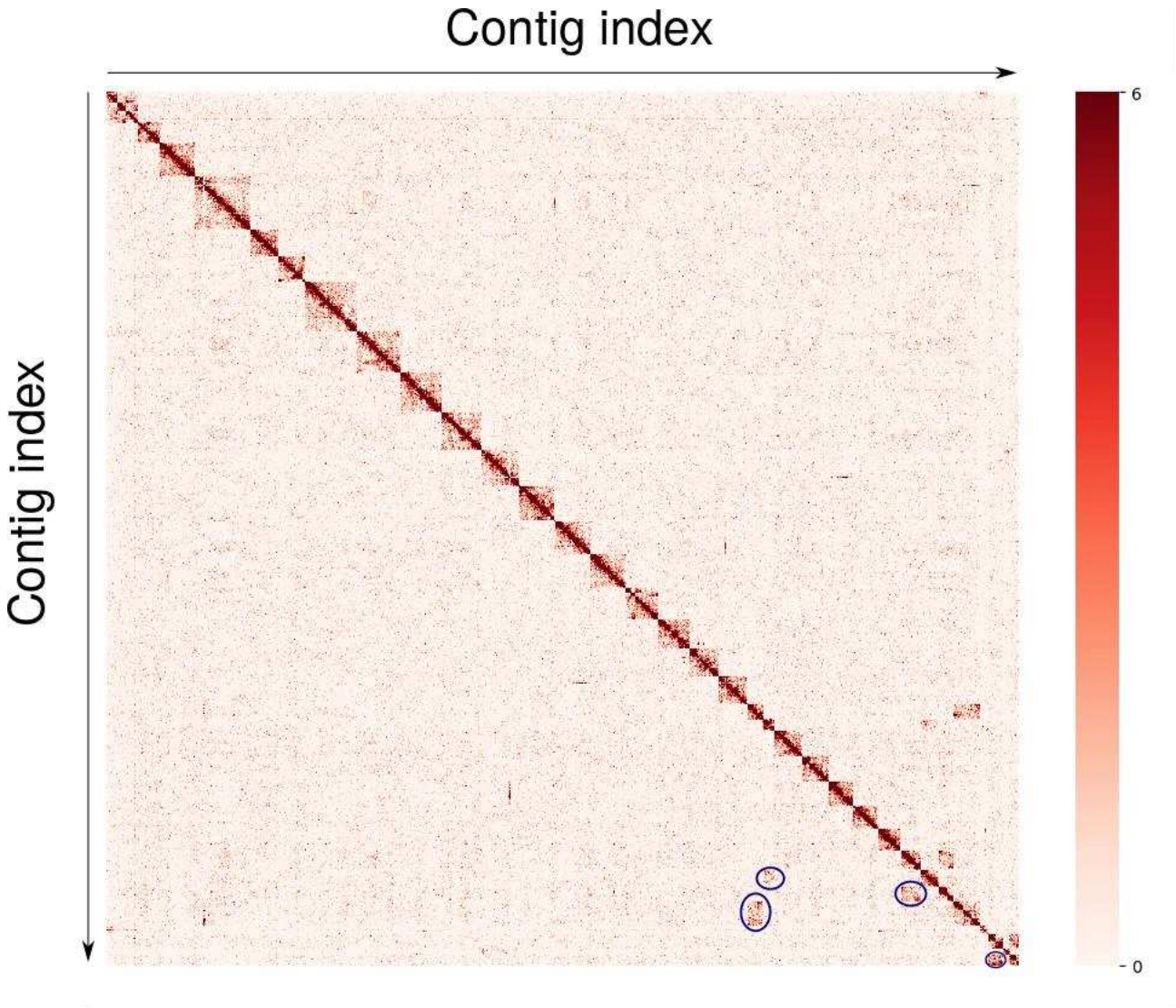


Figure S3: Contact map of the SALSA2 scaffolding. Large signal discrepancies have been marked. Smaller discrepancies are comparable to those obtained with instaGRAAL.

3.2 Assembling and detecting chromosomal rearrangements

After presenting our framework for Hi-C based scaffolding and its direct implementation, we present two use cases for investigating chromosome dynamics as revealed by joint assemblies:

- The first subsection concerns our published work on two lineages of *Trichoderma reesei* (QM6A and RUTC30). We successfully scaffolded both, yielding high quality chromosome-level assemblies. With that information, we identify a large-scale rearrangement.
- The second subsection concerns our main results on two lineages of *Cataglyphis hispanica*. Likewise, we scaffolded both genomes and revealed a chromosomal fusion. Work is still ongoing on the annotation of both assemblies and linking our newly acquired structural information to functional mechanisms underlying the peculiar hybridogenetic reproduction strategy of that species.

3.2.1 Rearrangements between two lineages of *Trichoderma reesei*

Jourdier *et al. Biotechnol Biofuels (2017) 10:151*
DOI 10.1186/s13068-017-0837-6

Biotechnology for Biofuels

RESEARCH

Open Access

Proximity ligation scaffolding and comparison of two *Trichoderma reesei* strains genomes

Etienne Jourdier[1†], Lyam Baudry[2,3†], Dante Poggi-Parodi[1], Yoan Vicq[1], Romain Koszul[2,3], Antoine Margeot[1], Martial Marbouty[2,3*‡] and Frédérique Bidard[1*‡]

Abstract

Background: The presence of low complexity and repeated regions in genomes often results in difficulties to assemble sequencing data into full chromosomes. However, the availability of full genome scaffolds is essential to several investigations, regarding for instance the evolution of entire clades, the analysis of chromosome rearrangements, and is pivotal to sexual crossing studies. In non-conventional but industrially relevant model organisms, such as the ascomycete *Trichoderma reesei*, a complete genome assembly is seldom available.

Results: The chromosome scaffolds of *T. reesei* QM6a and Rut-C30 strains have been generated using a contact genomic/proximity ligation genomic approach. The original reference assembly, encompassing dozens of scaffolds, was reorganized into two sets of seven chromosomes. Chromosomal contact data also allowed to characterize 10–40 kb, gene-free, AT-rich (76%) regions corresponding to the *T. reesei* centromeres. Large chromosomal rearrangements (LCR) in Rut-C30 were then characterized, in agreement with former studies, and the position of LCR breakpoints used to assess the likely chromosome structure of other *T. reesei* strains [QM9414, CBS999.97 (1-1, *re*), and QM9978]. In agreement with published results, we predict that the numerous chromosome rearrangements found in highly mutated industrial strains may limit the efficiency of sexual reproduction for their improvement.

Conclusions: The GRAAL program allowed us to generate the karyotype of the Rut-C30 strain, and from there to predict chromosome structure for most *T. reesei* strains for which sequence is available. This method that exploits proximity ligation sequencing approach is a fast, cheap, and straightforward way to characterize both chromosome structure and centromere sequences and is likely to represent a popular convenient alternative to expensive and work-intensive resequencing projects.

Keywords: *Trichoderma reesei*, Genome assembly, Hi-C, GRAAL, Centromere, Karyotype, Translocation, Chromosomal contact, Chromosome conformation capture

Background

Trichoderma reesei is one of the main industrial enzyme producers [1]. This Ascomycota naturally produces a full set of lignocellulosic biomass degrading enzymes, and carries high stakes for the food, textile, and bioenergy industries. Over the years, the enzyme production has been boosted through cycles of random mutageneses, with highly performing strains secreting up to 100 g L^{-1} of the natural enzyme mix [2]. *T. reesei* is also increasingly used as a versatile heterologous protein producer [3, 4]. In contrast to its industrial interest, the genetic tools available in *T. reesei* have developed at a slower pace than in other model filamentous fungi such as *Neurospora crassa* partly because of the small research community

*Correspondence: martial.marbouty@pasteur.fr; frederique.bidard-michelot@ifpen.fr
†Etienne Jourdier and Lyam Baudry contributed equally to this work
‡Martial Marbouty and Frédérique Bidard contributed equally to this work
[1] IFP Energies nouvelles, 1 et 4 Avenue de Bois-Préau, 92852 Rueil-Malmaison, France
[2] Groupe Régulation Spatiale des Génomes, Department Genomes and Genetics, Institut Pasteur, 75015 Paris, France
Full list of author information is available at the end of the article

sometimes constrained by industrial confidentiality imperatives. In addition, until recently [5], neither sexual crossings nor any annotated karyotype were available for this fungus.

Trichoderma reesei, described from a single wild-type isolate called QM6a, was believed to be devoid of a sexual cycle, whereas its teleomorph, *Hypocrea jecorina*, undergoes an heterothallic sexual cycle involving *MAT1-1* and *MAT1-2* loci [6]. The identification of a *MAT1-2* locus in the QM6a followed by a sexual crossing with a natural isolate of a *MAT1-1* type resulted in fertilized stromata and mature ascospores [5]. QM6a and its derivatives (of which QM9414, NG14, Rut-C30 [7]) are female sterile but male fertile and could nevertheless be crossed with a *MAT1-1* natural isolate acting as female partner, paving the way to the development of sexual crossing tools to generate genetic diversity, genetic cleanup, and strain improvement. Several groups have since built on this original finding by characterizing the receptor/pheromone system [8], uncovering the causes for female sterility [9] and studying meiosis [10] in this species. The latter study have demonstrated the biotechnological interest of crossings different industrial strains but also underlined their limits by pointing at the presence of segmental aneuploidies and chromosome rearrangements resulting in non-viable ascospores.

Chromosomal rearrangements in mutagenized *T. reesei* strains have been first described in the nineties [11, 12]. The karyotypes of industrial strains descending from the parental QM6a strain by several rounds of random mutagenesis displayed massive rearrangements, as revealed by pulse-field gel electrophoresis (PFGE). However, the relatively low resolution of the PFGE technique for chromosomes of similar sizes led to discrepancies between the original studies, and the precise karyotypes of the strains remained elusive. Years later, the draft sequence of the QM6a strain genome was released as a set of 89 scaffolds [13]. Subsequent efforts to obtain genomic wide information of other strains of the same lineage used either genome walking [14], oligonucleotide arrays [15], or short-reads sequencing platform [16–19] but did not improve the assembly. Even though the positions of chromosomal breakpoints were identified for several derivative strains [15], the impact on the chromosomal structure was difficult to assess because of the lack of a complete assembly. In addition, centromeres and telomeres positions remained unknown, as these regions are typically difficult to sequence and assemble because of their low complexity and, for centromeres, the lack of universal conserved sequence patterns. However, reaching at a full genome scaffolds remains an important goal for these model fungi [20]. In the case of *T. reesei*, getting the sequence and exact position of centromeres would provide invaluable information for the emerging sexual crossing field in this species. More broadly, information on centromeres in filamentous fungi remains sparse, and these sequences would bring interesting highlights onto their evolution and metabolism [21].

Using chromosome conformation capture data (3C; or also dubbed proximity ligation data) [22] and the homemade program GRAAL (Genome Re-Assembly Assessing Likelihood from 3D), our groups recently published the first proximity ligation scaffolding of an incomplete eukaryotic genome sequence. The 89 scaffolds of the *T. reesei* QM6a strain were re-scaffolded into seven chromosomes [23, 24]. In addition, the "Rabl" structure of chromosomes in fungi nuclei, where centromeres are clustered together at the microtubule organization center (spindle pole body in yeast), generates contacts enrichment between these sequences. When quantified, we also showed that the signal resulting from these 3D contacts allows the identification of centromere positions [25]. Although the QM6a contact map displayed such signal, we did not at the time characterize precisely these sequences. The published sequence from this past work was not thoroughly integrated within the JGI reference genome database, though it was nevertheless exploited in independent analyses by others [26].

Here, we provide an updated version of the QM6a chromosome scaffolding using an extra polishing step after GRAAL output. GRAAL is a scaffolding pipeline that processes pre-assembled contigs; as a result, the resulting assembly displays the same sequence as in the original genome. We also exploited the 3C contact map to identify the position and sequences of the QM6a centromeres [25], providing insight about *T. reesei* centromeres. The same pipeline was applied to the QM6a-derived strain Rut-C30, resulting in a genome scaffold in perfect agreement with previously identified chromosomal rearrangements between the two genomes [14, 15]. This result prompts us to put forward predictive karyotypes for several other *T. reesei* strains and to discuss the impact of such karyotypes on the emergence of segmental aneuploidy during crossing experiments [10].

Results

Improved QM6a chromosome assembly

The *T. reesei* QM6a genome was scaffolded into superscaffolds using the reference assembly from Martinez et al. [13] and the chromosome contact reads from Marie-Nelly et al. [23]. Scaffolding was performed using the latest version of GRAAL [27] run for 100 iterations. The scaffolding remains nearly identical to the one published previously, with seven superscaffolds matching the seven chromosomes [23]. Again, a fraction (0.5%) of the original assembly was not included in the superscaffolds,

as a result of low 3C sequencing coverage (lack of restriction sites and/or highly divergent GC content could account for such low coverage).

Because the resolution of the GRAAL scaffolding is limited by the distribution of restriction sites along the chromosome and the read coverage, a manual curation was necessary to complete the assembly. This step includes reinserting missing scaffold fragments, checking telomere repeats' orientations, and slightly shifting split locations to remain consistent with the presence of N gaps in the reference genome (see "Methods"). The resulting QM6a GRAAL scaffolding is fully consistent with the JGI reference genome, containing exactly the same sequences than original scaffolds. 65 scaffolds, comprising 99.5% of the genome, were scaffolded along seven chromosomes (Fig. 1). 22 scaffolds, representing 0.5% of the genome, were either too small (not enough restriction sites along their sequences) or insufficiently covered (not enough reads during 3C library sequencing) to be scaffolded within the chromosomes. We did not sequence the gaps between reassembled scaffolds, and instead 100 Ns were intercalated between scaffolds as a marker of GRAAL scaffolding position. Additional sequencing work would therefore be required to reach a final fully continuous genomic sequence. In a simultaneous and independent study from Ting-Fang Wang's team, a QM6a resequencing was performed (Wan-Chen Li et al. personnel communication). We agreed on the chromosome nomenclature (order by decreasing size, numbering with Roman numerals, and orientation with left arm shorter than right arm) so as our works are consistent.

Most scaffolds from the reference genome remained intact in the reassembly (in gray Fig. 1b). However, four scaffolds (1, 2, 5, and 28) were misassembled in the reference genome and were split by GRAAL into several segments in the new scaffolding (Fig. 1b) [23]. The split location of scaffold 28 and its reassembly with scaffolds 27 and 36 is consistent with deep sequencing of the CBS999.97 (1-2, *wt*) strain, whose genome is similar to QM6a [10]. We previously suggested that a fragment of scaffold_9 (≈1020–1045 kb) containing the ribosomal DNA units was duplicated on chromosome VI [23]. However, we were not able to determine the precise number of copies (probably three or four) and the exact sequence to assemble these copies, and we preferred to leave the exact sequence of scaffold_9 as in the JGI reference genome. Therefore, chromosome VI is in fact longer than chromosome VII (Wan-Chen Li et al. personnel communication).

Table 1 shows statistics on chromosome sizes, number of genes, and gene densities. Gene density in *T. reesei* is much more uniform than suggested [26], ranging from 0.26 to 0.28 genes per kb. Additional files 1, 2, 3, contain details on this reassembly (scaffold assembly, final sequence, gene annotation).

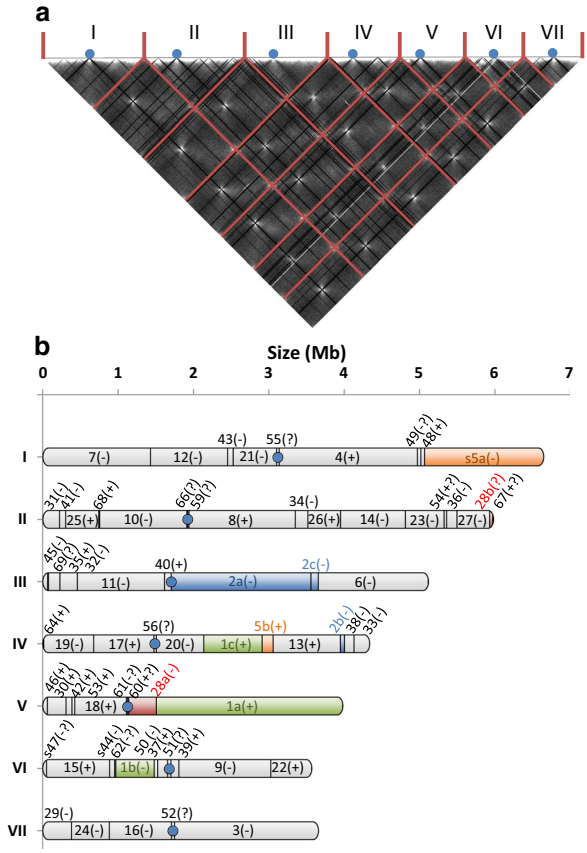


Fig. 1 *T. reesei* scaffold reassembly in seven chromosomes. *T. reesei* scaffolds from the JGI reference genome have been reassembled using chromosomal conformation capture (3C) sequencing data. **a** Contact matrix resulting from GRAAL reassembly. *Red bars* indicate the boundaries of the seven chromosome; centromere positions are represented by *blue dots*. **b** Order and orientation of the reassembled scaffolds in the seven chromosomes. Orientation uncertainties are noted with a question mark. Scaffolds 1, 2, 5, and 28 that were misassembled in the reference genome are shown in *green*, *blue*, *orange*, and *red*, respectively. Centromere positions are represented by *blue dots*

Centromere locations

Fungi chromosome organization typically follows a "Rabl" pattern, with the centromeres colocalizing at the microtubule organizing center. For instance, the strong *trans* contact signal between centromeres of *Saccharomyces cerevisiae* reflects this organization, resulting in discrete dots over the contact map of this species [28]. We have previously shown that centromere–centromere 3D contacts can be used to infer the positions of these regions along the 1D sequence [25]. The bright dots clearly visible in the contact map of the *T. reesei* QM6a

Table 1 Size (bp) , number of genes, and gene density (nb of genes per kb) of *T. reesei* QM6a chromosomes

Genetic element	Size	Number of genes	Gene density
Chromosome I	6,647,935	1817	0.27
Chromosome II	5,980,447	1701	0.28
Chromosome III	5,112,650	1336	0.26
Chromosome IV	4,337,413	1162	0.27
Chromosome V	3,979,336	1092	0.27
Chromosome VI	3,567,305	983	0.28
Chromosome VII	3,660,386	1022	0.28
Unassembled scaffolds	163,868	16	
Total	33,449,340	9129	

Gene annotation was based on the JGI Filtered Models set of genes

genome unveiled a clear Rabl organization (Fig. 1a), pointing at the centromeric regions in this species, and allowing us to identify their positions along the seven chromosomes (Table 2). These centromere signatures pointed at a set of 11 small scaffolds ranging in size from 11 to 43 kb (total length of 270 kb). Three of them (57, 58, and 65) could not be assigned to specific chromosomes (they are part of the 22 unassembled scaffolds), but the eight others were scaffolded within six of the seven chromosomes. For chromosome III, the centromere signature was found at the frontier between scaffolds 2 and 40, but we were not able to identify which centromere scaffold among scaffolds 57, 58, or 65, should be reassembled at this place. The centromeres of chromosome I, VI, and VII are metacentric, whereas the four others (chromosomes II to V) are submetacentric, with the longer (right) arm of the chromosome roughly twice as long as the shorter (left) arm.

AT content in centromeres

The average AT content of these centromere scaffolds is 76%, a much higher value than the average AT genomic content (48% [13]), and consistent with other fungal centromeres [21]. We checked whether this high AT content was specific to centromeres or telomeres by looking for AT-rich regions (%AT >65% and length >4 kb) in the whole genome. In addition to the 270 kb centromere scaffolds, 776 kb AT-rich regions were identified over the genome (98 kb at telomeres; 604 kb split over 72 intra-chromosomal regions; 74 kb in 12 unassembled scaffolds). Most AT-rich regions were positioned at the end of scaffolds, which may explain the previous assembly failures.

Genes in and around centromeres

Seventeen genes were annotated in these 11 scaffolds but all seems to be dubious Coding DNA Sequences (CDS) with many or very large introns, and their products are all annotated as putative proteins of unknown function. Using previously generated RNA-Seq data ([29] and Pirayre et al. to be published), we checked for transcription in these centromere scaffolds and we did not observe any transcription event. So it seems that most probably no gene is present on these scaffolds involved in *T. reesei* centromeres. Function enrichment analysis in close proximity to the centromeres (in a 50-kb window around centromeres) revealed significant enrichments in genes involved in nucleosome assembly (5 genes annotated with the GO term GO:0006334) and in genes linked to the respiratory chain (15 genes in the metabolic pathways of coenzyme Q biosynthesis, adenosine ribonucleotides de novo biosynthesis, and respiration). We can only make assumptions on the significance of this finding, but

Table 2 *T. reesei* QM6a centromeres

Chr	Location on chr (Mb)	Between scaffolds	Scaffolds involved	Size (kb)	%AT	Nb of genes (gene IDs)
Scaffolds with centromere signature reassembled in chromosomes						
chr I	3.12	21(−) and 4(+)	55	34	77.9	4 (112,674, 112,675, 112,676, 112,677)
chr II	1.93	10(−) and 8(+)	66 + 59	30	70.0	3 (71,146, 43,199, 42,942)
chr III	1.71	40(+) and 2a(−)	Unknown			
chr IV	1.48	17(+) and 20(−)	56	32	74.2	2 (112,678, 112,679)
chr V	1.12	18(+) and 28a(−)	60 + 61	32	77.0	1 (112,683)
chr VI	1.67	37(+) and 39(+)	51	43	76.3	2 (112,649, 73,103)
chr VII	1.73	16(−) and 3(−)	52	41	76.7	1 (112,651)
Other scaffolds with centromere signature but not reassembled						
			57	26	76.6	0
			58	21	78.7	3 (112,680, 112,681, 112,682)
			65	13	81.7	1 (112,689)

Chromosomal contact data were used to identify the location of the centromeres on the chromosomes. Centromeres were all identified in small scaffolds, not in the middle of well-assembled scaffolds

it could be that their presence in a zone of pericentric repression of crossover is a sign of their importance for the organism robustness and fitness [30]. Interestingly, the CenH3 (centromere-specific histone H3) encoding gene 57870 (orthologue of *N. crassa* NCU00145 and *S. cerevisiae* CSE4) was found on chromosome I at only 30 kb from the centromere (0.5% of the chromosome length). This feature is not conserved in other species, for example, *Schizosaccharomyces pombe* Cnp1 is found at 1.93 Mb from the centromere [31].

Inverted repeats

Although aware that centromeres were not fully assembled, we checked their sequences for homologies or repeats. We did not observe any sequence homology between centromere regions, which is consistent with the now accepted finding that most centromeres are epigenetically and not genetically maintained [32]. Remarkably, in four cases [scaffolds 51 (chr. VI), 56 (chr. IV), 57 and 58], we observed an inverted repeat structure with a central core region of 1–2 kb surrounded by an inverted repeat of 2.5–5 kb, which is quite similar to the centromere structure of *S. pombe* [31, 33, 34], *Candida albicans* [35], *Candida tropicalis* [36], and *Komagataella phaffii* (formerly *Pichia pastoris*) [37]. Details on this observation are available on Additional file 4 but a complete study on *T. reesei* centromeres structure would require a full assembly, and chromatin immunoprecipitation sequencing experiments.

Rut-C30 chromosome assembly

In order to get a chromosomal map of *T. reesei* Rut-C30, a 3C library of the Rut-C30 strain was generated, sequenced, and the resulting reads exploited to rescaffold the QM6a genome. Although a genomic sequence was available for *T. reesei* Rut-C30 strain [17], the JGI reference sequence of *T. reesei* QM6a strain was used to demonstrate that the approach could be applicable to any other non-sequenced strain, even if significant chromosomal rearrangements are expected.

GRAAL identified three chromosomal translocations and one large deletion (Table 3) present in Rut-C30 compared to the QM6a, in agreement with previous work [14, 15]. By design, and as stated before, GRAAL identifies rearrangement events with a precision limited by the sequencing coverage and the restriction pattern of the region (in this case, a couple of dozens of kb; "Methods"). Besides the rearrangements listed in Table 3, the two genome assemblies of Rut-C30 and QM6a were compared and did not present major differences: the reordering of the scaffolds not involved in chromosomal rearrangements (including the splitting of the misassembled scaffolds 1, 2, 5, and 28), as well as centromere positions, were fully consistent between the two assemblies (the Rut-C30 reassembly is available in Additional file 5). The fully scaffolded genomes of these two strains can then be compared in an attempt to have a better understanding of the evolutionary trajectories of the evolved Rut-C30 genome (Fig. 2). Different scenarios are possible from QM6a to Rut-C30, depending on the order of occurrence of the three translocation events, leading to the same chromosome structure. One possible scenario is shown Fig. 2c.

The three translocations resulted finally in the right arm of chromosome I (3′ end of scaffold 48 and main fragment of scaffold 5: 1.63 Mb and 442 genes in total), to be swapped with the right arm of chromosome I (3′ end of scaffold 22: 402 kb and 114 genes). But also in two fragments of chromosome I (one with a fragment of scaffold 4, and the other one with another fragment of scaffold 4, scaffold 49, and a small fragment of scaffold 48) to be inserted head to foot in the middle of the chromosome V (1.13 Mb and 310 genes in total for both fragments). Therefore, the whole sequence of chromosome III is still found on chromosome III. The 85 kb deletion is closed to the telomeric region of chromosome VI and

Table 3 Translocation and large deletion events found in GRAAL reassembly of *T. reesei* Rut-C30 with respect to QM6a

Translocation	Location on scaffolds (this study)	Location on scaffolds [15]	Mapping on QM6a chromosomes
n° 1	scaffold_2: 556 ± 22 kb	scaffold_2: 546,703 bp	chr III: 3,166,447
	scaffold_4: 1,197 ± 25 kb	scaffold_4: 1,204,862 bp	chr I: 4,342,096
n° 2	scaffold_4: 750 ± 27 kb	scaffold_4: 748,277 bp	chr I: 3,885,511
	scaffold_22: 138 ± 31 kb	scaffold_22: 139,515 bp	chr VI: 3,165,364
n° 3	scaffold_22: 138 ± 31 kb	scaffold_22: 139,476 bp	chr VI: 3,165,325
	scaffold_48: 0 ± 35 kb	scaffold_48: 1667 bp	chr I: 5,018,020
Large deletion	**Location on scaffold (this study)**	**Location on scaffold [14]**	**Mapping on QM6a chromosome**
85-kb deletion	scaffold_15: 0–85 ± 25 kb	scaffold_15: 1,555–86,603	chr VI: 52,198–137,246

Newly acquired 3C-seq data of *T. reesei* Rut-C30 strain were used to reassemble the reference genome. Comparison with QM6a reassembly allowed the identification of three chromosomal translocations and one large deletion. The position of these rearrangements is consistent with former work [14, 15]

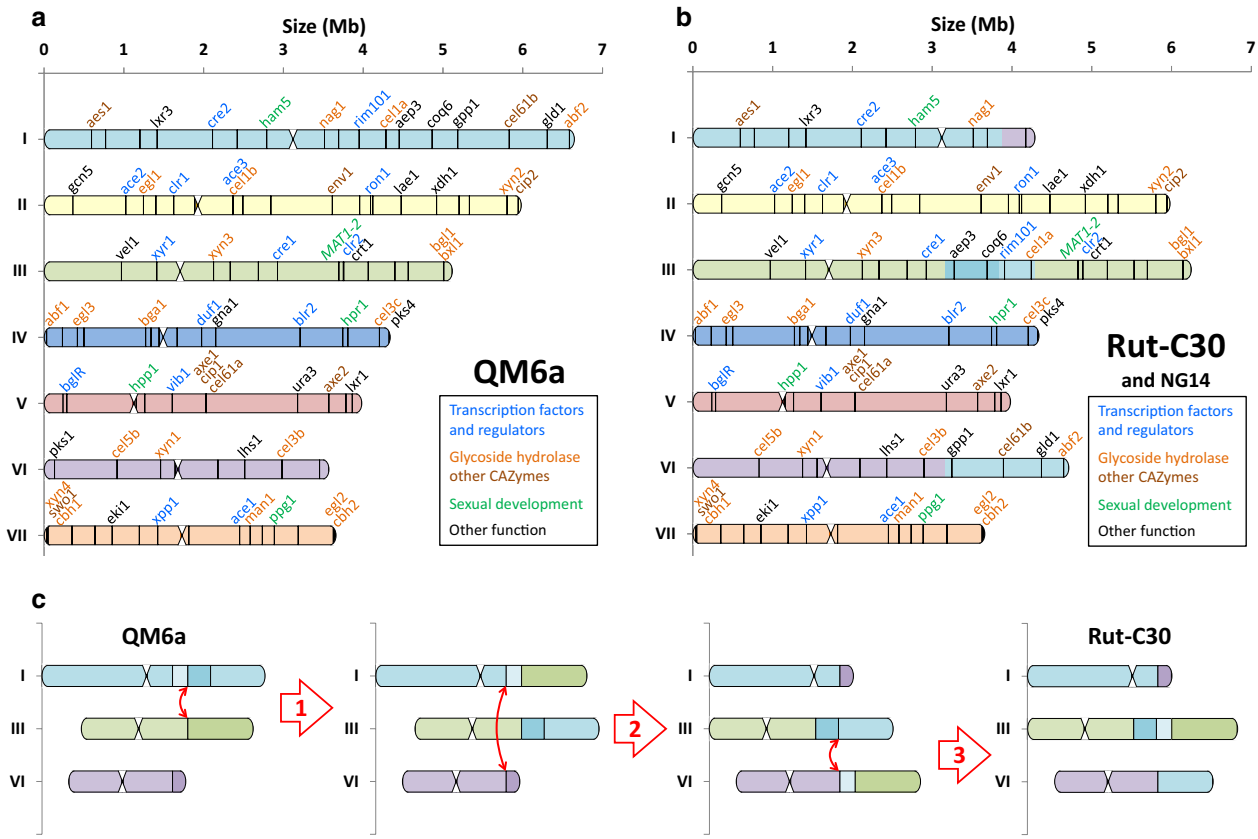


Fig. 2 Chromosome maps of *T. reesei* QM6a and Rut-C30 strains. Chromosome maps of *T. reesei* QM6a (**a**) and Rut-C30 and NG14 (**b**) strains were identified by reassembly of the JGI reference genome using 3C sequencing data for each strain. For Rut-C30 map, the *colors* of chromosome fragments are consistent with their *colors* in QM6a map to clearly show chromosomal rearrangements. Some emblematic genes were chosen along the sequence to be used as location markers (list available in Additional file 6). The Rut-C30 85 kb deletion event on chr. VI is shown by the lack of pks1 gene. Centromere locations are shown by restricted width. **c** Possible scenario (among others) from QM6a to Rut-C30. Translocations are numbered according to Table 3

therefore one of its flanking is an AT-rich region as previously described [14]. Except for the breakpoint chr I: 5,018,020 localized inside an AT-rich region, the %GC in a 1-kb window around the breakpoints displayed a similar or higher level than in the genome. The four events listed in Table 3 for Rut-C30 strain were already present in its ancestor NG14 [14, 15], so the chromosome structure of NG14 strain is most likely identical to Rut-C30 chromosome structure (Fig. 2b). The chromosomal rearrangements identified previously by the CGH array stud [15] and a genomics analysis [17] are in line with the contact map results obtained in this study. So it should be possible to reconstitute the karyotypes of other *T. reesei* strains for which this kind of information is available.

Inferring the chromosome structure of other *T. reesei* strains

We then confronted the QM6a chromosome structure with translocation events characterized in other *T. reesei* strains to reconstitute their expected karyotypes. Table 4 shows translocation breakpoints for the QM9414, QM9123 [15], CBS 999.97(1-1, *re*) [10], and QM9978 (Ivanova et al. to be published) strains, and their mapping on QM6a chromosomes. For each strain, the possible chromosome structure was assessed from these translocation events (Fig. 3). In QM9414 strain (Fig. 3b), two translocations involved chromosomes I, II, and VI, with among others, one fragment of chromosome II and one fragment of the VI being translocated onto chromosome I. In QM9978 (Fig. 3c), a reciprocal translocation event involved chromosomes V and VII, with the chromosome V breakpoint positioned 1.6 kb upstream the gene 54675 that encodes for the transcription factor VIB1. This rearrangement, by modifying the transcription of this gene, is responsible of the cellulase-negative phenotype of this strain (Ivanova et al. to be published). Finally, the translocation event in the diploid strain CBS 999.97 involved chromosomes II and IV, and resulted in the isolation of

Table 4 Translocation breakpoints of various *T. reesei* strains genomes

	Translocation breakpoints	Location on QM6a scaffolds [15]	Mapping on QM6a chromosomes
QM9414 & QM9123	n°1	scaffold_4: 1,190,139	chr I: 4,327,373
		scaffold_14: 118,472	chr II: 4,693,330
	n°2	scaffold_9: 787,779	chr VI: 2,237,971
		scaffold_27: 140,159	chr II: 5,788,998
CBS 999.97 (1-1, *re*)	Resulting in D-segment	scaffold_36: 54,323	chr II: 5,441,472
	Resulting in L-segment	scaffold_33: 33,249	chr IV: 4,304,165
QM9978	n°1	scaffold_1: 96,633	chr V: 1,604,851
		scaffold_16: 631,551	chr VII: 1,076,804

Translocation breakpoints were mapped on the superscaffolds generated by GRAAL

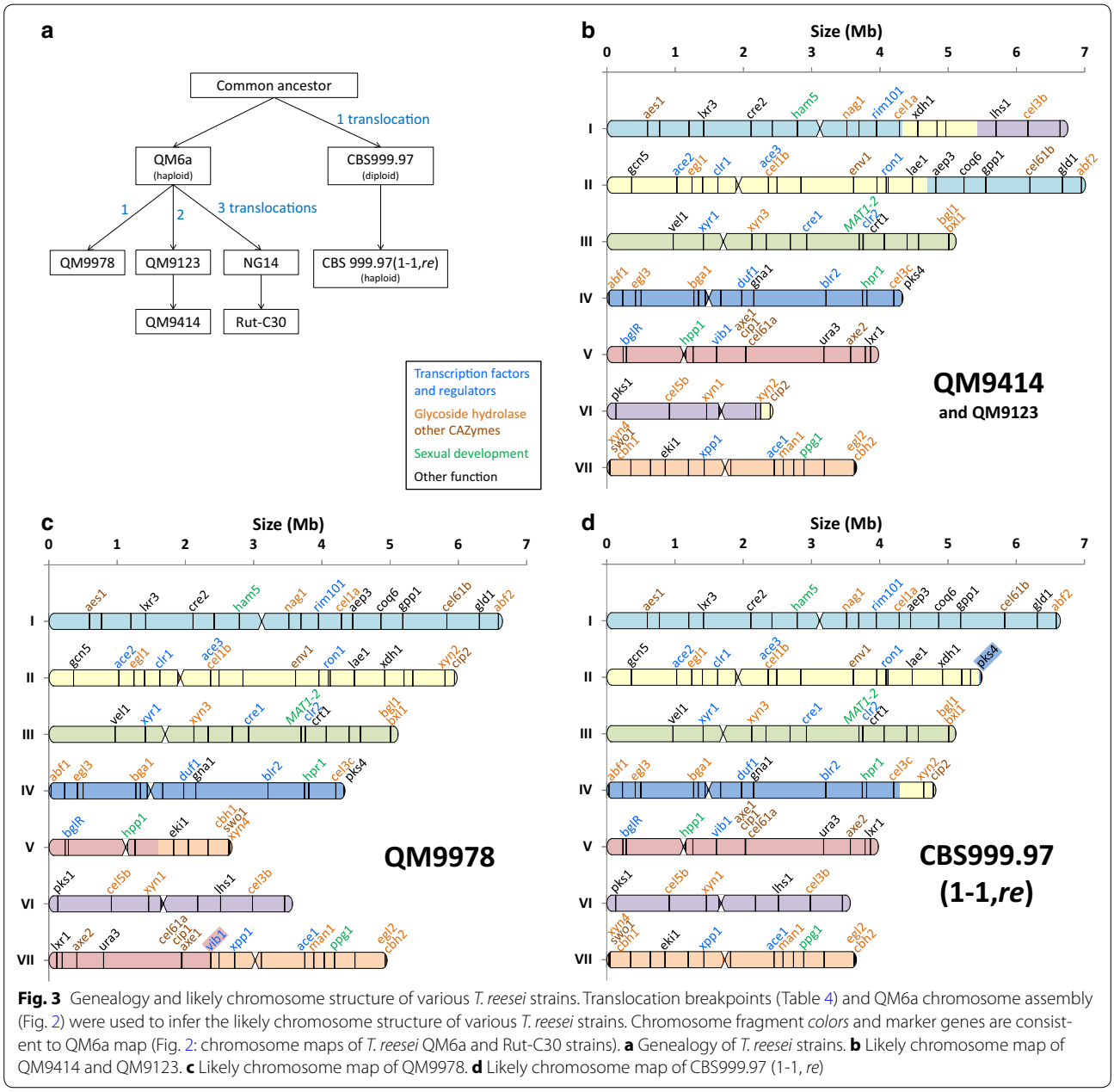


Fig. 3 Genealogy and likely chromosome structure of various *T. reesei* strains. Translocation breakpoints (Table 4) and QM6a chromosome assembly (Fig. 2) were used to infer the likely chromosome structure of various *T. reesei* strains. Chromosome fragment *colors* and marker genes are consistent to QM6a map (Fig. 2: chromosome maps of *T. reesei* QM6a and Rut-C30 strains). **a** Genealogy of *T. reesei* strains. **b** Likely chromosome map of QM9414 and QM9123. **c** Likely chromosome map of QM9978. **d** Likely chromosome map of CBS999.97 (1-1, *re*)

haploid strains either of WT or recombinant (*re*) karyotypes (Fig. 3d) [10].

Essentiality of the chromosomes fragments

When crossing CBS999.97 (1-1, *re*) with either CBS999.97 (1-2, *wt*) or QM6a, Chuang et al. showed that L-segment aneuploidy (containing 11 genes in our reassembly) is not lethal but results in a "white spore" phenotype because of the loss of the polyketide synthase 4 gene (tpks4, gene ID 82208) located on this segment [10]. On the other hand, loss of the D-segment (containing 167 genes in our reassembly) is not viable, most probably because essential genes are present on this segment. For each of the translocations listed in Tables 3 and 4, we computed the length and number of genes of the resulting chromosome fragments, from the breakpoint to the telomere (or to the next breakpoint in the case of QM9414 chromosome II and Rut-C30 chromosome I) (Table 5). Then we looked for essential genes in each of these chromosome fragments to verify whether their loss will be lethal or not.

In QM9414 strain, the fragment of chromosome II which has been translocated to chromosome VI contains only 63 genes, in which the ribosomal protein RPS24 (gene ID 81713) has been shown to be essential for 40S ribosomal subunit assembly in HeLa cells [38]. In Rut-C30 strain, the fragment of chromosome VI which has been translocated on chromosome I contains 114 genes, among which the acetyl-CoA carboxylase (geneID 81110) is presumably essential (its orthologue cut6 is essential in *S. pombe* [39]). All other chromosome fragments listed on Table 5 contain at least 290 genes. Assuming 18.7% of essential genes as in *S. cerevisiae* [40], the probability that these fragments do not contain an essential gene is below 10^{-26}. Therefore, the only translocated fragment which is not essential is the small previously described CBS999.97 (1-1, *re*) L-segment [10].

Table 5 Statistics on chromosome fragments

Strain	Chromosome	Fragment size (kb)	Nb of genes
CBS 999.97 (1-1, *re*)	chr II => chr IV (D-segment)	539	167 genes
	chr IV => chr II (L-segment)	33	11 genes
QM9414 & QM9123	chr I => chr II	2321	634 genes
	chr II => chr I	1096	322 genes
	chr II => chr VI	192	63 genes
	chr VI => chr I	1329	369 genes
QM9978	chr V => chr VII	2374	644 genes
	chr VII => chr V	1077	290 genes
Rut-C30	chr I => chr III	1133	309 genes
	chr I => chr VI	1630	442 genes
	chr III => chr III	1976	485 genes
	chr VI => chr I	402	114 genes

For each of the breakpoint described in Tables 3 and 4, the size and number of genes of the resulting chromosome fragment (from the breakpoint to the telomere or to the next breakpoint) were calculated. The only dispensable fragment is the L-segment described in CBS999.97 (1-1, *re*) [10]

Inferring lethal segmental aneuploidy in F1 progenies

Using the chromosome maps described in Figs. 2 and 3, we typically enumerated the possible chromosome structures in the F1 progeny for different crossing experiments (already described or not) involving as *MAT1-1* partner either CBS999.97 (1-1, *re*) [10] or a QM6a *MAT1-1* strain with restored female fertility [9] and checked for each structure whether it contains lethal segmental aneuploidy or not. An example of the enumeration is given on Fig. 4 for a *MAT1-1* female fertile QM6a strain crossed with Rut-C30 strain, and the results for other crossings are shown in Table 6.

When crossing CBS999.97 (1-1, *re*) with industrial strains QM9414 and Rut-C30, Chuang et al. observed much more meiotic lethality (asci with no or only four viable ascopores) than when crossing with QM6a. Our theoretical results are consistent with their experimental results: while enumerating the viable chromosome structures, we observed that whereas 75% of the possible chromosome structures are viable when crossing CBS999.97 (1-1, *re*) with QM6a, only 25–28% are viable when crossing with QM9414 or Rut-C30, respectively (Table 6). For

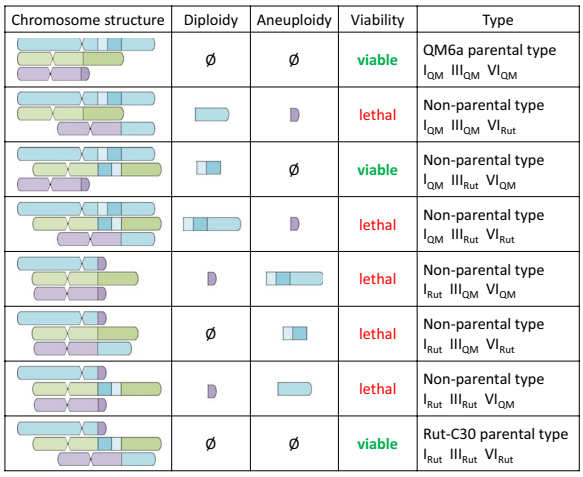


Fig. 4 Possible chromosome structures in F1 progeny resulting from a crossing between a *MAT1-1* female fertile QM6a strain and Rut-C30 strain. Using the chromosome structure of QM6a and Rut-C30 strains, we enumerated the possible chromosome structures in F1 progeny (only chromosomes I, III, and VI are shown here with *colors* consistent to Fig. 3c). For each possible structure, the fragmental diploidy or aneuploidy is shown. Since the chromosome fragments contain essential genes, segmental aneuploidy results in inviable progeny

Table 6 Analyses of possible chromosome structures for different crossing experiments

Crossing experiment	Nb ≠ chr	Total possible structures	Non-viable	Viable	Possible viable structures different from parental ones
CBS999.97 (1-1, *re*) × CBS999.97 (1-2, *wt*) or × QM6a	2	$2^2 = 4$	1	3 (75%)	1 structure with chr II fragment (D-segment) diploidy
CBS999.97 (1-1, *re*) × QM9414	4	$2^4 = 16$	12	4 (25%)	1 structure with chr II fragment diploidy
					1 structure with chr II fragment diploidy and chr VI fragment diploidy
CBS999.97 (1-1, *re*) × Rut-C30	5	$2^5 = 32$	23	9 (28%)	1 haploid with QM6a structure,
					1 crossed-haploid,
					4 structures with 1 chr fragment diploidy,
					1 structure with 2 chr fragment diploidy
QM6a (*MAT1-1, ff*) × QM6a	0	1	0	1 (100%)	None
QM6a (*MAT1-1, ff*) × QM9414	3	$2^3 = 8$	6	2 (25%)	None
QM6a (*MAT1-1, ff*) × Rut-C30	3	$2^3 = 8$	5	3 (38%)	1 structure with chr I fragment diploidy

The first three cases have already been experimentally described [10]. The next 3, involving a *MAT1-1* female fertile (ff) QM6a strain, have not yet been described. We assumed that crossing-over were possible but not in translocated parts

the not yet described crossings involving a *MAT1-1* female fertile QM6a strain, we similarly noticed that only 25 and 38% of the possible structures are viable when crossing with QM9414 and Rut-C30, respectively (Table 6). When crossing with Rut-C30, only one non-parental chromosome structure is viable (Fig. 4). When crossing with QM9414, the only possible chromosome structures are the two parental structures (Table 6). Using CBS999.97 (1-1, *re*), Chuang et al. had suggested that crossing should be used cautiously to improve industrial strains [10]. Our analysis shows that this is not due to the specific chromosome structure of this strain: using QM6a as a *MAT1-1* partner for crossing with industrials strains will result in almost the same meiotic lethality.

Discussion

Chromosome assembly

Chromosome contact data resulting from the sequencing of 3C/Hi-C libraries represent powerful information to improve or complete genome scaffolding [23]. Genome reassembly algorithms like GRAAL are based on polymer physics principles, and as such, give trustworthy, statistically sound, information about the relative position of each pair of fragments along each chromosome sequence, even when the fragments' regions are separated by gaps which had failed to be sequenced and assembled previously. In that regard, this pipeline based on contact data outperforms current deep sequencing when trying to prove that two sequences are neighboring. For instance, GRAAL successfully integrated 63 pairs of such scaffold fragments into the QM6A reassembly which had failed to be assembled during the initial sequencing. Moreover, it was able to identify six misassemblies in the initial genome. Here, we showed that GRAAL was able to reassemble the Rut-C30 chromosomes using the QM6a sequence as a reference, and to correctly identify the six breakpoint locations specific to Rut-C30 (in addition to the misassemblies commonly found in QM6a). The pipeline can therefore identify a chromosome structure even when its sequence is not precisely known or when numerous chromosomal rearrangements occur. It could be applied with great potential to other strains, e.g., ones resulting from sexual crossing, without the need to get a sequence of these strains beforehand. Because the Rut-C30 contact map reflects the average genome organization of this strain (independently of the QM6a chromosome structure since only the reference scaffolds were used), the data could also be used for a more in-depth investigation of variations in the chromosomal contacts/interactions pattern between the two strains. However, since GRAAL is a reassembly pipeline, it does not give new information about the sequence in itself, so additional sequencing or computational work is required to fill-in the gaps between reassembled scaffolds. Misassemblies or translocation breakpoints are here identified with a ≈10 kb precision, which is sufficient here given the precise breakpoints have already been sequenced. In the case of a new strain, a chromosome walking iterative alignment of 3C-seq reads on the sequence should probably allow the identification of translocation breakpoints with the same base-pair precision.

Centromere location and composition

Centromeres are defined as "chromosomal elements that are both necessary and sufficient for chromosome segregation" [32]. These regions display a remarkable diversity in size and structure, ranging from the so-called point 125-bp centromeres in *S. cerevisiae* to several megabases

sequence of satellite DNA in plants and animals. Fungal centromeres typically range from 30 to 450 kb in size. While the point centromeres sequences seem sufficient to provide centromeric function, bigger centromeres seem to be defined epigenetically. The lack of sequence consensus even between centromeres of the same organism, and the low complexity of these AT-rich sequences have made identification and sequencing of centromeres challenging. The discovery of the centromeric histone CenH3 as the landmark of centromere regions has made chromatin immunoprecipitation the method of choice to functionally distinguish centromeric regions from other low complexity repeated regions. Here, the "Rabl" pattern of chromosomal structure in *T. reesei* observed in our previous work [23] prompted us to take advantage of the physical proximity between centromeres in this specific spatial chromosome organization for the identification of their location along the sequence [25]. The chromosomal contact data are therefore a functional proof of the centromeric nature of these sequences. As expected, the centromeric regions we determined were nearly devoid of coding sequences [21].

Interestingly, we observed in four centromeric regions (scaffolds 51, 56, 57, 58) a 7- to 10-kb long inverted repeat regions, reminiscent of inverted repeats organization found in yeasts *C. albicans, C. tropicalis*, *K. phaffii*, or *S. pombe* [31, 33–37]. To our knowledge, such an organization has not been described in filamentous fungi, as most data come from the study of *N. crassa*, whose centromeric region are 150–300 kb long and consist in degenerate transposon sequences. This raises the question of whether at least some centromeres in *Trichoderma* are sequence- or at least inverted repeat-defined, as recently hypothesized for *C. tropicalis* [37] and not only epigenetically defined. Such observation could have an influence on efforts to develop a plasmid transformation system in this fungus. Apparently, these large inverted repeat features are not unique to *Trichoderma*, as we were able to make similar observations in *Fusarium graminearum* by analyzing the latest genome sequence [41] (see Additional file 4).

Importance of chromosome structure for analyses of crossing experiment

Knowing QM6a karyotype and chromosome translocations in some of its derivatives, we were able to predict the karyotypes of other *T. reesei* strains, from three lineages different from the NG14/Rut-C30 lineage, and to infer the possible chromosome structure in the F1 progeny for different crossing experiments involving these strains. Doing so, we managed to explain the higher meiotic lethality observed by Chuang et al. when crossing CBS999.97 (1-1,*re*) with industrial strains QM9414 and Rut-C30 compared to crossing with the natural isolate QM6a [10]. Chromosomal rearrangements resulted in chromosome structures which are not completely compatible any more in the two parents, producing lethal segmental aneuploidy in F1 progeny and conversely producing viable F1 progeny with a limited diversity in chromosome structure. This will obviously result in a limited diversity of sequence in the viable F1 progeny, since translocated fragments will undergo much less crossing-over, if any, than other parts of the genome. This imbalance may be an issue for genetics analysis-based experiments like bulk segregant analysis and for industrial strains improvement.

Conclusions

In this work, we exploited chromosome contact data and the program GRAAL to both complete the assembly/scaffolding of the *T. reesei* reference genome, and identify its centromeres positions. That the method is robust was supported by performing the same analysis on the Rut-C30 strain, a derivative of the reference strain, which confirmed both centromeres identification and previously identified chromosome translocations in this strain. Finally, given chromosomal translocations occurred in different strain lineages of this fungus, we illustrated the importance of our data by showing predicted karyotypes of several strains and predicted consequences on crossing experiments between strains. The recent possibilities offered by strain crossings in *T. reesei* will possibly make such data and similar analyses essential in future industrial fungal research.

Methods

Strain and cultures

Trichoderma reesei Rut-C30 (strain ATCC 56765) strain was cultured in bioreactor as described previously [29].

Construction of 3C libraries

For *T. reesei* QM6a, the construction of 3C library has already been described previously [23]. For Rut-C30 strain, the 3C library was constructed following exactly the same protocol and restriction enzyme (DpnII).

GRAAL assembly

Genome (Re)Assembly Assessing Likelihood from 3D (GRAAL) is an algorithm which uses chromosome conformation capture (3C) data to rescaffold contigs and improve genome assembly [23]. Briefly, the original genome is first split into bins containing the same number of restriction fragments (a restriction fragment is a genome region between two restriction sites of the enzyme used for the 3C library construction), then the reads from the 3C library are mapped onto these bins

so as to compute an initial contact matrix, each entry therein representing the contact frequency between each bin pair and bins being ordered along the initial genome assembly. This matrix shows contact discrepancies since the original genome is not fully assembled. Finally, GRAAL reorders the bins so as to get the most likely matrix based on what contact frequency distribution is expected from chromosomes according to a standard polymer physics model [42]. The *T. reesei* QM6a chromosome sequence we previously published is an example of the raw output from the algorithm.

Manual corrections

Several GRAAL computations were performed with different bin sizes to assess the assembly's robustness. Then manual corrections were performed to go beyond the limitations of GRAAL and other reassembly programs. Since scaffolds were split into bins with the same number of restriction fragments, scaffold ends were too small (sequencing coverage too low) to be included in the computation, so were lost in the raw output sequence. We manually added them so as to get the entire scaffolds in the reassembly. When a scaffold is misassembled in the original genome, GRAAL is able to find the splitting location at an accuracy depending on the size of bins involved in the splitting (around 10–50 kb depending on the definition of the bins, and on the location of the restriction sites). We checked the sequence around the splits and most of the time we noticed nearby the presence of ≈1 kb NNN sequences, so we manually corrected the split location to be consistent with this gap location. Reassembly programs like GRAAL easily reorder bins using contact data, but they may fail in finding the correct bin orientation, so many bins were switched (by comparison with the neighboring bins from the same original scaffold) in the raw output sequence. We manually corrected them to get the scaffolds as in the original assembly without switching bins. However, some scaffolds were too small to get a reliable orientation, in this case, we arbitrarily chose the forward direction for the sequence available in Additional file 2. Seven telomere repeats were identified in the original sequence [13] and six of them were assembled in the chromosomes, as noticed previously [26] although they were not at chromosome ends in the raw output sequence. We checked their presence at chromosome ends, and used them three times to identify the correct bin directions (for scaffold 45, 46, and 64 in chromosomes III, V, and IV, respectively). As for scaffold 31 on chromosome VI, we deleted 7 kb at the 3′ end because they were duplications of the telomere sequence. Around 20–30 bins (<4% of the total number of bins) had not been reassembled because their signature in the contact matrix was not strong enough for GRAAL. We manually checked the contact matrix and reassembled these bins in the final sequence depending on their contact signature (telomere, centromere, standard). Finally, the gene annotation from the JGI (gtf file for the Filtered Models set of genes, [43]) was mapped to the reassembled sequence in order to get the coordinates of the 9129 genes on the chromosomes.

Centromere positions

Centromere positions along the chromosomes have been manually identified using the contact data (see Additional file 7 for raw data contact frequencies over the entire genome). Because of their Rabl organization, centromeres have stronger interaction with each other than with their neighboring sequences.

Gene enrichment analysis

To calculate the enrichment in genes close to the centromeres, we used the gene annotations (GO terms and EC numbers) from the JGI [43] and from the FungiPath database [44–46], and performed the enrichment analysis with the Pathway Tools software [47]. A 50-kb window was defined around the centromeres, which resulted in a set of 238 genes (2.6% of the genome).

Additional files

Additional file 1. Details on QM6a reassembly.

Additional file 2. QM6a reassembly sequence in fasta format (7 chromosomes + unassembled scaffolds).

Additional file 3. Annotation file describing the location on the chromosomes of i) the original scaffolds, ii) the centromeres and iii) the 9129 genes from the JGI Filtered Models set of genes.

Additional file 4. Identification of inverted repeats in *T. reesei* and *F. graminearum* centromeres.

Additional file 5. Details on Rut-C30 reassembly.

Additional file 6. List of gene markers used on Figs. 2 and 3, with their names, IDs, locations on scaffolds and chromosomes, and functional annotations.

Additional file 7. Raw data contact frequencies over the entire genome.

Abbreviations

PFGE: pulse-field gel electrophoresis; 3C: chromosome conformation capture; CDS: coding sequence.

Authors' contributions

DPP, YV, and MM prepared the Rut-C30 3C library. LB and MM performed the GRAAL assemblies. EJ performed the manual curation and the assemblies analysis, and is a major contributor in writing the manuscript. FB, MM, AM, and RK supervised the study and contributed in writing the manuscript. All authors read and approved the final manuscript.

Author details

[1] IFP Energies nouvelles, 1 et 4 Avenue de Bois-Préau, 92852 Rueil-Malmaison, France. [2] Groupe Régulation Spatiale des Génomes, Department Genomes and Genetics, Institut Pasteur, 75015 Paris, France. [3] UMR 3525, CNRS, 75015 Paris, France.

Acknowledgements
We acknowledge Ting-Fang Wang for sharing with us some data about his independent work on *T. reesei* resequencing, so as we could match up our nomenclature.

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

Availability of data and materials
The Rut-C30 3C library sequencing data analyzed during this study were deposited on NCBI BioProject Database under the Accession Number PRJNA355969.

Funding
LB and MM were supported by funding to R.K. from the European Research Council under the 7th Framework Program (FP7/2007-2013)/ERC Grant Agreement 260822.

Publisher's Note

Received: 17 December 2016 Accepted: 31 May 2017
Published online: 12 June 2017

References

1. Bischof RH, Ramoni J, Seiboth B. Cellulases and beyond: the first 70 years of the enzyme producer *Trichoderma reesei*. Microb Cell Fact. 2016;15:106.
2. Ben Chaabane F, Chaussepied B. Process for the continuous production of cellulases by a filamentous fungus using a carbon substrate obtained from an acid pretreatment. US Patent 9249402 B2; 2016.
3. Nevalainen H, Peterson R. Chapter 7—heterologous expression of proteins in *Trichoderma*. In: Gupta VG, Schmoll M, Herrera-Estrella A, Upadhyay RS, Druzhinina I, Tuohy M, editors. Biotechnology and biology of *Trichoderma*. Amsterdam: Elsevier; 2014. p. 89–102.
4. Landowski CP, Mustalahti E, Wahl R, Croute L, Sivasiddarthan D, Westerholm-Parvinen A, et al. Enabling low cost biopharmaceuticals: high level interferon alpha-2b production in *Trichoderma reesei*. Microb Cell Fact. 2016;15:104.
5. Seidl V, Seibel C, Kubicek CP, Schmoll M. Sexual development in the industrial workhorse *Trichoderma reesei*. Proc Natl Acad Sci. 2009;106:13909–14.
6. Lieckfeldt E, Kullnig C, Samuels GJ, Kubicek CP. Sexually competent, sucrose- and nitrate-assimilating strains of *Hypocrea jecorina* (*Trichoderma reesei*) from South American soils. Mycologia. 2000;92:374–80.
7. Peterson R, Nevalainen H. *Trichoderma reesei* RUT-C30—thirty years of strain improvement. Microbiology. 2012;158:58–68.
8. Seibel C, Tisch D, Kubicek CP, Schmoll M. The role of pheromone receptors for communication and mating in *Hypocrea jecorina* (*Trichoderma reesei*). Fungal Genet Biol. 2012;49:814–24.
9. Linke R, Thallinger GG, Haarmann T, Eidner J, Schreiter M, Lorenz P, et al. Restoration of female fertility in *Trichoderma reesei* QM6a provides the basis for inbreeding in this industrial cellulase producing fungus. Biotechnol Biofuels. 2015;8:155.
10. Chuang Y-C, Li W-C, Chen C-L, Hsu PW-C, Tung S-Y, Kuo H-C, et al. *Trichoderma reesei* meiosis generates segmentally aneuploid progeny with higher xylanase-producing capability. Biotechnol Biofuels. 2015;8:30.
11. Carter GL, Allison D, Rey MW, Dunn-Coleman NS. Chromosomal and genetic analysis of the electrophoretic karyotype of *Trichoderma reesei*: mapping of the cellulase and xylanase genes. Mol Microbiol. 1992;6:2167–74.
12. Mäntylä AL, Rossi KH, Vanhanen SA, Penttilä ME, Suominen PL, Nevalainen KH. Electrophoretic karyotyping of wild-type and mutant *Trichoderma longibrachiatum* (*reesei*) strains. Curr Genet. 1992;21:471–7.
13. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, et al. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat Biotech. 2008;26:553–60.
14. Seidl V, Gamauf C, Druzhinina IS, Seiboth B, Hartl L, Kubicek CP. The *Hypocrea jecorina* (*Trichoderma reesei*) hypercellulolytic mutant RUT C30 lacks a 85 kb (29 gene-encoding) region of the wild-type genome. BMC Genom. 2008;9:327.
15. Vitikainen M, Arvas M, Pakula T, Oja M, Penttilä M, Saloheimo M. Array comparative genomic hybridization analysis of *Trichoderma reesei* strains with enhanced cellulase production properties. BMC Genom. 2010;11:441.
16. Le Crom S, Schackwitz W, Pennacchio L, Magnuson JK, Culley DE, Collett JR, et al. Tracking the roots of cellulase hyperproduction by the fungus *Trichoderma reesei* using massively parallel DNA sequencing. Proc Natl Acad Sci. 2009;106:16151–6.
17. Koike H, Aerts A, LaButti K, Grigoriev IV, Baker SE. Comparative Genomics Analysis of *Trichoderma reesei* Strains. Ind Biotechnol. 2013;9:352–67.
18. Porciuncula JdO, Furukawa T, Mori K, Shida Y, Hirakawa H, Tashiro K, et al. Single nucleotide polymorphism analysis of a Trichoderma reesei hyper-cellulolytic mutant developed in Japan. Biosci Biotechnol Biochem. 2013;77:534–43.
19. Nitta M, Furukawa T, Shida Y, Mori K, Kuhara S, Morikawa Y, Ogasawara W. A new Zn(II)(2)Cys(6)-type transcription factor BglR regulates beta-glucosidase expression in *Trichoderma reesei*. Fungal Genet Biol. 2012;49:388–97.
20. Thomma BP, Seidl MF, Shi-Kunne X, Cook DE, Bolton MD, van Kan JA, Faino L. Mind the gap; seven reasons to close fragmented genome assemblies. Fungal Genet Biol. 2016;90:24–30.
21. Smith KM, Galazka JM, Phatale PA, Connolly LR, Freitag M. Centromeres of filamentous fungi. Chromosome Res. 2012;20:635–56.
22. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing Chromosome Conformation. Science. 2002;295:1306–11.
23. Marie-Nelly H, Marbouty M, Cournac A, Flot J-F, Liti G, Parodi DP, et al. High-quality genome (re)assembly using chromosomal contact data. Nat Commun. 2014;5:5695.
24. Flot J-F, Marie-Nelly H, Koszul R. Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. FEBS Lett. 2015;589:2966–74.
25. Marie-Nelly H, Marbouty M, Cournac A, Liti G, Fischer G, Zimmer C, Koszul R. Filling annotation gaps in yeast genomes using genome-wide contact maps. Bioinformatics. 2014;30:2105–13.
26. Druzhinina IS, Kopchinskiy AG, Kubicek EM, Kubicek CP. A complete annotation of the chromosomes of the cellulase producer *Trichoderma reesei* provides insights in gene clusters, their expression and reveals genes required for fitness. Biotechnol Biofuels. 2016;9:75.
27. GRAAL GitHub repository. http://github.com/koszullab/GRAAL. Accessed 1 Dec 2016.
28. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, et al. A three-dimensional model of the yeast genome. Nature. 2010;465:363–7.
29. Poggi-Parodi D, Bidard F, Pirayre A, Portnoy T, Blugeon C, Seiboth B, et al. Kinetic transcriptome analysis reveals an essentially intact induction system in a cellulase hyper-producer *Trichoderma reesei* strain. Biotechnol Biofuels. 2014;7:173.
30. Nambiar M, Smith GR. Repression of harmful meiotic recombination in centromeric regions. Semin Cell Dev Biol. 2016;54:188–97.
31. Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, et al. The genome sequence of *Schizosaccharomyces pombe*. Nature. 2002;415:871–80.
32. Malik HS, Henikoff S. Major evolutionary transitions in centromere complexity. Cell. 2009;138:1067–82.
33. Nakaseko Y, Adachi Y, Funahashi SI, Niwa O, Yanagida M. Chromosome walking shows a highly homologous repetitive sequence present in all the centromere regions of fission yeast. EMBO J. 1986;5:1011–21.
34. Fishel B, Amstutz H, Baum M, Carbon J, Clarke L. Structural organization and functional analysis of centromeric DNA in the fission yeast *Schizosaccharomyces pombe*. Mol Cell Biol. 1988;8:754–63.
35. Sanyal K, Baum M, Carbon J. Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. Proc Natl Acad Sci USA. 2004;101:11374–9.
36. Chatterjee G, Sankaranarayanan SR, Guin K, Thattikota Y, Padmanabhan S, Siddharthan R, Sanyal K. Repeat-associated fission yeast-like regional centromeres in the ascomycetous budding yeast *Candida tropicalis*. PLoS Genet. 2016;12:e1005839.

37. Coughlan AY, Hanson SJ, Byrne KP, Wolfe KH. Centromeres of the yeast *Komagataella phaffii* (*Pichia pastoris*) have a simple inverted-repeat structure. Genome Biol Evolut. 2016;8:2482–92.
38. Choesmel V, Fribourg S, Aguissa-Touré A-H, Pinaud N, Legrand P, Gazda HT, Gleizes P-E. Mutation of ribosomal protein RPS24 in Diamond-Blackfan anemia results in a ribosome biogenesis disorder. Hum Mol Genet. 2008;17:1253–63.
39. Saitoh S, Takahashi K, Nabeshima K, Yamashita Y, Nakaseko Y, Hirata A, Yanagida M. Aberrant mitosis in fission yeast mutants defective in fatty acid synthetase and acetyl CoA carboxylase. J Cell Biol. 1996;134:949–61.
40. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature. 2002;418:387–91.
41. King R, Urban M, Hammond-Kosack MCU, Hassani-Pak K, Hammond-Kosack KE. The completed genome sequence of the pathogenic ascomycete fungus *Fusarium graminearum*. BMC Genom. 2015;16:544.
42. Rippe K. Making contacts on a nucleic acid polymer. Trends Biochem Sci. 2001;26:733–40.
43. *Trichoderma reesei* v2.0, on the JGI Genome Portal. http://genome.jgi.doe.gov/Trire2/Trire2.home.html. Accessed 8 Nov 2016.
44. FungiPath database. http://fungipath.i2bc.paris-saclay.fr. Accessed 8 Nov 2016.
45. Grossetête S, Labedan B, Lespinet O. FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. BMC Genom. 2010;11:81.
46. Pereira C, Denise A, Lespinet O. A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genom. 2014;15(Suppl 6):S16.
47. Karp PD, Paley S, Romero P. The pathway tools software. Bioinformatics. 2002;18:S225–32.

T. reesei QM6a reassembly

65 scaffolds from the JGI reference genome (33.3Mb - 99.5% of the genome) have been reassembled in 7 chromosomes, as follows :

CHR I

scaffold		start	end	direction	
7	full	1	1 429 972	-1	
12	full	1	1 022 062	-1	
43	full	1	74 996	-1	
21	full	1	576 034	-1	
55	full	1	33 670	?	centromere - direction uncertainty
4	full	1	1 832 615	1	
49	full	1	46 304	-1?	direction uncertainty
48	full	1	48 367	1	
5	fragment	1	1 583 115	-1	split location identified by Ns gap

size : 6 647 935 (with 100bp Ns spacers between each scaffold)

CHR II

scaffold		start	end	direction	
31	fragment	1	224 034	-1	telomere repeats (4 duplications of the telomere sequence have been deleted, 7kb)
41	full	1	80 626	-1	
25	full	1	439 677	1	
68	full	1	10 734	1	
10	full	1	1 156 739	-1	
66	full	1	11 200	?	centromere - direction and order uncertainty
59	full	1	18 517	?	centromere - direction and order uncertainty
8	full	1	1 408 331	1	
34	full	1	166 473	-1	
26	full	1	433 400	1	
14	full	1	861 070	-1	
23	full	1	512 080	-1	
54	full	1	34 758	1?	direction uncertainty
36	full	1	136 855	-1	
27	full	1	433 262	-1	
28	fragment	367 024	407 093	?	split location identified by Ns gap - direction uncertainty
67	full	1	11 021	1?	direction uncertainty

size : 5 980 447 (with 100bp Ns spacers between each scaffold)

CHR III

scaffold		start	end	direction	
45	full	1	65 952	-1	telomere repeats OK
69	full	1	10 696	?	direction uncertainty
35	full	1	152 537	1	
32	full	1	230 370	-1	
11	full	1	1 155 933	-1	
40	full	1	89 857	1	
					centromere position (no centromere scaffold reliable assembled)
2	fragment	154 748	2 007 204	-1	split location uncertainty (154 748 was chosen here after alignment with fragment 1-98434)
2	fragment	1	98 434	-1	split location identified by Ns gap
6	full	1	1 455 714	-1	telomere repeats OK

size : 5 112 650 (with 100bp Ns spacers between each scaffold)

CHR IV

scaffold		start	end	direction	
64	full	1	14 482	1	telomere repeats
19	full	1	663 018	-1	
17	full	1	797 352	1	
56	full	1	32 194	?	centromere - direction uncertainty
20	full	1	629 213	-1	
1	fragment	2 981 735	3 756 989	1	split location identified by Ns gap
5	fragment	1 584 116	1 729 360	1	split location identified by Ns gap
13	full	1	891 309	-1	
2	fragment	99 435	154 747	-1	split location uncertainty (around 152 to 158 kb, 154747 chosen after alignment of the 2 other fragments)
38	full	1	125 035	-1	
33	full	1	207 997	-1	

size : 4 337 413 (with 100bp Ns spacers between each scaffold)

CHR V

scaffold		start	end	direction	
46	full	1	62 252	1	telomere repeats
30	full	1	247 268	1	
42	full	1	78 584	1	
53	full	1	36 593	1	
18	full	1	685 578	1	
61	full	1	15 406	-1?	centromere - direction and order uncertainty
60	full	1	15 714	1?	centromere - direction and order uncertainty
28	fragment	1	366 023	-1	split location identified by Ns gap
1	fragment	1	2 471 118	1	telomere repeats - split location identified by Ns gap and presence of telomere repeats

size : 3 979 336 (with 100bp Ns spacers between each scaffold)

CHR VI

scaffold		start	end	direction	remarks
47	full	1	50 543	-1?	direction uncertainty
15	full	1	837 556	1	
44	full	1	66 247	-1	
62	full	1	15 337	-1?	direction uncertainty
1	fragment	2 471 169	2 980 271	-1	split location identified by Ns gap and presence of telomere repeats on the other side
50	full	1	45 663	-1	
37	full	1	132 540	1	
51	full	1	43 169	?	centromere - direction uncertainty
39	full	1	105 148	1	
9	full	1	1 219 543	-1	
22	full	1	541 456	1	

size : 3 567 305 (with 100bp Ns spacers between each scaffold)

CHR VII

scaffold		start	end	direction	
29	full	1	382 182	-1	
24	full	1	501 049	-1	
16	full	1	824 923	-1	
52	full	1	41 083	?	centromere - direction uncertainty
3	full	1	1 910 749	-1	

size : 3 660 386 (with 100bp Ns spacers between each scaffold)

>scaffold_57

scaffold		start	end	direction	
57	full	1	25 756	1	centromere signature but not reliably assembled

>scaffold_58

scaffold		start	end	direction	
58	full	1	21 040	1	centromere signature but not reliably assembled

>scaffold_63

scaffold		start	end	direction
63	full	1	14 539	1

>scaffold_65

scaffold		start	end	direction	
65	full	1	12 580	1	centromere signature but not reliably assembled

>scaffold_70

scaffold		start	end	direction
70	full	1	8 513	1

>scaffold_71

scaffold		start	end	direction
71	full	1	6 846	1

>scaffold_72

scaffold		start	end	direction
72	full	1	6 811	1

>scaffold_73

scaffold		start	end	direction
73	full	1	6 421	1

>scaffold_74

scaffold		start	end	direction
74	full	1	5 890	1

>scaffold_75

scaffold		start	end	direction
75	full	1	5 683	1

>scaffold_76

scaffold		start	end	direction
76	full	1	5 459	1

>scaffold_77

scaffold		start	end	direction
77	full	1	5 371	1

>scaffold_78

scaffold		start	end	direction
78	full	1	5 154	1

>scaffold_79

scaffold		start	end	direction
79	full	1	4 691	1

>scaffold_80

scaffold		start	end	direction
80	full	1	4 619	1

>scaffold_81

scaffold		start	end	direction
81	full	1	4 614	1

>scaffold_82

scaffold		start	end	direction
82	full	1	4 370	1

>scaffold_83

scaffold		start	end	direction
83	full	1	3 796	1

>scaffold_84

scaffold		start	end	direction
84	full	1	3 468	1

>scaffold_85

scaffold		start	end	direction
85	full	1	3 089	1

>scaffold_86

scaffold		start	end	direction
86	full	1	3 000	1

>scaffold_87

scaffold		start	end	direction
87	full	1	2 158	1

Additional file 4 : identification of inverted repeats in *T. reesei* and *F. graminearum* centromeres

In 4 cases (scaffolds 51 (chr I), 56 (chr IV), 57 and 58), we observed in centromere scaffolds an inverted repeat structure with a central core region of 1 to 2 kb surrounded by an inverted repeat of 2.5 to 5 kb (Figure S2A). This structure seems quite similar to the centromere structure of *S. pombe* [1–3], *Candida tropicalis* [4] and *Komagataella phaffii* (formerly *Pichia partoris*) [5] (Figure S2B).

We annotated these sequences “mid” for the central cores and , “LR” for the left repeat, and “RR” for the right repeat, consistently with *C. tropicalis* and *K. phaffii* [4, 5], followed by the chromosome or scaffold number (Figure S2A below). The LR4 and RR4 sequences of the inverted repeat of chr IV centromere (scaffold 56) share 92% identity on 4kb without any gaps. In the 3 other cases, the LR and RR sequences share ≈58% identity but with large gaps (identity reaches 84 to 92% while excluding gaps).

While these observations could result from a misassembly of these AT-rich regions, they suggest that centromere structure in *Trichoderma* is significantly different from what is described in *Neurospora* and other filamentous fungi [6], and share some similarities with structures observed in *Taphrinomycotina* and *Saccharomycetales*.

Moreover, using the latest *Fusarium graminearum* genome release [7], we observed undescribed similar inverted repeats in the centromeres of *F. graminearum* chromosomes 1 and 2 (Figure S2B).

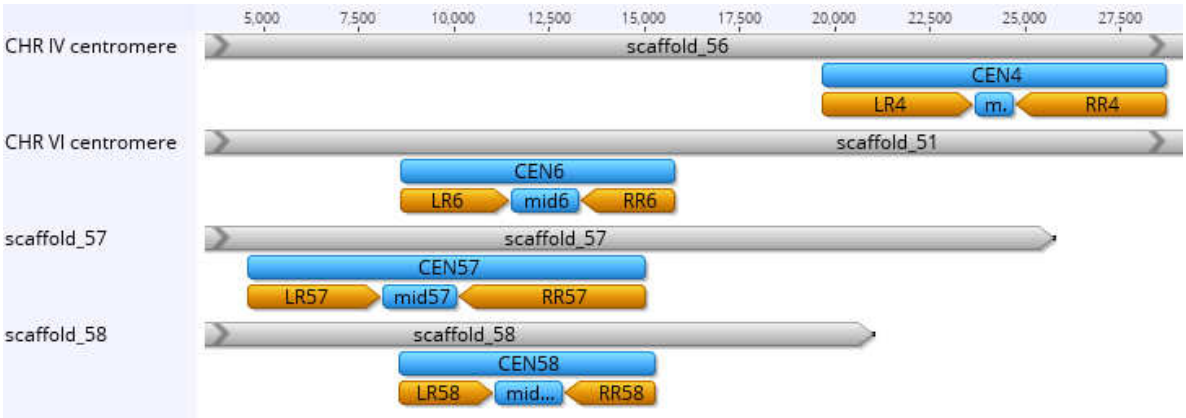
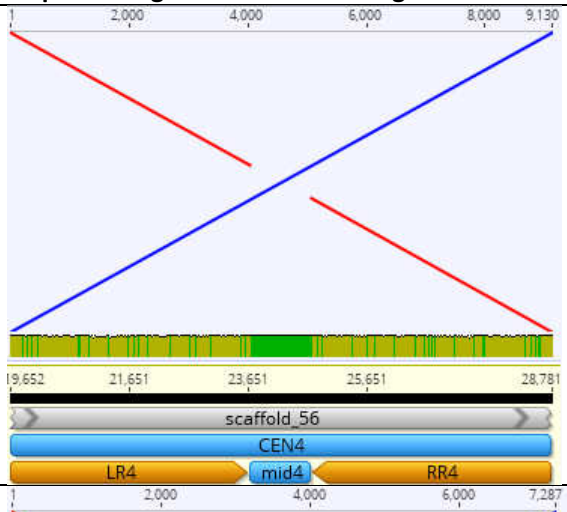
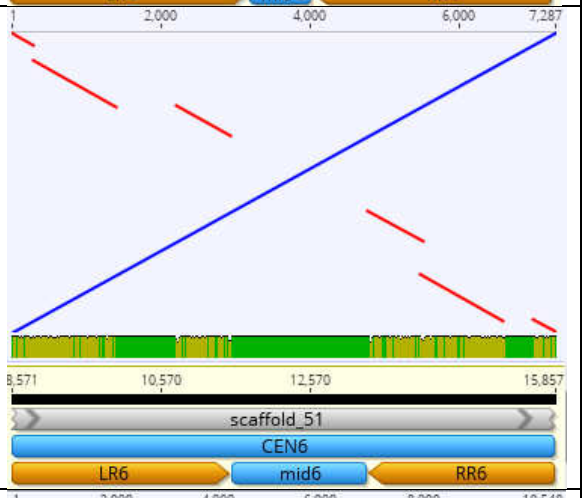
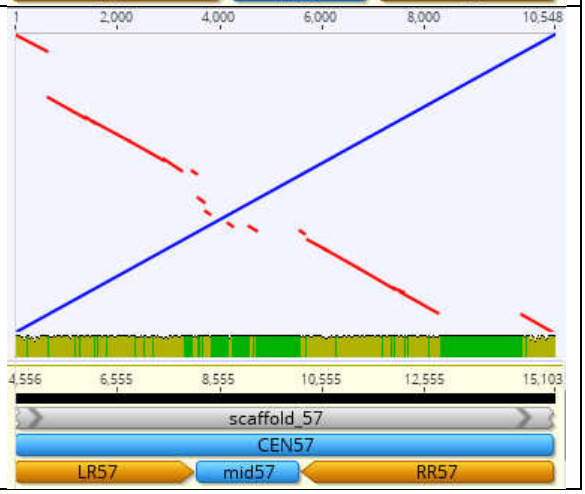
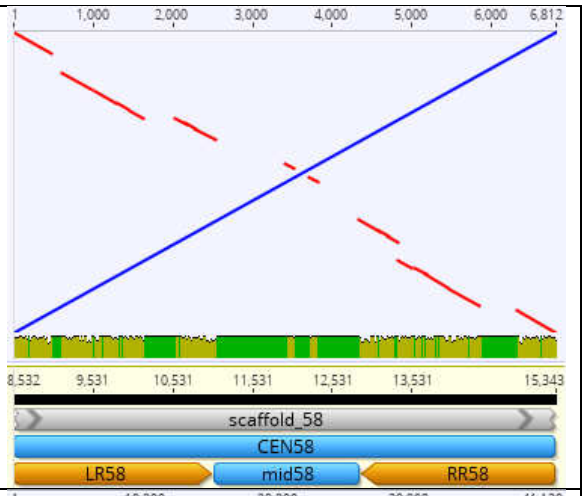
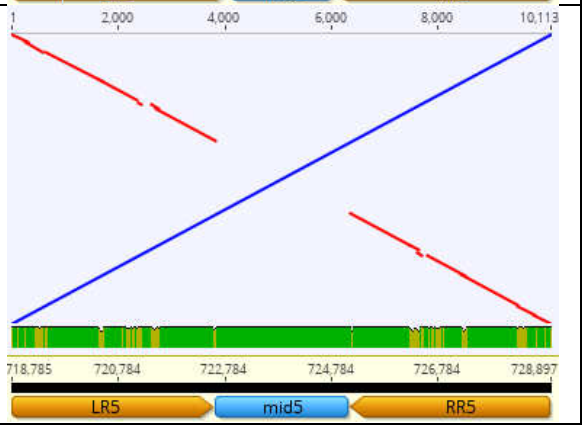


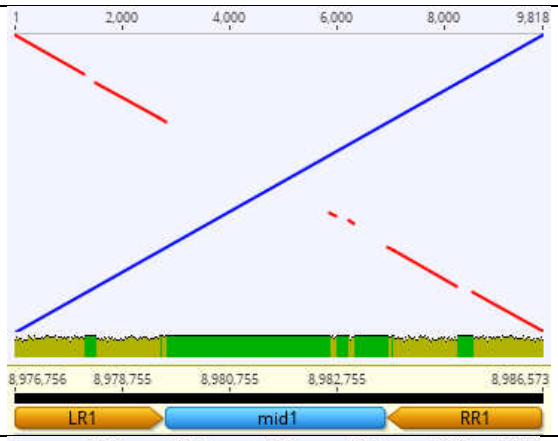
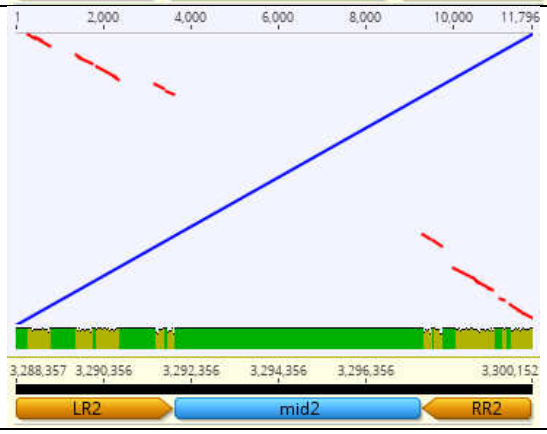
Figure S2A: Inverted repeats found on centromere-involved scaffolds
Four similar structures with a central core (mid) region surrounded by an inverted repat (LR and RR) sequences were identified on 4 scaffolds involved in *T. reesei* centromeres (scaffold 56 in chr IV centromere, scaffold 51 in chr VI centromere, and scaffolds 57 and 58 with centromere signature but not assembled).

Figure S2B: Sequence alignment of centromeres on themselves

Core centromere sequences (containing LR, mid and RR sequences) have been aligned against themselves using the LASTZ software [8, 9] with default parameters, in order to show the inverted repeats. This figure includes the 4 sequences from *T. reesei*, and arbitrary chosen sequences from *S. pombe, K. phafii, C. tropicalis* and *F. graminearum*.

Strain and location	Sequence alignment on itself using LASTZ
T. reesei CEN4 9,130 bp Scaffold_56:19,652-28,781	
T. reesei CEN6 7,287 bp Scaffold_51:8,571-15,857	
T. reesei CEN57 10,548 bp Scaffold_57:4,556-15,103	

T. reesei CEN58 6,812 bp Scaffold_58:8,532-15,343	
Schizosaccharomyces pombe CEN2 41,139 bp Chromosome II: 1,602,264-1,6447,747	
Komagataella phaffii (Pichia pastoris) CEN2 6,655 bp Chromosome 2 (FR839629.1): 843,845-850,499	
Candida tropicalis CEN5 10b113 bp Supercontig3.5 (GG692399.1): 718,785-728,897	

Fusarium graminearum CEN1 9,818 bp Chromosome 1: 8,976,756-8,986,573	
Fusarium graminearum CEN2 11,796 bp Chr2: 3,288,357-3,330,152	

References

1. Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, et al. The genome sequence of *Schizosaccharomyces pombe*. Nature 2002;415:871–80. doi:10.1038/nature724.
2. Nakaseko Y, Adachi Y, Funahashi S-i, Niwa O, Yanagida M. Chromosome walking shows a highly homologous repetitive sequence present in all the centromere regions of fission yeast. The EMBO Journal 1986;5:1011–21.
3. Fishel B, Amstutz H, Baum M, Carbon J, Clarke L. Structural organization and functional analysis of centromeric DNA in the fission yeast *Schizosaccharomyces pombe*. Molecular and Cellular Biology 1988;8:754–63.
4. Chatterjee G, Sankaranarayanan SR, Guin K, Thattikota Y, Padmanabhan S, Siddharthan R, Sanyal K. Repeat-Associated Fission Yeast-Like Regional Centromeres in the Ascomycetous Budding Yeast *Candida tropicalis*. PLoS Genet 2016;12:e1005839. doi:10.1371/journal.pgen.1005839.
5. Coughlan AY, Hanson SJ, Byrne KP, Wolfe KH. Centromeres of the Yeast *Komagataella phaffii* (*Pichia pastoris*) Have a Simple Inverted-Repeat Structure. Genome Biology and Evolution 2016;8:2482–92. doi:10.1093/gbe/evw178.
6. Smith KM, Galazka JM, Phatale PA, Connolly LR, Freitag M. Centromeres of filamentous fungi. Chromosome Research 2012;20:635–56. doi:10.1007/s10577-012-9290-3.

7. King R, Urban M, Hammond-Kosack MCU, Hassani-Pak K, Hammond-Kosack KE. The completed genome sequence of the pathogenic ascomycete fungus *Fusarium graminearum*. BMC Genomics 2015;16:1–21. doi:10.1186/s12864-015-1756-1.
8. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human–Mouse Alignments with BLASTZ. Genome Research 2003;13:103–7. doi:10.1101/gr.809403.
9. Harris RS. Improved pairwise alignment of genomic DNA [PhD thesis]: Pennsylvania State University; 2007.

Additional file 5

T. reesei Rut-C30 reassembly (based on genome sequence from *T. reesei* QM6a)

65 scaffolds from the JGI reference genome (33.3Mb - 99.5% of the genome) have been reassembled in 7 chromosomes, as follows :

CHR I

	scaffold	start	end	direction	comment
7	full	1	1 429 972	-1	
12	full	1	1 022 062	-1	
43	full	1	74 996	-1	
21	full	1	576 034	-1	
55	full	1	33 670	?	centromere (consistent with QM6a)
4	translocation	1	748 277	1	manual correction according to (Vitikainen et al. 2010)
22	translocation	139 515	541 456	1	

size : 4 287 553 (with 100bp Ns spacers between each scaffold)

CHR II

	scaffold	start	end	direction	comment
31	fragment	1	224 034	-1	split location corrected as in QM6a reassembly
41	full	1	80 626	-1	
25	full	1	439 677	1	
68	full	1	10 734	1	
10	full	1	1 156 739	-1	
66	full	1	11 200	?	centromere (consistent with QM6a)
59	full	1	18 517	?	centromere (consistent with QM6a)
8	full	1	1 408 331	1	
34	full	1	166 473	-1	
26	full	1	433 400	1	
14	full	1	861 070	-1	
23	full	1	512 080	-1	
54	full	1	34 758	1?	
36	full	1	136 855	-1	
27	full	1	433 262	-1	
28	fragment	367 024	407 093	?	split location corrected as in QM6a reassembly
67	full	1	11 021	1?	

size : 5 980 447 (with 100bp Ns spacers between each scaffold)

CHR III

	scaffold	start	end	direction	comment
45	full	1	65 952	-1	
69	full	1	10 696	-1?	
35	full	1	152 537	1	
32	full	1	230 370	-1	
11	full	1	1 155 933	-1	
40	full	1	89 857	1	
					centromere location consistent with QM6a (no scaffold realiably assembled)
2	translocation	546 704	2 007 221	-1	manual correction according to (Vitikainen et al. 2010)
4	translocation	748 278	1 204 862	-1	
4	translocation	1 204 863	1 832 615	1	manual correction according to (Vitikainen et al. 2010)
49		1	46 304	?	
48	translocation	1	1 666	1	
2	translocation	154 748	546 703	-1	manual correction according to (Vitikainen et al. 2010)
2	fragment	1	98 434	-1	split location corrected as in QM6a reassembly
6	full	1	1 455 714	-1	

size : 6 245 475 (with 100bp Ns spacers between each scaffold)

CHR IV

	scaffold	start	end	direction	comment
64	full	1	14 482	1	
19	full	1	663 018	-1	
17	full	1	797 352	1	
56	full	1	32 194	?	centromere (consistent with QM6a)
20	full	1	629 213	-1	
1	fragment	2 981 735	3 756 989	1	split location corrected as in QM6a reassembly
5	fragment	1 584 116	1 729 360	1	split location corrected as in QM6a reassembly
13	full	1	891 309	-1	
2	fragment	99 435	154 747	-1	split location corrected as in QM6a reassembly
38	full	1	125 035	-1	
33	full	1	207 997	-1	

size : 4 337 413 (with 100bp Ns spacers between each scaffold)

CHR V

	scaffold	start	end	direction	comment
46	full	1	62 252	1	telomere repeats
30	full	1	247 268	1	
42	full	1	78 584	1	
53	full	1	36 593	-1?	
18	full	1	685 578	1	
61	full	1	15 406	?	centromere (consistent with QM6a)
60	full	1	15 714	?	centromere (consistent with QM6a)
28	fragment	1	366 023	-1	split location corrected as in QM6a reassembly
1	fragment	1	2 471 118	1	split location corrected as in QM6a reassembly

size : 3 979 336 (with 100bp Ns spacers between each scaffold)

CHR VI

	scaffold	start	end	direction	comment
47	full	1	50 543	-1	
15	deletion	1	1 555	1	manual correction according to (Seidl et al. 2008)
15	deletion	86 603	837 556	1	
44	full	1	66 247	-1?	
62	full	1	15 337	?	
1	fragment	2 471 169	2 980 271	-1	split location corrected as in QM6a reassembly
50	full	1	45 663	-1	
37	full	1	132 540	1	
51	full	1	43 169	-1?	centromere (consistent with QM6a)
39	full	1	105 148	1	
9	full	1	1 219 543	-1	
22	translocation	1	139 476	1	manual correction according to (Vitikainen et al. 2010)
48	translocation	1 667	48 367	1	
5	fragment	1	1 583 115	-1	split location corrected as in QM6a reassembly

size : 4 710 394 (with 100bp Ns spacers between each scaffold)

CHR VII

	scaffold	start	end	direction	comment
29	full	1	382 182	-1	
24	full	1	501 049	-1	
16	full	1	824 923	-1	
52	full	1	41 083	?	centromere (consistent with QM6a)
3	full	1	1 910 749	-1	

size : 3 660 386 (with 100bp Ns spacers between each scaffold)

>scaffold_57

	scaffold	start	end	direction	
57	full	1	25 756	1	centromere signature but not reliably assembled

>scaffold_58

	scaffold	start	end	direction	
58	full	1	21 040	1	centromere signature but not reliably assembled

>scaffold_63

	scaffold	start	end	direction
63	full	1	14 539	1

>scaffold_65

	scaffold	start	end	direction
65	full	1	12 580	1

centromere signature but not reliably assembled

>scaffold_70

	scaffold	start	end	direction
70	full	1	8 513	1

>scaffold_71

scaffold		start	end	direction
71	full	1	6 846	1

>scaffold_72

	scaffold	start	end	direction
72	full	1	6 811	1

>scaffold_73

	scaffold	start	end	direction
73	full	1	6 421	1

>scaffold_74

	scaffold	start	end	direction
74	full	1	5 890	1

>scaffold_75

scaffold		start	end	direction
75	full	1	5 683	1

>scaffold_76

	scaffold	start	end	direction
76	full	1	5 459	1

>scaffold_77

	scaffold	start	end	direction
77	full	1	5 371	1

>scaffold_78

	scaffold	start	end	direction
78	full	1	5 154	1

>scaffold_79

	scaffold	start	end	direction
79	full	1	4 691	1

>scaffold_80

	scaffold	start	end	direction
80	full	1	4 619	1

>scaffold_81

	scaffold	start	end	direction
81	full	1	4 614	1

>scaffold_82

	scaffold	start	end	direction

82	full	1	4 370	1

>scaffold_83

	scaffold	start	end	direction
83	full	1	3 796	1

>scaffold_84

	scaffold	start	end	direction
84	full	1	3 468	1

>scaffold_85

	scaffold	start	end	direction
85	full	1	3 089	1

>scaffold_86

	scaffold	start	end	direction
86	full	1	3 000	1

>scaffold_87

	scaffold	start	end	direction
87	full	1	2 158	1

Jourdier *et al. Biotechnol Biofuels (2018) 11:163*
https://doi.org/10.1186/s13068-018-1161-5

Biotechnology for Biofuels

CORRECTION

Open Access

Correction to: Proximity ligation scaffolding and comparison of two *Trichoderma reesei* strains genomes

Etienne Jourdier[1†], Lyam Baudry[2,3†], Dante Poggi-Parodi[1], Yoan Vicq[1], Romain Koszul[2,3], Antoine Margeot[1], Martial Marbouty[2,3*‡] and Frédérique Bidard[1*‡]

Correction to: Biotechnol Biofuels (2017) 10:151
https://doi.org/10.1186/s13068-017-0837-6

Following publication of the original article [1], the authors reported a problem in the drawing of Rut-C30 chromosome III in Fig. 2b of the original article [1]. The two fragments of chromosome I inserted inside chromosome III should be swapped, and the direction of the fragment containing rim101 and cel1a genes should be inverted. This reversed insertion indicates that at least 2 rearrangements occurred simultaneously, so the possible scenario proposed in Fig. 2c of the original article was inaccurate. The corrected Fig. 2 with modified panels b and c is available in this erratum. The detailed description of Rut-C30 assembly in the Additional file 5 of the original article is correct.

The authors also noticed two mistakes in chromosome numbering in the description of these translocations. The correct description is

> "The three translocations resulted finally in the right arm of chromosome I (3′ end of scaffold 48 and main fragment of scaffold 5: 1.63 Mb and 442 genes in total), to be swapped with the right arm of chromosome **VI** (3′ end of scaffold 22: 402 kb and 114 genes). But also in two fragments of chromosome I (one with a fragment of scaffold 4, and the other one

*Correspondence: martial.marbouty@pasteur.fr; frederique.bidard-michelot@ifpen.fr
†Etienne Jourdier and Lyam Baudry contributed equally to this work
‡Martial Marbouty and Frédérique Bidard contributed equally to this work
[1] IFP Energies nouvelles, 1 et 4 Avenue de Bois-Preau, 92852 Rueil-Malmaison, France
[2] Groupe Régulation Spatiale des Génomes, Department Genomes and Genetics, Institut Pasteur, 75015 Paris, France
Full list of author information is available at the end of the article

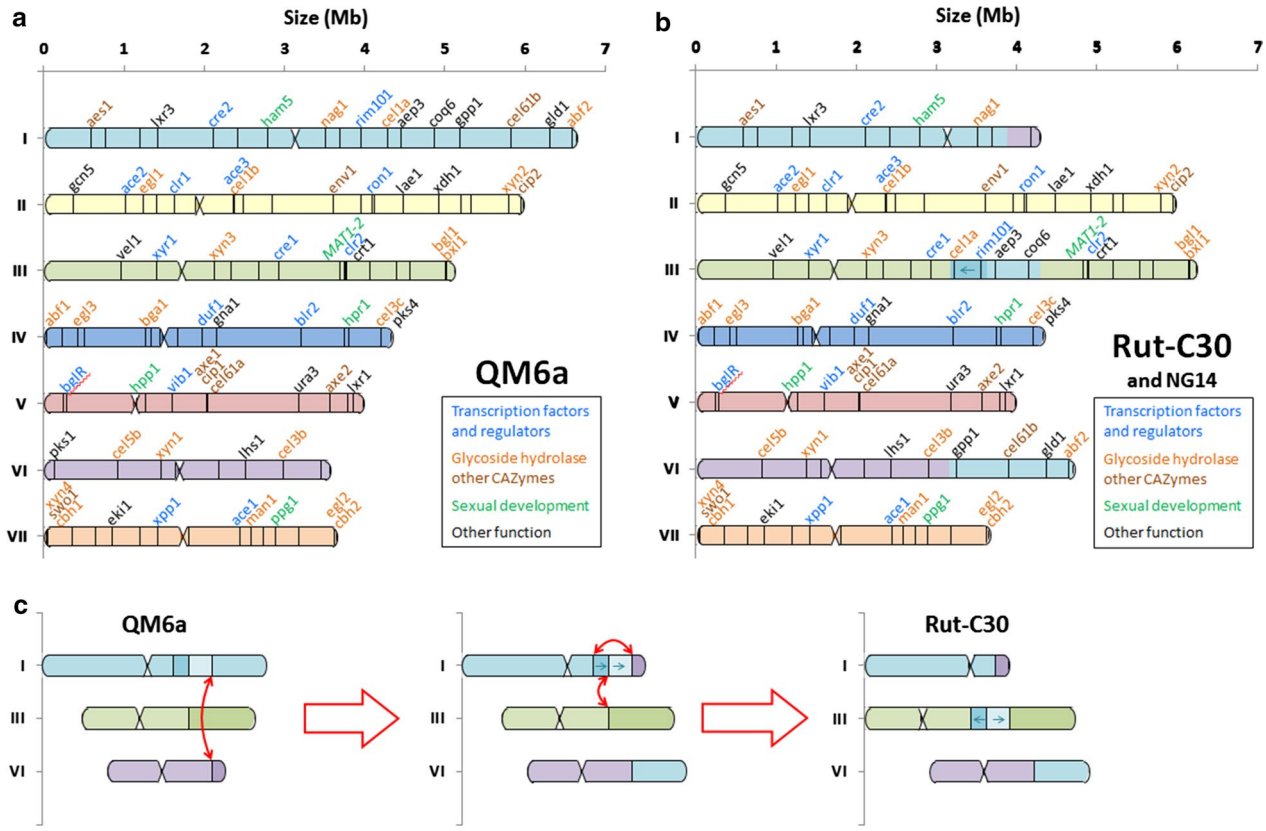


Fig. 2 Chromosome maps of *T. reesei* QM6a and Rut-C30 strains. Chromosome maps of *T. reesei* QM6a (**a**) and Rut-C30 and NG14 (**b**) strains were identified by reassembly of the JGI reference genome using 3C sequencing data for each strain. For Rut-C30 map, the *colors* of chromosome fragments are consistent with their *colors* in QM6a map to clearly show chromosomal rearrangements. Some emblematic genes were chosen along the sequence to be used as location markers (list available in Additional file 6). The Rut-C30 85 kb deletion event on chr. VI is shown by the lack of pks1 gene. Centromere locations are shown by restricted width. **c** Possible scenario (among others) from QM6a to Rut-C30

with another fragment of scaffold 4, scaffold 49 and a small fragment of scaffold 48) to be inserted head to foot in the middle of the chromosome **III** (1.13 Mb and 310 genes in total for both fragments)."

Author details
[1] IFP Energies nouvelles, 1 et 4 Avenue de Bois-Preau, 92852 Rueil-Malmaison, France. [2] Groupe Régulation Spatiale des Génomes, Department Genomes and Genetics, Institut Pasteur, 75015 Paris, France. [3] UMR 3525, CNRS, 75015 Paris, France.

The original article can be found online at https://doi.org/10.1186/s13068-017-0837-6.

Publisher's Note

Received: 29 May 2018 Accepted: 1 June 2018
Published online: 13 June 2018

Reference

1. Jourdier E, Baudry L, Poggi-Parodi D, Vicq Y, Koszul R, Margeot A, Marbouty M, Bidard F. Proximity ligation scaffolding and comparison of two *Trichoderma reesei* strains genomes. Biotechnol Biofuels. 2017;10:151. https://doi.org/10.1186/s13068-017-0837-6.

3.2.2 Joint assembly of two *Cataglyphis hispanica* lineages reveals chromosome fusion

In this section we discuss the scaffolding project of the desert ant *Cataglyphis hispanica* and the rearrangements it unveiled.

3.2.2.1 Overview of *Cataglyphis hispanica*

Nearly all ants species (and many other species from the order *Hymenoptera* such as bees or wasps) are eusocial and live in large colonies where the reproductive function is monopolized by one or a select few females called *queens*. Sex determination is unique in Hymenoptera in that it is determined by ploidy: females are diploid while males are haploid. Almost all individuals within a colony are female; the non-reproducing ones are called *workers*, while males, whose lifespan is usually limited to a reproductive season and die off soon after mating, are called *drones* [350]. Many species are *polymorphic*, with distinct morphological differences between workers, queens and drones, as figure 51 shows. Reproductive individuals often have wings, in which case they are called *alates*. This is a only broad descriptive outline, as ants as a taxonomic group are extremely diverse in behavior, societal organization, and reproductive strategies [351].

Figure 51: Polymorphism in *C. hispanica* individuals: a winged male (left), a queen surrounded by workers (center) and a worker (right).

Source: Taken with permission from Hugo Darras at `https://www.flickr.com/people/fourmis/`

The *Cataglyphis* genus is especially notable for its diversity in breeding systems, including *hybridogenesis*. In this system, females and males from close species or lineages reproduce, but the males' chromosomes are discarded in germinal cells. Males only transmit their genetic material on the somatic level. A unique variant of this system, called *social hybridogenesis*, was observed in *Cataglyphis* species, among which features the desert ant *C. hispanica* [352]. It is illustrated in figure 52.

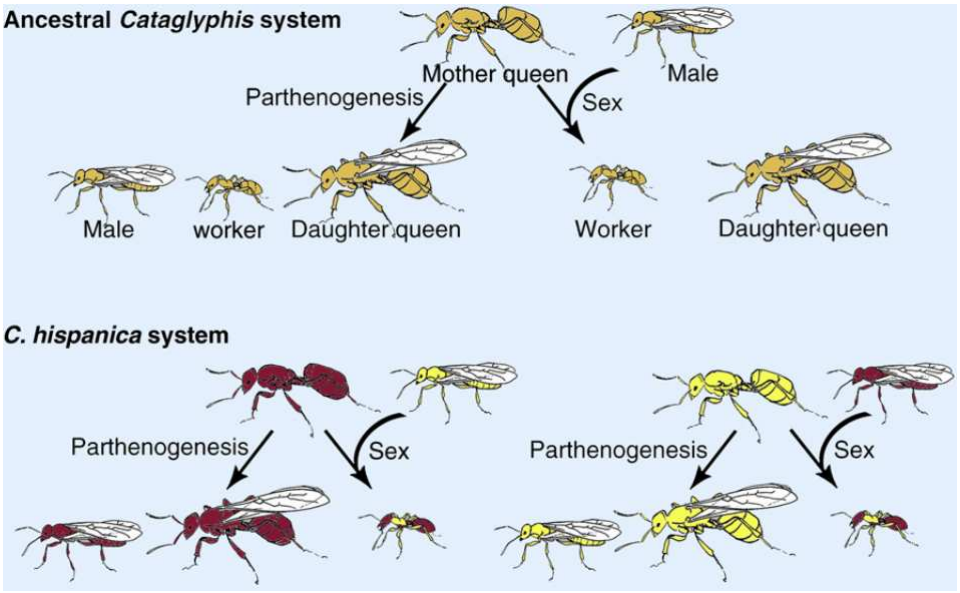


Figure 52: Social hybridogenesis in *C. hispanica*.

Source: Adapted from *Darras et al., 2012*, [352].

In this species, individuals are split into two lineages, and workers are produced from the interbreeding between a male and a female from either lineage. Sexual males and females themselves are "pure-breeds", stemming from asexual reproduction though parthenogenesis and only bearing maternal genetic material. Social hybridogenesis can be therefore thought of as a "generalization" of standard hybridogenesis to the caste level (as opposed to the somatic/germinal distinction within a single individual).

This strategy is relatively rare, only documented in a few other *Cataglyphis* species and taxa, such as *Solenopsis* fire ants and *Pogonomyrmex* seed harvester ants [353]. It is however remarkably consistent, as figure 53 shows that *C. hispanica* is widespread over Spain and Portugal and this behavior was observed among all such colonies [354]. Moreover, the exact mechanisms are not well known at the genomic level. Part of this lack of knowledge stems from the absence of a chromosome-level genome, so that any structural dynamics underlying such mechanisms would go unnoticed.

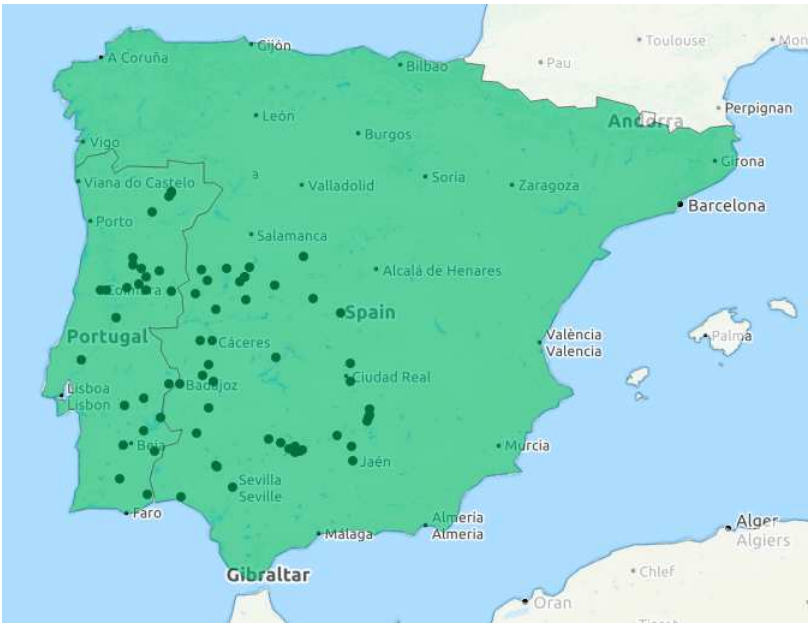


Figure 53: Geographical habitat of *C. hispanica*.

Source: Adapted from `antmaps.org`.

Here we present preliminary results in our attempt to solve the question. They can be articulated into three steps:

- Obtain high-quality chromosome-level assemblies for both lineages of *C. hispanica*, complete with extensive validation and metrics assessment
- Investigate potential large-scale chromosome rearrangements between both lineages
- Match this newly acquired structural data to functional annotations so as to form a comprehensive picture of the mechanisms behind social hybridogenesis

We have decisive results on the first two steps and work is still ongoing for the third.

3.2.2.2 Joint Hi-C based scaffolding

Preliminary assembly We first set out to assemble the genomes of two individual queens, one for either lineage (subsequently referred to as lineages 1 and 2). The strategy we used includes both Hi-C and long reads; a complete workflow is illustrated in figure 54. The following preliminary steps had been performed prior to our Hi-C based work:

- A set of nanopore reads assembled with the Flye, a long read based assembler. Flye includes a polishing step, whereby long reads are aligned onto the final assembly to correct errors. It is an iterative process, as the reads can be aligned again onto

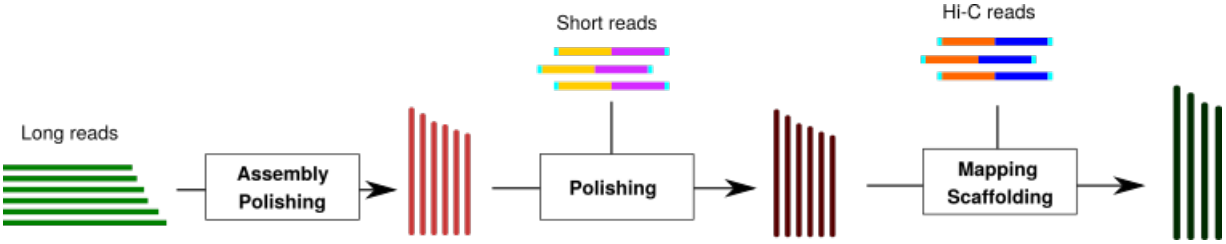


Figure 54: *C. hispanica* reassembly workflow.

the updated assembly for further corrections until no visible increase in quality is observed. Six rounds of this polishing were performed.

- A set of Illumina short-reads was used for additional polishing on the long read assembly with the help of Pilon. Eight rounds of polishing were perfomed onto the first lineage assembly, and five rounds onto the second lineage.

This assembly (subsequently referred to as the *nanopore* or *hybrid assembly*) served as the reference for comparative and validation purposes. In addition, a previous assembly solely based on short read data was available (subsequently referred to as the *short read* or *Illummina assembly*.

Hi-C mapping and scaffolding We mapped 64,691,140 paired-end reads onto the DpnII restriction fragments of the nanopore assemblies of lineages 1 and 2, respectively, using our own Hi-C pipeline [1] with Bowtie 2 in the back-end (using the option `--very-sensitive-local`). Alignments with mapping quality below 30 were discarded. This resulted in 4,417,135 (resp. 2,944,341) Hi-C contacts. We then filtered some fragments out of the contact map distribution prior to binning: fragments below 50 bp were discarded, as well as fragments with coverage below the mean standard deviation. They were kept aside so as not to disrupt the global contact distribution, with the intent of re-integrating them later during the polishing step. Then, each contact map was recursively sum-pooled fragment-wise five times (as described in section 1.4.4.3) so that each bin comprised $3^5 = 243$ times.

We then ran instaGRAAL for 100 cycles on both lineages. We reasoned that the initial nanopore assembly structure was a good starting point and didn't split it prior to reassembly. This resulted in 26 (resp. 27) main scaffolds above 1 Mb. We then polished each assembly using instaGRAAL's own implementation as explained in the methods of section 3.1. Briefly, the internal structure of contigs is reconstructed internally so as to correct artifact inversions or relocations within each newly formed scaffold. Lastly, we manually corrected all remaining discrepancies with the reference that weren't due to mapping issues (*i.e.* presumably false breakpoints).The cumulative length of each newly formed scaffolding (compared to the reference) is shown in figure 55.

[1] HiC-Box, available at `https://github.com/koszullab/HiC-Box`

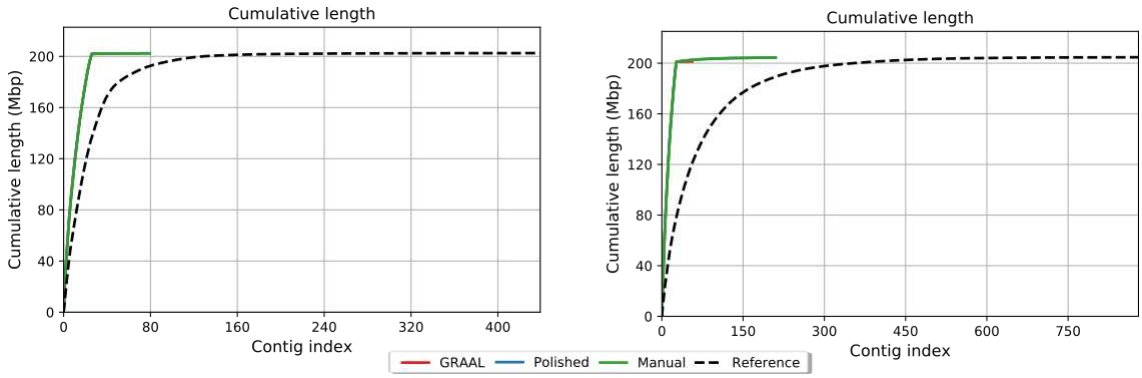


Figure 55: Cumulative lengths of *C. hispanica* lineage assemblies (lineage 1 on the left, lineage 2 on the right): the raw GRAAL assembly, the instaGRAAL polished assembly and the instaGRAAL polished assembly with manual corrections are compared with respect to the hybrid/long-read assembly as reference.

We then re-mapped the Hi-C reads onto the final assembly, after the manual corrections. The final contact maps before and after reassembly are shown in figure 56.

3.2.2.3 Investigating rearrangements

A striking feature arising from a comparison of both lineages is the apparent fusion of two chromosomes. This is shown in a similarity dotplot between both genomes (figure 57).

Overall, and except for the fusion, the dotplot shows a rough one-to-one mapping between chromosomes from one lineage to the other. Smaller-scale rearrangements are less clear and could be due to artifacts. Notably, regions with a lot of disorderly arranged small sequences are repeated (or otherwise homologous) stretches typically found in telomeric regions. Additional polishing could be needed to properly resolve them.

On the other hand, the fusion is confirmed by comparing the Hi-C contact maps themselves (figure 58). The signal is strong enough that the rearrangement could not have arisen from an instaGRAAL artifact alone. Moreover, subsequent re-runs of the software consistently showed this modification. On the other hand, the presence of extraneous repeated sequences confined in one scaffold is also confirmed in figure 58. Overall, similarity data is consistent with contact data.

These results strongly suggest a physical fusion between chromosomes 5 and 8 of lineage 2, which would become chromosome 1 of lineage 1. However, the presence of artifacts in scaffold 9 of lineage 1 (resp. scaffold 10 of lineage 2) also suggests an extensive validation of all genomes involved is required in order to confirm our assemblies are indeed high-quality and suitable for comparative analysis.

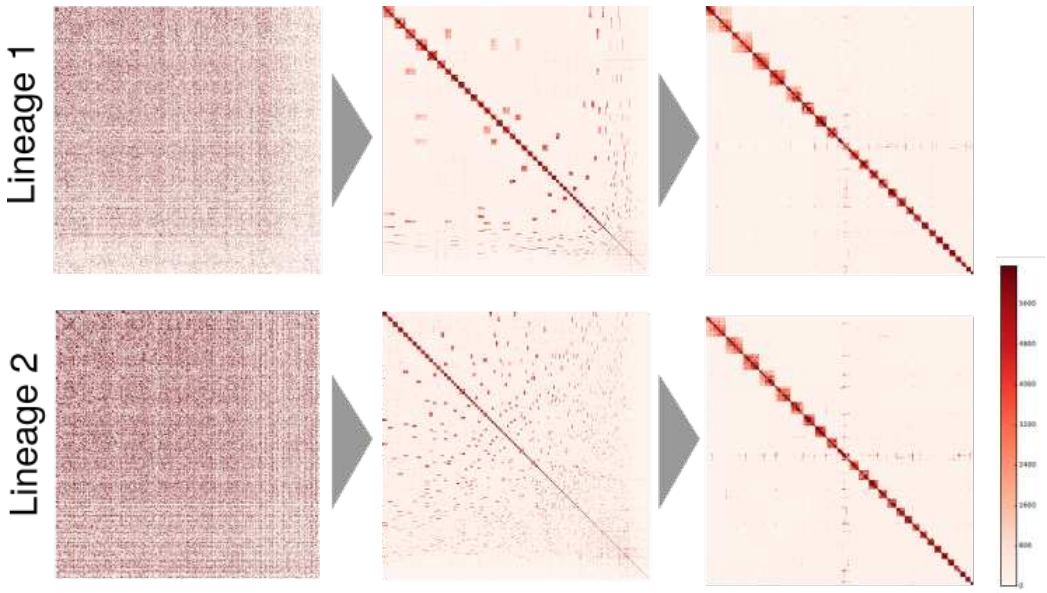


Figure 56: *C. hispanica* contact maps at different stages of the workflow: Illumina short-read based (left), long-read/hybrid based (center) and after instaGRAAL scaffolding and polishing (right).

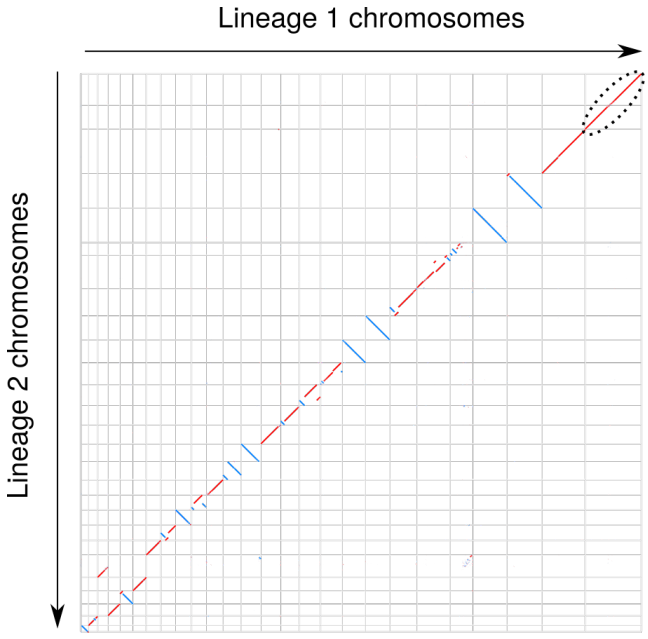


Figure 57: Similarity plots of scaffoldings for each *C. hispanica* lineage. The apparent fusion has been marked on the top right corner.

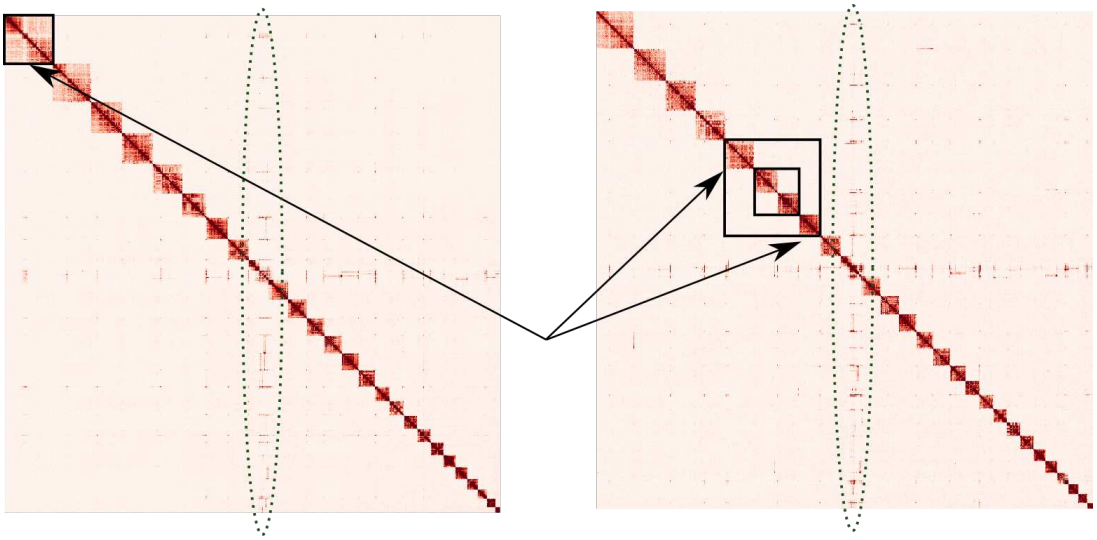


Figure 58: Evidence of chromosome fusion on Hi-C contact maps between both lineages of *C. hispanica*. Merged chromosomes are indicated by arrows, whereas potential artifacts caused by repeated sequences have been marked in green.

3.2.2.4 Genome validation

We performed a comparative validation of all assemblies at our disposal. For each lineage, we had at our disposal an old short read based assembly and a long read/hybrid assembly. We first performed a preliminary scaffolding work with an early version of GRAAL; we later implemented GRAAL polishing and an additional number of features that would transform it into instaGRAAL, and applied it to the assembly. We then corrected by hand all discrepancies between the instaGRAAL scaffolding and the long read/hybrid assembly. In summary, we did a comparative analysis on the following:

- The preliminary short-read assemblies
- The long read/hybrid assemblies
- The raw GRAAL assembly, without polishing
- The instaGRAAL assembly, with polishing
- The instaGRAAL assembly, with polishing and manual error correction

The main metrics to be assessed were Nx (and related), discrepancy with the long read assemblies and BUSCO completeness (using a database of $n = 4415$ orthologs from ODBv9), as shown in table 1. We also considered (and included) k-mer completeness.

Overall, we observed in each case a tenfold improvement in N50. As cumulative plots in figure 55 have shown, more than 99% of each genome's lineage was successfully reintegrated into the main 26 (resp. 27) scaffolds. These 26/27 chromosome counts were

	Short-read	Long-read	GRAAL	instaGRAAL	Manual
Total length (Mb)	211.6	202.5	202.2	202.5	202.5
Contig/scaffold count	6039	439	223	224	222
NG50 (Mb)	0.24	4.4	8.25	8.34	8.18
NG75 (Mb)	0.104	2.5	5.87	5.95	5.98
L50	245	17	9	9	9
Misassemblies/discrepancies	N/A	0	537	22	0
K-mer completeness (%)	100	100	99.79	99.92	99.92
BUSCO completeness (%)	94.5	98.1	97.7	97.9	97.9

Table 1: Assembly metrics for lineage 1 genomes. Discrepancies are accounted with respect to the long-read based reference assembly.

	Short-read	Long-read	GRAAL	instaGRAAL	Manual
Total length (Mb)	221.3	204.7	201.6	204.7	204.7
Contig/scaffold count	14930	880	304	392	392
NG50 (Mb)	0.22	1.44	8.15	8.18	8.1
NG75 (Mb)	0.079	0.609	5.64	5.54	5.56
L50	264	42	10	10	10
Misassemblies/discrepancies	N/A	0	688	19	1
K-mer completeness (%)	100	100	98.48	99.89	99.89
BUSCO completeness (%)	94.1	97.3	97.3	97.6	97.6

Table 2: Assembly metrics for lineage 2 genomes. Discrepancies are accounted with respect to the long-read based reference assembly.

observed in repeated control instaGRAAL runs, strongly indicating that they are the actual number of chromosomes for each lineage. This overall shows instaGRAAl was successful at scaffolding the genome of either lineage.

Because we trusted the initial long-read based assembly enough not to induce artifact breakpoints with our scaffolding, we performed extensive polishing. This decreased discrepancies by an order of magnitude in each lineage (537 to 22, resp. 688 to 19). After manually reviewing each of the remaining discrepancies, we set out to correct them by hand if they were not due to mapping issues, thus resulting in the 0 (resp. 1) discrepancies in the final assembly. Corresponding Nx metrics show these modifications had little, if any, impact on the global scaffolding structure.

Lastly, completeness metrics (k-mer and BUSCO orthologs) were found to be overall satisfactory or otherwise unchanged. In lineage 1, the GRAAL scaffolding induces a slight loss in BUSCO completeness, presumably due to artifact breakpoints. Polishing (automatic or manual) alleviates this somewhat, although a 0.2% decrease (9 orthologs) is still observed with respect to the initial long-read assembly. In lineage 2, the scaffolding did not alter the initial completeness, and polishing actually improved it. This stresses the importance of re-injecting initial data into one's scaffolding when it is considered trustworthy enough. Notably, all assemblies were markedly more complete than the Illumina short-read based one, thereby validating our entire workflow based on long reads and Hi-C.

3.2.2.5 Ongoing work

In this section we have presented our main results on the joint study of *C. hispanica* lineages:

- We have successfully obtained high quality, chromosome-level assemblies for both lineages, complete with extensive validations
- We have very strong evidence from Hi-C data that two chromosomes in one lineage have been merged in the other.

Current work is still ongoing for the annotation of either genome: we need to link the structural genomics results to functional data so as to identify the genes of interest that could be responsible for the mechanisms underlying the social hybridogenesis.

Our genomes, although the best available quality for *C. hispanica*, still contain errors and a number of artifacts, notably among repeated sequences that could be potentially misplaced: additional data is necessary to further polish the assemblies. Lastly, additional verifications on a cytological level could be needed to confirm the chromosome counts inferred from Hi-C data.

4 Metagenome assembly and network dynamics

In the introduction we have underlined the potential of a 3C based approach for metagenome binning and assembly as a necessary step for understanding genome dynamics among complex communities. In this section we present our main results, published and submitted respectively, showcasing our approach:

- We first demonstrate the effectiveness of the meta3C framework with the first 3C experiment on an *in vivo* sample; the method builds upon the proof-of-concept works detailed earlier as well as 3C-based scaffolding. We successfully scaffold more than a hundred bacterial genomes, identify features of interest within the genomes, and isolate phage-host relationships.
- Then, we implement the design we have built into a full-fledged pipeline, dubbed metaTOR. We benchmark it against state-of-the-art traditional methods and prove it outperforms all of them in terms of completeness and contamination. The successful application of this pipeline yields again more than a hundred bacterial genomes.

4.1 Scaffolding bacterial genomes and probing host-phage interactions

MICROBIOLOGY

Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay

Martial Marbouty,[1,2] Lyam Baudry,[1,2] Axel Cournac,[1,2] Romain Koszul[1,2]*

The biochemical activities of microbial communities, or microbiomes, are essential parts of environmental and animal ecosystems. The dynamics, balance, and effects of these communities are strongly influenced by phages present in the population. Being able to characterize bacterium-phage relationships is therefore essential to investigate these ecosystems to the full extent of their complexity. However, this task is currently limited by (i) the ability to characterize complete bacterial and viral genomes from a complex mix of species and (ii) the difficulty to assign phage sequences to their bacterial hosts. We show that both limitations can be circumvented using meta3C, an experimental and computational approach that exploits the physical contacts between DNA molecules to infer their proximity. In a single experiment, dozens of bacterial and phage genomes present in a complex mouse gut microbiota were assembled and scaffolded de novo. The phage genomes were then assigned to their putative bacterial hosts according to the physical contacts between the different DNA molecules, opening new perspectives for a comprehensive picture of the genomic structure of the gut flora. Therefore, this work holds far-reaching implications for human health studies aiming to bridge the virome to the microbiome.

INTRODUCTION

High-throughput DNA sequencing technologies developed over the past decade have set a milestone for the analysis of microbial communities in natural environments. Metagenomic approaches provide an overview of the diversity of DNA or RNA molecules directly isolated from natural mixes of species (*1*–*4*). Large-scale exploratory studies have revealed that complex communities are ubiquitous in all environments (*5*, *6*), where they hold diverse and important roles, including contributions to animal and plant metabolisms (*7*–*10*). These developments have greatly accelerated the discovery of new bacteria (*3*, *4*, *11*–*14*), plasmids (*15*, *16*), and virus/phages (*17*–*20*). However, some limitations persist despite constant technological improvements. Notably, the difficulty to assemble complete genomes and full episome sequences (*21*) and the inability to characterize the interactions between those different molecules impair the full resolution of the genomic structure of these populations. For instance, bacteria-phage relationships remain poorly characterized, despite the impact of phages on the balance of microbial communities (*22*, *23*). The presence of phages, which are considered the most abundant and diverse biological entities on earth (*24*), in these ecosystems, has far-reaching consequences beyond particular pairwise interactions (*25*), influencing everything from bacterial virulence (*26*) to cell physiology (*27*). However, the characterization of a phage genome from sequencing data is usually not sufficient to identify its bacterial host(s). As a result, understanding the interplay between phages and the overall microbial community remains limited or out of reach (*28*). Therefore, new approaches alleviating these limitations are needed to better understand phage-bacteria relationships in complex ecosystems (*29*).

One way to address this challenge is to exploit the physical collisions experienced by DNA segments along one and/or between multiple DNA molecules. The frequencies of cis contacts between pairs of loci within a chromosome are higher than the trans contacts between segments located in different chromosomes. These contacts generate a predictive three-dimensional (3D) signature that can be exploited to improve chromosome scaffolding (*21*, *30*, *31*). Recent studies suggest that metagenomic analyses could also benefit from these approaches (*32*–*35*). A blind clustering analysis of the contacts experienced by DNA molecules isolated from controlled or seminatural mixes of microorganisms showed that most contacts involve pairs of DNA regions coming from the same genome (*34*). These contacts were quantified using meta3C (*34*), a derivative of the chromosome conformation capture method (3C; Materials and Methods) (*36*). Briefly, DNA molecules within a mixture of microbial species are frozen in space with a cross-linking agent. The DNA trapped within cross-linked protein complexes is then digested with a restriction enzyme. The resulting restriction fragments (RFs) are then religated together. Ligation events will mostly involve RFs that were in close vicinity in space before the fixation step and, therefore, that were very likely to share the same cell compartment. The quantification of these events is done using paired-end (PE) sequencing. Meta3C reads can be used to perform a de novo assembly that will generate contigs reflecting the genetic content of the community, as well as the clustering and scaffolding steps that will provide a glimpse of the genomic structure of the population [reviewed by Marbouty and Koszul (*35*)]. Fortuitous hints have suggested that chromosomes and other kinds of DNA molecules, such as plasmids (*34*), could be identified from the meta3C data and assigned to their host cells. However, no large-scale exploration of the genomic structure of a truly natural complex community had been undertaken so far using this approach.

Here, we investigated the ability of meta3C to bring new insights into the genomic structure of a natural and complex mammalian gut microbiota, including its phage-host interactions. Starting with a single, unknown natural complex microbial ecosystem, a computational workflow was designed to allow the de novo assembly and scaffolding of dozens of bacterial genome scaffolds. Moreover, the pipeline also leads to the assembly of large bacteriophage sequences, including a large genome phylogenetically close to the phiKZ phage family (*37*, *38*) and never fully characterized before in the mammalian gut (*39*). Finally, these phage sequences were assigned to bacterial chromosome scaffold(s) based on their physical contact frequencies, providing information

[1]Institut Pasteur, Department Genomes and Genetics, Groupe Régulation Spatiale des Génomes, 75015 Paris, France. [2]CNRS, UMR 3525, 75015 Paris, France.
*Corresponding author. Email: romain.koszul@pasteur.fr

about the coexistence of bacteriophages within multiple species and/or strains. These results show that DNA collisions allow the tracking of mobile genetic elements of interest within complex microbial populations, opening the way to high-resolution monitoring of horizontal transfer events within populations and dynamic studies of microbiota genomic structure.

RESULTS AND DISCUSSION

Applying meta3C to a mice gut microbiota

To investigate the genomic structure of the mouse gut microbiome, a single feces sample from a healthy control male mouse (C57BL/6) from the Institut Pasteur animal facility was collected, split, and processed through two meta3C protocols that solely differed by the restriction enzyme being used: either Hpa II [C′CGG] or Mlu CI [′AATT] (Fig. 1A and Materials and Methods). As discussed before (*34*, *35*), using enzymes differing in the GC content of the corresponding restriction sites (RSs) is expected to improve contact coverage for GC- and AT-rich genomes. The Hpa II and Mlu CI libraries were sequenced on an Illumina NextSeq machine [2 × 75 base pairs (bp)], with 114 and 71 million PE reads recovered, respectively. Reads from both libraries were pooled and assembled into contigs using the IDBA-UD program (*40*), resulting in 374,363 contigs (cumulated size, 580 Mb; N50, 3783 bp; maximum size, 490 kb; mean size, 1402 bp). Translated coding sequences resulting from this assembly [594,648 open reading frames (ORFs) detected—MetaGeneMark; (*41*)] were analyzed at the taxon and functional levels using the metagenomics RAST (MG-RAST) pipeline (Materials and Methods) (*42*). As expected from a gut metagenome, the major clades in the sample were Firmicutes (70%) and Bacteroidetes (15%) (Fig. 1B) (*43*). An analysis of DNA sequences using the Kraken program (*44*) (Materials and Methods) confirmed these results with, also as expected, ~80% of the sequences not attributed to a specific genome (*43*). Coding sequences were then annotated for essential genes, phages, and conjugative elements using repository databases (*14*, *45*, *46*), and the annotated contigs were then split into 1-kb fragments. This step has two objectives: first, to limit the impact of misassembly errors (such as chimeric contigs) arising during the assembly step, and second, to normalize the contact signal with respect to the influence of contig size on their representation during the segmentation of the network. Contigs under 500 bp were discarded, leading to a global set of 553,310 contigs (513 Mb total). An internal control for the network segmentation step was implemented by introducing meta3C reads of a chosen mix of three bacterial species (Materials and Methods and fig. S1A), resulting in a final set of 569,146 contigs (526 Mb total). The contact network was then generated by aligning meta3C PE reads against the contigs. Whereas in most (75%) instances both reads of a pair mapped within the same contig, in 46 million instances each read of a pair aligned along a different contig, resulting in a pair of contigs bridged by at least one contact. Contact frequencies between contigs were then normalized by the read coverage of the contigs (Materials and Methods), resulting in a large network of 569,146 nodes and 20,557,427 weighted edges. Contigs showing enriched contacts are likely to correspond to DNA molecules sharing the same cell compartment (*34*).

Iterative segmentation of the meta3C contact map into core communities

The global network was then segmented into communities (in a network analysis sense) using the Louvain clustering algorithm (Materials and Methods) (*47*). A community is a subnetwork, or partition, of contigs having enriched contacts with each other as opposed to the other contigs. After one clustering step, 515 Mb (>98%) of the total DNA was spread among 93 communities ranging in size from 500 to 61,000 contigs. By design, the Louvain algorithm cannot attribute a node to multiple communities and is nondeterministic: When a segmentation is performed twice on the same network, some nodes will be assigned to distinct partitions if these communities share the elements represented by these nodes. We reasoned that this inherent property could be exploited to strengthen the analysis and identify DNA sequences shared by two or more large communities. To do so, we performed the segmentation independently 100 times, delineating core communities (CCs) made of contigs that systematically cluster together for each of these iterations (Fig. 1C, fig. S2, and Materials and Methods). The distribution of the sizes of CCs recovered after iterative segmentation was computed, showing that the number of CCs larger than 500 contigs (that is, of approximately 500 kb or more) quickly converges toward 124 clusters encompassing ~90% of the total DNA (Fig. 1, C and D, and table S1). The iteration procedure also led to a reduction of the contact background between communities of contigs, suggesting a better resolution of the network (Fig. 1E). The control contigs containing three bacterial species were segmented into three well-defined CCs (black triangles in fig. S1B), confirming that the Louvain iterative procedure conveniently segregates genomes from the meta3C network. The presence of very large CCs containing more than 10,000 contigs nevertheless suggests that some CCs encompass more than one genome of closely related species, potentially due to the presence of numerous shared sequences (below). Finally, the influence of the choice of the restriction enzyme on the contig representation is made clear when the contact map is binned into a fixed number of RFs for each enzyme, illustrating the interest of combining two different restriction enzymes to cover both AT- and GC-rich sequences (fig. S1C).

Characterization of meta3C CCs

To investigate the genetic nature of CCs, we computed gene ontology distribution based on contig annotation for different classes of genetic elements (Fig. 1F). Contigs carrying essential genes (n = 24,896) (*48*) or lineage-specific markers [single genes copy (SGCs), n = 7104] (*49*), all specific of bacterial chromosomes, were predominantly found in the larger CCs. On the other hand, contigs carrying genes related to conjugative elements (n = 4676) (*50*) and phages (n = 4796) (*20*) were significantly enriched in small CCs as opposed to the previous categories. This analysis indicates that large CCs contain contigs of sequences belonging to bacterial chromosomes (table S1) and mobile elements (table S2), whereas small CCs represent mostly independent episomes or mobile elements, such as plasmids and phages (table S3).

Metagenomic data are often analyzed in light of covariance analyses of genetic elements over multiple samples (*14*, *43*, *48*). These approaches have led to the characterization of co-abundance groups of genes (CAGs) (*14*, *43*), which are clusters of genes whose sequencing coverage covaries within the samples. Among CAGs, groups containing more than 700 coding sequences have been dubbed metagenomic species (MGS). It was suggested that MGS clusters represent species-specific groups of genes. To compare both approaches, meta3C reads were aligned against the gene catalog of mouse microbiota MGS (*43*). Genes were then clustered, either through their MGS index or through the Louvain iterative procedure, and contact maps of the 100 largest MGS and meta3C CCs were generated (fig. S2). A strong diagonal revealed important contact signal within MGS, confirming that, to a large extent, MGS do group together DNA molecules belonging to the same cellular compartment,

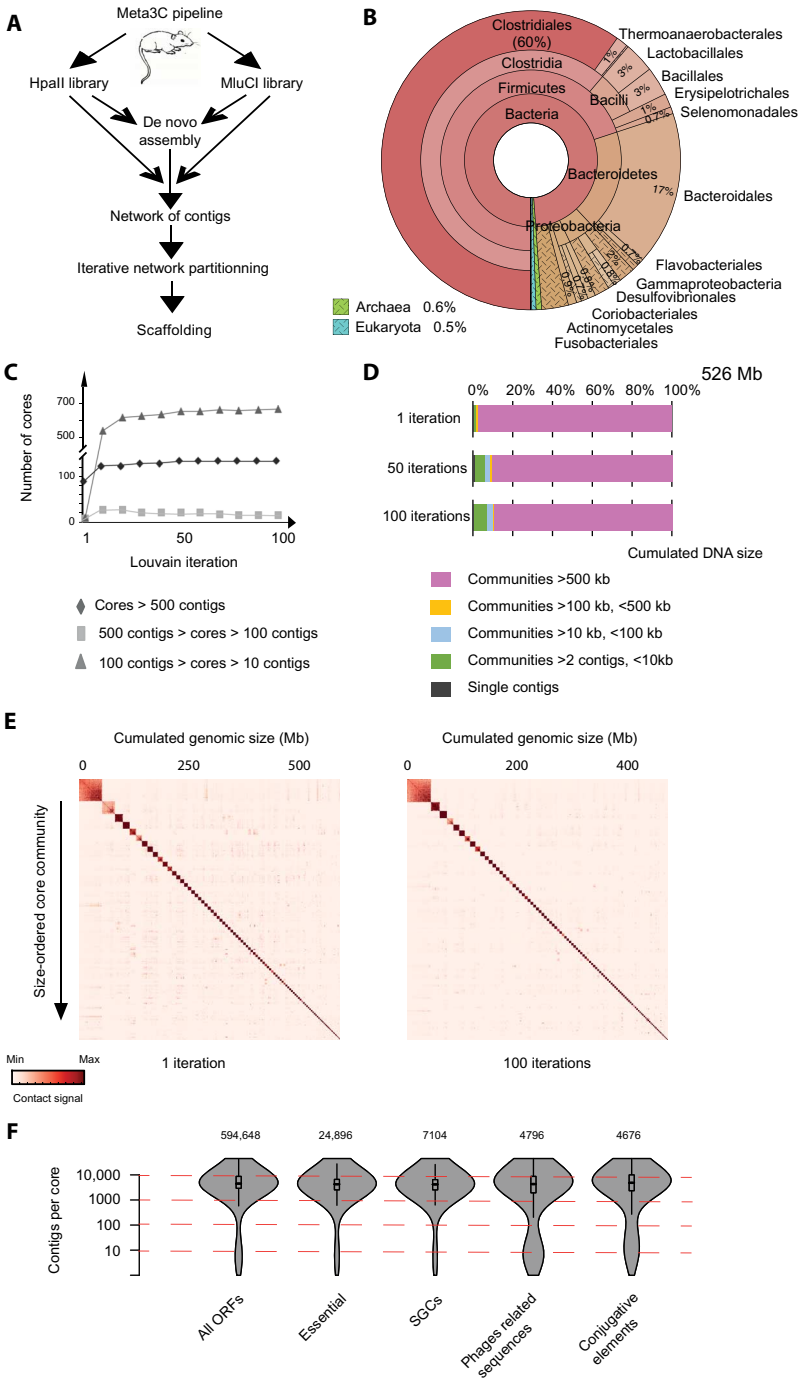


Fig. 1. Meta3C analysis of the mice gut microbiome. (**A**) Flowchart representing the computational analysis steps of a meta3C experiment. First, the reads from two sequenced meta3C libraries are assembled de novo into contigs. The meta3C contact information from both data sets is then used to generate a contact network between all contigs. The Louvain algorithm is then applied iteratively to segment the global network into CCs. (**B**) MG-RAST taxonomy analysis of the contigs generated from the de novo assembly step. (**C**) Evolution of the distribution of CC sizes over 100 Louvain iterations (*x* axis). Triangles, CCs with 10 to 99 contigs; squares, CCs with 100 to 499 contigs; diamonds, CCs with 500 contigs or more. (**D**) Stacked bar chart of the distribution of CC sizes for 1, 50, or 100 Louvain iterations. Categories of CCs are indicated under the histograms. (**E**) Contact maps of the 100 largest CCs recovered after a single and 100 Louvain iterations (1 vector = 200 kb). The *x* and *y* axes are labeled with the cumulated DNA size and the index of the community, respectively. (**F**) Vioplot of different functional contig annotations as a function of their CC size (in number of contigs) (*y* axis = log scale). The number of annotated elements is indicated for each category.

thus the same species. This map also immediately pointed at MGS exhibiting potent physical contacts with each other, strongly suggesting that these groups of sequences share at least one cellular compartment in the population and hence belong to the same species. On the other hand, meta3C CCs hardly exhibit any contacts between each other, as expected if these CCs correspond to phased genomic sequences of discrete species. A comparison of both methods reveals that around half of the genes present in a given MGS are found in a CC, a difference that may result from

the fact that MGS are computed over more than a hundred samples, whereas CCs are generated with a single sample. Therefore, the two approaches complement each other for well-studied ecosystems for which many samples are already available.

De novo assembly and scaffolding of bacterial genomes

The content of large CCs was then investigated qualitatively. First, contigs from each of the 121 CCs encompassing more than 500 contigs (excluding the 3 control CCs) were used as an index to align all raw meta3C reads using Bowtie2 (mapping parameters: -local -sensitive, ambiguous matches allowed -parameters a-; Fig. 2A and Materials and Methods). When at least one member of a read pair mapped onto one of these contigs, both sequences were retained. All PE reads with a good quality score (Materials and Methods) recovered for each CC were then assembled de novo with IDBA-UD to generate a new set of contigs (no precorrection option, default parameters). For each final assembly above 500 kb, all contigs above 500 bp were retrieved. The quality of bacterial genome assemblies can be assessed by looking for the presence of a standardized set of marker genes (*51*). The pools of contigs generated for each CC were therefore screened using the CheckM pipeline for these markers (*49*). Most assemblies had a marker gene content typical of what is expected from a single bacterial genome, although some of the largest communities contained multiple copies of marker genes, suggesting that they contained more than one genome (see below).

The contigs from each of the 121 CCs were then scaffolded using the program GRAAL (Fig. 2B) (*30*). Briefly, GRAAL exploits contacts between DNA regions to assess for their colinearity. The program progresses by successive iterations to converge toward the 1D genome structure that best accounts for the 3D data. For instance, the 3264 contigs present in CC #63 were reordered by GRAAL into a large, 3.2-Mb scaffold (Fig. 2B). These scaffolds can then be compared to chromosomal contact maps of single species, which have been described before and are schematically represented in Fig. 2C (*34*, *52*). These maps display typical patterns. First, a main diagonal reflects enriched local contacts all along the chromosome, a consequence of neighboring DNA regions interacting more often together than distant ones. Second, a strong signal in each corner of the map indicates a circular chromosome (pink arrowheads in Fig. 2C). Finally, secondary features that are specific to bacterial chromosome metabolism are also sometimes visible, notably a secondary diagonal (Fig. 2C) (*53*). This feature reflects the cohesion of replichores initiated at the origin of replication and has been described in *Caulobacter crescentus* (*52*), *Bacillus subtilis* (*53*), and *Vibrio cholerae* (*54*). It is present in other species as well but not in *Escherichia coli* (*35*). GRAAL was run for 100 iterations on each newly assembled CC (Fig. 2, D and E; fig. S4; table S1; and Materials and Methods). Two-thirds (80) of the 121 assemblies resulted in a marked increase in the N50 of the sequences present in the corresponding CC, with the generation of one (or more) large, megabase-scaled scaffold(s). The resulting contact maps of these large scaffolds were inspected for any potential remaining inconsistencies left out by the probabilistic nature of GRAAL's algorithm (fig. S5 and Materials and Methods). The features displayed by these contact maps were often highly consistent with published contact maps of bacterial genomes. Notably, the continuous main diagonal and the presence of a circularization signal suggest that no large DNA regions are missing in many of the scaffolds. In addition, a secondary diagonal was often present on some of the maps (Fig. 2, i, ii, iv, and v; see also fig. S4). Finally, *dnaA* homologs were often identified at the crossing between this secondary and the main diagonal (Fig. 2, i, ii, iv, and v). *dnaA* is found at the origin of replication (*ori*) in most bacteria, and its presence at the edge of the secondary diagonal is highly consistent with recent analyses describing the role of the replication origin during the cell cycle of *B. subtilis* in chromosome folding (*53*, *55*). Moreover, the position of these putative *ori* sites correlates with the highest coverage in PE reads of the scaffold (Fig. 2, i, ii, iv, and v), suggesting that this procedure also allows one to infer the growth status of these species. The recovered scaffolds and the assembled large CCs were again analyzed through the CheckM pipeline (table S1), revealing a clear improvement in the quality, with respect to both completeness and contamination level, of the recovered genomes. For instance, each of the two large scaffolds retrieved after processing CC #6 (Fig. 2E, v) shows a nearly complete bacterial gene catalog, pointing at the presence of two individual genomes belonging to the same clade. The global conservation of gene order between these two scaffolds (Fig. 2F) suggests that these two species are closely related, and therefore highlights the potential of the meta3C approach [see also CC #22 for another example; Fig. 2, E (vi) and F].

Annotation and analysis of prophages in bacterial genomes

The annotation of the large scaffolds, using the Phaster pipeline (*56*), also pointed at the presence of putative prophage sequences integrated within bacterial genome scaffolds (Fig. 2, D and E, small red and green rectangles on the right side of all matrices). Here, again, our recent work (*53*) proved convenient to interpret the corresponding contact maps (Fig. 3A). The prophages present within the *B. subtilis* genome appear in the contact map as discrete regions with peculiar contact patterns [fig. S6A; see Marbouty *et al.* (*53*) for discussion]. The SPβ prophage sequence is particularly apparent in the contact maps of exponentially growing cells. This prophage appears to get activated upon exposition of the cells to the rifampicin drug, as revealed by the increase in read coverage of the phage genome, resulting in a strong increase in 3C contact signal (fig. S6B) (*53*). In addition, enriched contacts between the extremities of the phage genome were also characterized, suggesting a possible circular form. The phage sequences encompassed within the genomic scaffolds retrieved after GRAAL processing display contact patterns reminiscent of these observations (see, for instance, CC #25; Fig. 3A). This observation suggests that the contact map patterns could be exploited to refine predictions from the Phaster pipeline and to help in the characterization of prophage sequences. For CC #25, the contact pattern and read coverage of the two prophage loci are consistent with a silent pattern (fig. S6A). On the contrary, one of the two scaffolds retrieved from CC #6 (Fig. 2E, vi; scaffold 2 in red) exhibits a peculiar locus, isolated from the rest of the scaffold, more covered and annotated as an incomplete prophage. The contact pattern and read coverage of this region are consistent with an active phage similar to *B. subtilis* SPβ in the presence of rifampicin (fig. S6B). More analyses will be needed to further validate the presence and activity of these phages in these bacteria, but this analysis nevertheless suggests the meta3C data point at silent and active prophages among complex communities. However, one must note that it remains unclear whether the approach has the ability to trap phage genomes present in phage particles outside the bacteria cellular compartment or if it traps virulent phages infecting and killing bacteria in a short amount of time; more experiments will be needed to answer these important questions.

In some instances, the scaffolding step results in multiple scaffolds that do not seem to correspond to large, fully individualized bacterial chromosomes. These scaffolds sometimes display contact patterns

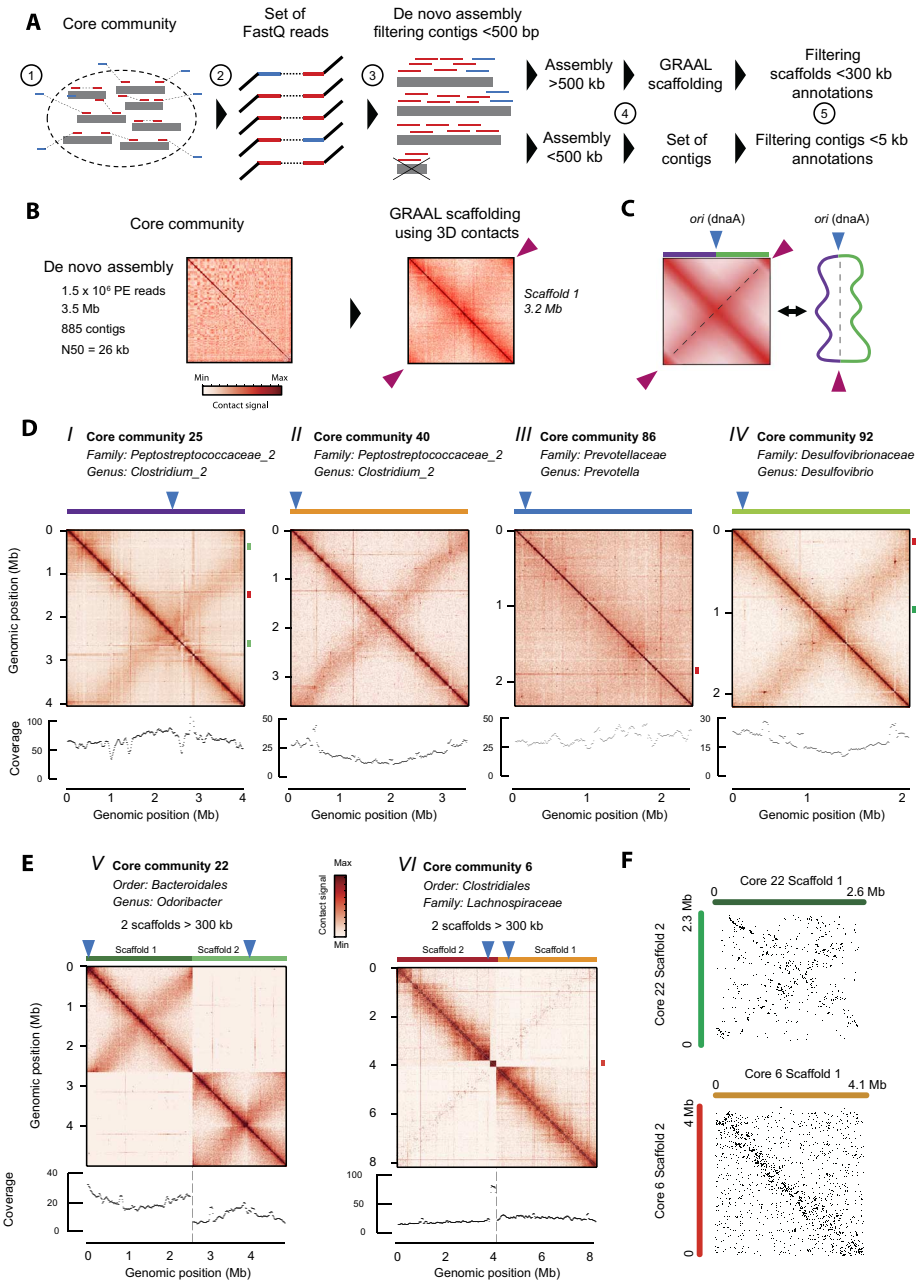


Fig. 2. De novo scaffolding of bacterial genomes from large CCs. (**A**) Pipeline describing the computational processing of CCs. Contigs pooled together within a CC are used to build a genome index (step 1). All PE reads from meta3C libraries are aligned against this index (step 2). If one read of a pair maps onto these contigs, then both reads are retained for the de novo assembly using IDBA-UD (step 3). If the cumulated size of the newly assembled contigs of 5 kb or more reaches at least 500 kb, then they are processed with the GRAAL scaffolding program (step 4). For each CC, the resulting scaffolds and/or contigs are then annotated for taxonomy or the presence of phage sequences (step 5). (**B**) Example of CC63: The 3264 newly assembled contigs [step 2 in (A)] are processed by GRAAL [step 4 in (A)]. Left: Contact map of the newly assembled contigs. Right: Contact map of the 3.2-Mb scaffold obtained after GRAAL processing. Pink triangles point at the circularization signal in the map, consistent with a bacterial circular chromosome. (**C**) Schematic representation of the typical primary and secondary features found on a bacterial contact map (left), alongside a diagram of the corresponding chromosome organization (right). Beside the circularization signal (purple triangles), a secondary diagonal is often found (dotted black lines) as a result of contacts between the left (violet) and right (green) replichores. The secondary diagonal crosses the main diagonal at the origin of replication (blue triangles). (**D**) Contact maps (10-kb bins) of the largest (>500 kb) GRAAL scaffolds retrieved in four CCs, displaying patterns characteristic of bacterial chromosomes [with (i, ii, and iv) or without (iii) a secondary diagonal]. Taxonomic annotation, distribution of read coverage, and position of *dnaA* (blue triangles) are indicated for each scaffold. The read coverage distribution can be used to infer the growth state of the corresponding bacterium. When present, putative prophage loci are represented on the right vertical axis with green (complete prophage) or red (incomplete prophage) rectangles. (**E**) Same analysis as in (D) but for two CCs each containing two large and distinct scaffolds [core 22 (v); core 6 (vi)]. Scaffold 2 from core 6 (vi) exhibits a discrete, more covered (see red rectangle on the coverage distribution) region annotated as an incomplete prophage. (**F**) Comparison of the positions of orthologous genes in the scaffolds obtained in (E). Orthologous genes are displayed as dots based on their position along scaffolds 1 and 2 represented in the *x* and *y* axes, respectively (top, core 22; bottom, core 6). The conservation of synteny between the two scaffolds is apparent from the higher density of orthologous genes (dots) in the diagonal of the graph.

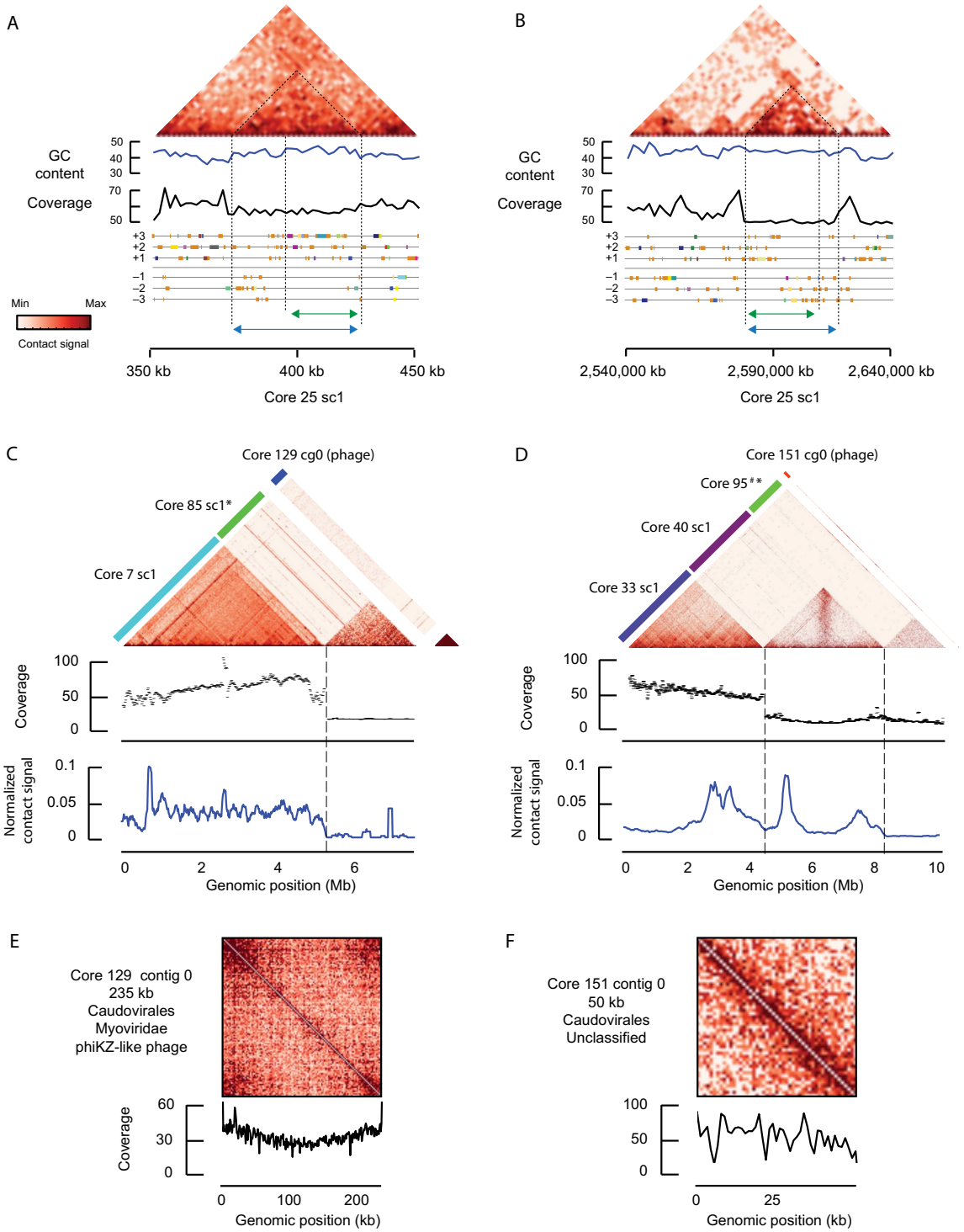


Fig. 3. Analysis of phage-bacteria interactions. (**A** and **B**) Putative prophage sequences in bacterial scaffolds. Magnification of the main diagonal and annotations of the two genomic loci characterized as intact prophages by Phaster in the core 25 scaffold (green rectangles, Fig. 2D). GC content, read coverage distribution, and the predicted ORF annotations (six-frame translation) are indicated under each matrix. Orange genes encode for hypothetical proteins and are enriched in this genomic region. The peculiar contact signals displayed by prophages in contact matrices (see fig. S6) suggest that the border of the prophage locus predicted by Phaster (green double arrows) can be refined because of the meta3C data (dotted black lines and blue double arrows). (**C** and **D**) Representative contact maps between large independent phage contigs (cores 129 and 151) and bacterial scaffolds of interest either (i) display enriched contacts or (ii) present clustered regularly interspaced short palindromic repeats (CRISPR) spacer sequences also found in the phage sequence (scaffold labeled with an asterisk). The read coverage of the bacterial scaffolds and the normalized contact frequencies between the phage contigs and the bacterial scaffolds are plotted under the maps (black and blue graphs, respectively). "#" indicates a set of contigs not scaffolded by GRAAL. (**E** and **F**) Cis contact map and read coverage distribution for the candidate phage contigs from (C) and (D), respectively. A circularization signal appears on the large (235 kb) core 129 contig. The corresponding coverage also points at the possible multiplication of this genomic structure from a discrete position.

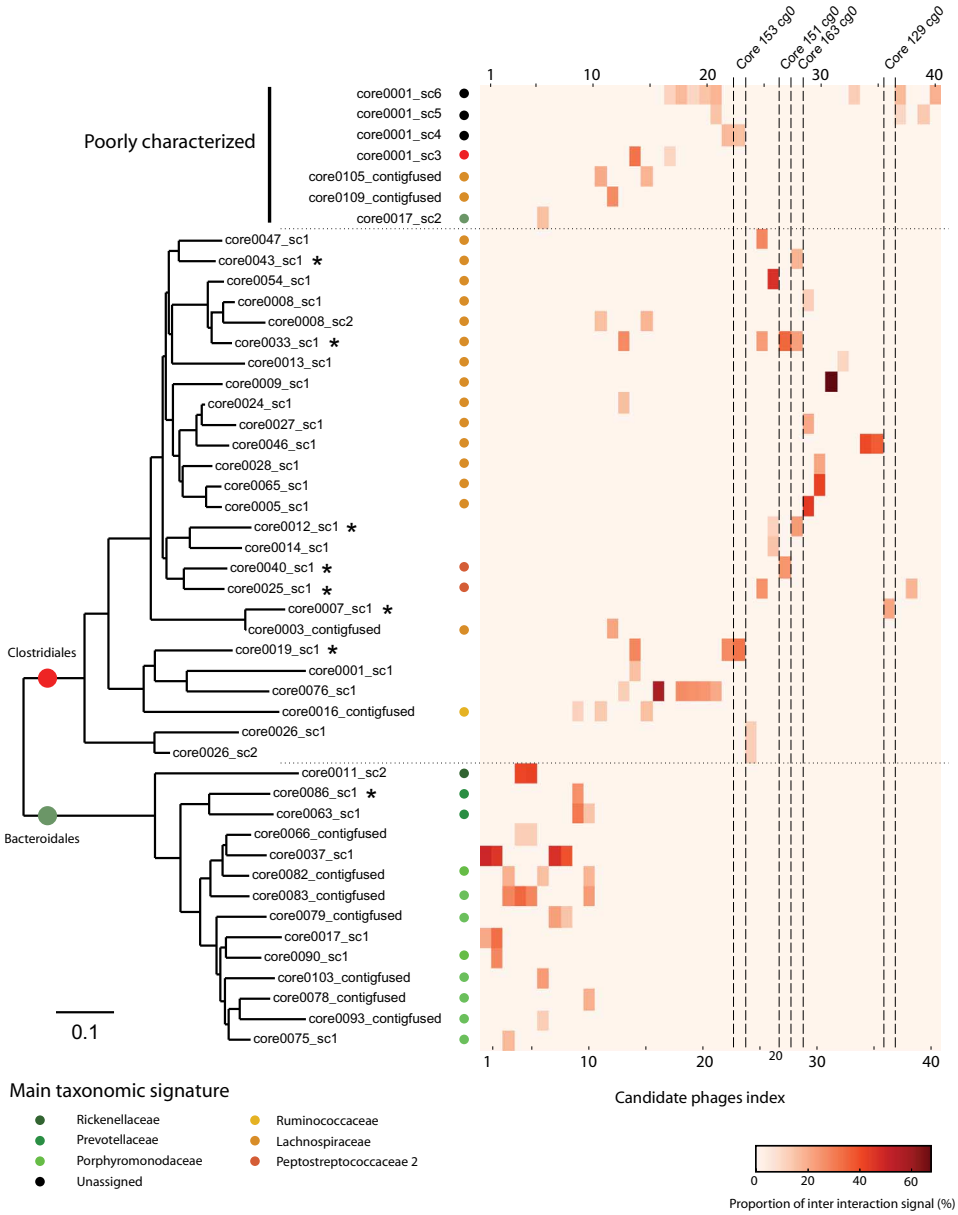


Fig. 4. Overview of phage-bacteria interactions through meta3C. Normalized contact map between the 40 candidate phage contigs in the *x* axis (obtained from the reassembly of small CCs) and the 47 bacterial genome scaffolds/assemblies in the *y* axis. An interaction had to represent at least 10% of the total contacts made by a candidate phage with a bacterial genome scaffold/assembly to be retained. Bacterial genome scaffolds/assemblies were ordered according to their phylogeny relationships (tree on the left of the map). Main taxonomic annotations based on genetic marker analysis are indicated with colored circles next to each predicted bacterial genome. The color scale reflects the contact frequencies, in % of total contacts made by the phage sequence. The stars points at CCs of bacterial genome scaffolds emphasized in Figs. 2 and 3 and fig. S8. The phage contigs outlined along the *x* axis correspond to those described in Fig. 3 and fig. S8.

consistent with the presence of small genomic entities (for instance, see the squares in the upper left corner of cores 11 and 14, contact maps in fig. S4; table S2 and data set S3), leaving room for unexpected or surprising results, such as the identification of new viruses or genetic elements (*28*). However, the exploration of this "dark matter" will require deeper analyses.

Phage assembly and analysis

The annotation of the contigs contained in the small CCs revealed an enrichment in phage sequences, suggesting that some of these pools of contigs correspond to viral genomes. To further investigate these communities, we performed a new round of assemblies (Fig. 2A) on these CCs (see Materials and Methods for details; no GRAAL scaffolding was performed at this stage). Contigs above 10 kb were annotated with a BLASTP search against two National Center for Biotechnology Information (NCBI) databases of viral sequences [Phage Orthologous Group (POG) and Viral databases; $E < 10^{-4}$; Materials and Methods] (*57*). Forty-three contigs ranging from 10 to 235 kb displayed at least one significant hit against the POG database (table S3) and multiple hits against the Viral database (table S3). For instance, 11 putative encoded proteins from the 218 ORFs identified within the largest contig (235 kb, core129 contig0) presented a similarity with proteins from the POG

database, including a genetic marker associated to the phiKZ giant phage family known to infect *Pseudomonas aeruginosa* (POG 3254) (table S3 and fig. S7) (*37*, *38*, *58*). This contig's genomic organization is typical of phage genomes, with ORFs that are largely co-oriented and organized in sizable blocks encoded on the same strand (fig. S7) (*28*, *59*). The contact map of this contig displays a circularization signal, as well as a skewed read coverage, suggesting that bidirectional replication is taking place (Fig. 3E). This large contig was not present in its full length in the first assembly (95% of the sequence was contained within three large contigs), confirming the interest of our approach to assemble and scaffold metagenomes.

Phage-host interactions

As discussed above, assigning phages to their bacterial hosts remains a challenge in metagenomic studies. To see whether quantifying DNA collision events between the phage and the host genome could alleviate this limitation, we computed the normalized contacts between the phiKZ-like contig and the 140 bacterial genome scaffolds (that is, from large CCs). A single bacterial scaffold belonging to the Clostridiales phylum (core7 scaffold1) presented enriched contacts with this long contig (Fig. 3C). This result suggests that this phage genome has frequent contacts with the genome of this bacterial species; hence, this bacterial species hosts the phage. We performed the same analysis of several other putative phage contigs (Fig. 3, D and F, and fig. S8). Notably, we identified a contig (core151 contig0) harboring typical markers from the Caudovirales family and exhibiting several enriched contacts with reconstructed bacterial genomes (Fig. 3D). A refined analysis of those contacts indicates the existence of hot interaction spots of this contig with different loci and points to possible multiple integration sites into the bacterial scaffolds (core33 scaffold1, Clostridiales and core40 scaffold1, Clostridiales). In parallel, we searched for CRISPR spacers found in the different bacterial scaffolds that would present a match on the candidate phage contigs (fig. S9) (*23*, *28*, *29*). We identified 1575 putative spacers and 55 significant blast matches in our candidate contigs (table S4). With only one exception, none of the bacterial scaffolds detected by this analysis displayed enriched contacts with the phage contigs (Fig. 3, C and D, and fig. S8, A and B; bacterial cores labeled with asterisks). For instance, a perfect match was found for a spacer present on the phiKZ-like contig and on the scaffold retrieved from CC #85, but no contacts between the two sequences were detected. One possible explanation is that this bacterium maintains this spacer in its genome as a defense against future infections and therefore contacts between the two genomes are very limited. CRISPR spacer–based predictions are known to detect high rates of false positives, especially when only one hit is detected between the host and its phage (*29*). Additional meta3C data will help to understand these observations and to provide new insights into the ecology of phages and bacteria in the gut.

To broaden the analysis, we studied the contacts between the 43 candidate phage contigs and all 140 bacterial CCs. A host-phage interaction was considered significant when it accounted for at least 10% of all contacts made by the phage sequence. All but three phage candidates displayed at least one, sometimes more, preferred bacterial scaffold(s). An "infection heatmap" was generated to represent the contacts between the phage genomes and the putative host genomes (47 potential hosts were detected), ordered according to their phylogenetic relationship (CheckM pipeline; table S1 and Fig. 4). The infection spectrum of phages in this bacterial community emerges from this representation. Boundaries between clades are consistent with previous studies (*60*). Overall, this first viral-host contact map illustrates the approach's interest and enables further analyses of phage infection dynamics as well as mobile element propagation in complex communities.

CONCLUSION

Overall, the first meta3C experiment performed on a truly complex natural microbiome highlights the power of contact genomics/proximity ligation approaches to study phages and bacterial interactions (*21*). It is worth noting again that this approach does not require multiple experiments: A single meta3C library generated with a single restriction enzyme will bring an important amount of information. Therefore, meta3C could significantly contribute to the full characterization of the genomic structure of complex environmental microbial communities and the analysis of their dynamic changes. The experiment so far does not provide an exhaustive overview of the phage population, mostly because virulent phages that kill bacteria quickly were not sought for. In the future, the present experiment could be backed by the sequencing and genomic analyses of the population of viral particles. That way, one would expect to be able to confront viral particle genomes and phage genomes in contact with bacterial chromosomes, to reach a truly exhaustive characterization of the entire population. Performed over time, the genomes of the different species within a population and the dynamics of mobile elements within the population could be generated, providing valuable insights into the adaptation/evolution of the species present in the ecosystem.

MATERIALS AND METHODS

Generation of meta3C libraries

Feces from a C57BL/6 male mouse were recovered and immediately suspended in 30 ml of 1× tris-EDTA buffer supplemented with 3% of fresh formaldehyde. Fixation proceeded for 1 hour under gentle agitation. Ten milliliters of glycine (2.5 M) was added to the tube, and the quenching was performed for 20 min. The pellet was recovered by centrifugation and stored at −80°C until use. Meta3C libraries were then prepared and sequenced (2 × 75 bp, Illumina NextSeq, 10 first bases as index), as described by Marbouty *et al.* (*34*).

Metagenome assembly

Raw reads were filtered using the QIIME software, as described by Bokulich *et al.* (*61*). A de novo assembly was generated using IDBA-UD v1.0.9 (*40*) with default parameters but without any pre-correction option (raw reads, 193 million PE reads; filtered reads, 169 million PE reads) (resulting assembly, 374,363 contigs; cumulated size, 580 Mb; N50, 3783 bp; maximum size, 490 kb; mean size, 1402 bp). After filtration of contigs of sizes under 500 bp, the total assembly was 521 Mb.

Metagenome analysis

Contigs from the metagenomic assembly were analyzed with the MG-RAST and Kraken pipelines. The MG-RAST server (*42*) allowed automated annotations of complete or draft microbial genomes and provided information on phylogenetic and functional classification of the contigs. Kraken (*44*) is a program that assigns taxonomic labels to short DNA sequences using exact *k*-mer alignments.

Generation of internal control

Concurrently with the mice gut meta3C process, 4 million PE reads from a previous meta3C experiment performed onto a controlled mix of three

bacteria (*E. coli*, *V. cholera*, and *B. subtilis*) (*34*) were used to perform an assembly using the same parameters as above. The resulting contigs were added to the final set of 553,310 contigs from the mice gut assembly, providing a set of 569,146 contigs corresponding to an assembly of 526 Mb.

Identification of CCs

An approach based on the Louvain algorithm (v0.3) (*47*) was used to pool contigs into CCs (fig. S2). Before clustering, contigs were split into 1-kb chunks (without a sliding window). Again, contigs smaller than 500 bp were discarded at this stage (this process resulted in a small loss of 8 Mb of sequences, with a total assembly left of 513 Mb). The resulting 553,310 contigs covered ~90% of the initial assembly (569,146 contigs with the ones from the control experiment, corresponding to an assembly of 526 Mb). Raw reads (plus the 4 million PE reads of the control) were then independently realigned against this set of contigs using Bowtie2 (parameters: –very-sensitive-local) coupled with an iterative procedure, and no ambiguous matches were allowed (*53*). PE information was then included: Whereas two reads of a pair often mapped onto the same contig, 46 million contig pairs were nevertheless bridged by at least one pair of reads. For each pair of contigs, the weighted interaction was normalized by the square root of the product of their respective read coverages.

The Louvain algorithm was run 100 times independently. Its nondeterministic heuristics were exploited to weigh and improve the reliability and stability of the clustering. Each group of contigs that systematically clustered together over the 100 iterations defined a CC (fig. S2). Topologically, this means that the Jaccard distance between every contig index vector (that is, a vector whose components are the indices of the Louvain community to which the vector's contig was assigned for that Louvain iteration) belonging to a single CC is 0.

Contig annotations

Putative coding sequences on the assembled contigs were determined using the MetaGeneMark v3.26 software (*41*) and annotated using BLASTP v2.2.30 and two protein databases (ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/extendedPOGs-10/blastdb/ and ftp://ftp.ncbi.nih.gov/refseq/release/viral/), as well as published hidden Markov models (HMM) [CONJscan (*47*) and MultiMetaGenome (*48*)] and the HMMER software (*62*). Positive hits (+1 positions) were then assigned to the processed contigs (500 bp to 1 kb). Sequences from contigs (>5 kb) recovered after the reassembly of small CCs were annotated using the same databases and HMM models. Among those sequences, the 43 contigs carrying at least one homolog contained in the POG database were considered to be a candidate phage contig (*57*).

Comparison with CAGs

A catalog of mice microbiota genes was retrieved from Xiao *et al.* (*43*) and used as a genome index to map the reads from the two meta3C libraries. Consistent with this work, approximately 60% of PE reads could be aligned unambiguously to this index. Genes were then clustered on the basis of either their CAG index (*43*) or their CC indexes. Contact matrices of the 100 largest groups for each category were then generated (contact scores were normalized by the coverage of each gene).

Assembly of CCs

Contigs from each CC were used as an index to align all meta3C reads with Bowtie2 (mapping parameters: -local -sensitive, ambiguous matches allowed -parameters a-). When at least one member of the PE reads mapped onto one of these indexes, both read sequences were retained. Raw PE sequences recovered for each CC were quality-filtered (see above) and then processed using IDBA-UD v1.0.9 (same parameters as above) to generate a de novo assembly. For each CC, if the cumulated assembly size was larger than 500 kb, then all contigs above 500 bp were retrieved and processed by the scaffolding program GRAAL (*30*). For assemblies smaller than 500 kb, which, for instance, can represent a poorly assembled chromosome because of low coverage, the resulting contigs were directly annotated (see above).

GRAAL scaffolding

GRAAL was run for 100 iterations on the set of contigs (>500 bp) present in a given CC, as described by Marie-Nelly *et al.* (*30*). Briefly, the algorithm fitted the contact data onto a classic DNA polymer model (*63*) and then altered the relative positions and orientations of pairs of DNA sequences to gradually converge toward the most likely 1D genome according to the said model. The model was then readjusted to better fit the new data, and a new iteration began. The duplication mode described by Marie-Nelly *et al.* was not activated. Table S1 summarizes the outcome of this scaffolding step and the generation of large (>500 kb) scaffolds exhibiting the properties of bacterial genomes. The contact signal generated by some of these idiosyncratic properties, such as circularity or the presence of a secondary diagonal, was not predicted by GRAAL's general polymer model. It can sometimes induce scaffolding errors (such as flips of large blocks) readily visible because of the incongruous signal they generate in the contact map of the scaffold (see Marie-Nelly *et al.* for more examples). Hence, manual corrections were added. These are mainly simple modifications of the same nature as GRAAL's (that is, inversions and transpositions) that alleviate incongruities in a self-evident way on the contact map [fig. S5 shows how two modifications (one inversion and one transposition) alleviate all incongruities from a GRAAL scaffold].

Genome completion analysis

The scaffolds generated by GRAAL were analyzed using CheckM (*49*). This program assesses the quality of a genome assembly by checking for the presence of lineage-specific gene markers. This pipeline was also used to build phylogenetic trees and assign taxonomy annotation to the CCs and scaffolds retrieved. Scaffolds/assemblies with less than 10 characterized genetic markers were removed from the phylogenetic tree construction.

Bacterial genome comparison

Scaffolds ranging from CC #6 to CC #22 were annotated and compared using RAST v2.0 (http://rast.nmpdr.org/) (*64*).

Genome annotations

Bacterial scaffolds obtained after GRAAL processing were screened for prophage sequences using the Phaster software (*56*). The putative coding sequences of the phiKZ-like genome (core129 contig0) were annotated using BLASTP v2.2.30 and the NCBI nonredundant RefSeq protein database.

Phage-host prediction through CRISPR spacer analysis

The pilecr v1.06 program was used to screen the different assembled bacterial genomes and to identify 1575 CRSIPR spacers. The candidate phage contigs were then screened for the presence of these spacers using BLASTN v2.2.30 with short query parameters (*28*, *29*). Hits with *E* values lower than 0.1 were retained and are displayed in fig. S9 and table S4.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/3/2/e1602105/DC1

fig. S1. Generation of raw CCs.
fig. S2. Iterative Louvain procedure and characterization of CCs.
fig. S3. Comparison of CAGs and meta3C approaches.
fig. S4. Scaffolding of dozens of bacterial chromosomes.
fig. S5. Example of post-GRAAL scaffold correction.
fig. S6. Structural behavior of phage SPβ in *B. subtilis* genome.
fig. S7. Schematic representation of the phiKZ-like genome.
fig. S8. Interactions of phages with bacterial genomes.
fig. S9. CRISPR spacers' blast output.
table S1. Description of the 140 largest genomic structures (>500 kb) detected in the mice gut microbiome and their assembly/scaffolding statistics.
table S2. Description of the 59 contigs corresponding to candidate phages hailing from the unscaffolded output of the GRAAL software.
table S3. Description of the 43 contigs hailing from the reassembly of small CCs and corresponding to candidate phages.
table S4. CRISPR spacers' blast output (format #6).
data set S1. Contig data (contigs_id, contig_name, GC content, coverage, core_community_index, core_size).
data set S2. Normalized contig network (contig_1, contig_2, normalized interaction).
data set S3. This file contains all the GRAAL scaffolds larger than 300 kb (FASTA format).
data set S4. This file, in complement of data set S3, contains all the contigs not included in the scaffolds larger than 300 kb (FASTA format).
data set S5. This file contains all the CC assemblies (contigs >5 kb, FASTA format) that were not scaffolded by GRAAL because of their small size (cumulated size, <500 kb; see steps 4 and 5 in fig. S2).

REFERENCES AND NOTES

1. J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, R. M. Goodman, Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol.* **5**, R245–R249 (1998).
2. P. Hugenholtz, B. M. Goebel, N. R. Pace, Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774 (1998).
3. J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach; MetaHIT Consortium, P. Bork, S. D. Ehrlich, J. Wang, A human gut microbial gene catalog established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
4. J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, H. O. Smith, Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
5. N. R. Pace, A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
6. F. Partensky, W. R. Hess, D. Vaulot, Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63**, 106–127 (1999).
7. A. L. Kau, P. P. Ahern, N. W. Griffin, A. L. Goodman, J. I. Gordon, Human nutrition, the gut microbiome and the immune system. *Nature* **474**, 327–336 (2011).
8. J. F. Cryan, T. G. Dinan, Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* **13**, 701–712 (2012).
9. L. Philippot, J. M. Raaijmakers, P. Lemanceau, W. H. van der Putten, Going back to the roots: The microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.* **11**, 789–799 (2013).
10. J. Yang, J. W. Kloepper, C.-M. Ryu, Rhizosphere bacteria help plants tolerate abiotic stress. *Trends Plant Sci.* **14**, 1–4 (2009).
11. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
12. S. Guermazi, P. Daegelen, C. Dauga, D. Rivière, T. Bouchez, J. J. Godon, G. Gyapay, A. Sghir, E. Pelletier, J. Weissenbach, D. Le Paslier, Discovery and characterization of a new bacterial candidate division by an anaerobic sludge digester metagenomic approach. *Environ. Microbiol.* **10**, 2111–2123 (2008).
13. P. Bork, C. Bowler, C. de Vargas, G. Gorsky, E. Karsenti, P. Wincker, *Tara* Oceans studies plankton at planetary scale. *Science* **348**, 873 (2015).
14. H. B. Nielsen, M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. L. Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J.-M. Batto, M. B. Quintanilha dos Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Doré, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas, S. Kennedy, K. Kristiansen, J. R. Kultima, P. Léonard, F. Levenez, O. Lund, B. Moumen, D. Le Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sørensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada; MetaHIT Consortium, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, S. D. Ehrlich, Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
15. A. Schlüter, L. Krause, R. Szczepanowski, A. Goesmann, A. Pühler, Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. *J. Biotechnol.* **136**, 65–76 (2008).
16. V. Sentchilo, A. P. Mayer, L. Guy, R. Miyazaki, S. G. Tringe, K. Barry, Community-wide plasmid gene mobilization and selection. *ISME J.* **7**, 1173–1186 (2013).
17. L. A. Ogilvie, B. V. Jones, The human gut virome: A multifaceted majority. *Front. Microbiol.* **6**, 918 (2015).
18. E. S. Lim, Y. Zhou, G. Zhao, I. K. Bauer, L. Droit, I. M. Ndao, B. B. Warner, P. I. Tarr, D. Wang, L. R. Holtz, Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).
19. M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, F. Rohwer, Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14250–14255 (2002).
20. D. M. Kristensen, A. R. Mushegian, V. V. Dolja, E. V. Koonin, New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.* **18**, 11–19 (2010).
21. J.-F. Flot, H. Marie-Nelly, R. Koszul, Contact genomics: Scaffolding and phasing (meta) genomes using chromosome 3D physical signatures. *FEBS Lett.* **589** (20 Pt. A), 2966–2974 (2015).
22. C. Canchaya, G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann, H. Brüssow, Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).
23. A. Stern, E. Mick, I. Tirosh, O. Sagy, R. Sorek, CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22**, 1985–1994 (2012).
24. C. A. Suttle, Viruses in the sea. *Nature* **437**, 356–361 (2005).
25. K. D. Seed, M. Yen, B. J. Shapiro, I. J. Hilaire, R. C. Charles, J. E. Teng, Evolutionary consequences of intra-patient phage predation on microbial populations. *eLife* **3**, e03497 (2014)
26. B. M. Davis, M. K. Waldor, Filamentous phages linked to virulence of *Vibrio cholerae*. *Curr. Opin. Microbiol.* **6**, 35–42 (2003).
27. S. L. Welkos, R. K. Holmes, Regulation of toxinogenesis in *Corynebacterium diphtheriae*. I. Mutations in bacteriophage β that alter the effects of iron on toxin production. *J. Virol.* **37**, 936–945 (1981).
28. B. E. Dutilh, N. Cassman, K. McNair, S. E. Sanchez, G. G. Z. Silva, L. Boling, J. J. Barr, D. R. Speth, V. Seguritan, R. K. Aziz, B. Felts, E. A. Dinsdale, J. L. Mokili, R. A. Edwards, A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
29. R. A. Edwards, K. McNair, K. Faust, J. Raes, B. E. Dutilh, Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2015).
30. H. Marie-Nelly, M. Marbouty, A. Cournac, J.-F. Flot, G. Liti, D. P. Parodi, S. Syan, N. Guillén, A. Margeot, C. Zimmer, R. Koszul, High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5**, 5695 (2014).
31. N. Kaplan, J. Dekker, High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
32. C. W. Beitel, L. Froenicke, J. M. Lang, I. F. Korf, R. W. Michelmore, J. A. Eisen, A. E. Darling, Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
33. J. N. Burton, I. Liachko, M. J. Dunham, J. Shendure, Species-level deconvolution of metagenome assemblies with Hi-C–based contact probability maps. *G3* **4**, 1339–1346 (2014).
34. M. Marbouty, A. Cournac, J.-F. Flot, H. Marie-Nelly, J. Mozziconacci, R. Koszul, Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife* **3**, e03318 (2014).
35. M. Marbouty, R. Koszul, Metagenome analysis exploiting high-throughput chromosome conformation capture (3C) data. *Trends Genet.* **31**, 673–682 (2015).
36. J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
37. V. V. Mesyanzhinov, J. Robben, B. Grymonprez, V. A. Kostyuchenko, M. V. Bourkaltseva, N. N. Sykilinda, V. N. Krylov, G. Volckaert, The genome of bacteriophage φKZ of *Pseudomonas aeruginosa*. *J. Mol. Biol.* **317**, 1–19 (2002).
38. A. Cornelissen, S. C. Hardies, O. V. Shaburova, V. N. Krylov, W. Mattheus, A. M. Kropinski, Complete genome sequence of the giant virus OBP and comparative genome analysis of the diverse ϕKZ-related phages. *J. Virol.* **86**, 1844–1852 (2012).

39. A. S. Waller, T. Yamada, D. M. Kristensen, J. R. Kultima, S. Sunagawa, E. V. Koonin, P. Bork, Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME J.* **8**, 1391–1402 (2014).
40. Y. Peng, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin, IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
41. W. Zhu, A. Lomsadze, M. Borodovsky, Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132 (2010).
42. F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, R. A. Edwards, The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
43. L. Xiao, Q. Feng, S. Liang, S. B. Sonne, Z. Xia, X. Qiu, X. Li, H. Long, J. Zhang, D. Zhang, C. Liu, Z. Fang, J. Chou, J. Glanville, Q. Hao, D. Kotowska, C. Colding, T. R. Licht, D. Wu, J. Yu, J. J. Y. Sung, Q. Liang, J. Li, H. Jia, Z. Lan, V. Tremaroli, P. Dworzynski, H. B. Nielsen, F. Bäckhed, J. Doré, E. Le Chatelier, S. D. Ehrlich, J. C. Lin, M. Arumugam, J. Wang, L. Madsen, K. Kristiansen, A catalog of the mouse gut metagenome. *Nat. Biotechnol.* **33**, 1103–1108 (2015).
44. D. E. Wood, S. L. Salzberg, Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
45. S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
46. J. Guglielmini, B. Néron, S. S. Abby, M. P. Garcillán-Barcia, F. de la Cruz, E. P. C. Rocha, Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* **42**, 5715–5727 (2014).
47. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
48. M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, P. H. Nielsen, Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
49. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
50. J. Guglielmini, L. Quintais, M. P. Garcillán-Barcia, F. de la Cruz, E. P. C. Rocha, The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLOS Genet.* **7**, e1002222 (2011).
51. J. M. Lang, A. E. Darling, J. A. Eisen, Phylogeny of bacterial and archaeal genomes using conserved genes: Supertrees and supermatrices. *PLOS ONE* **8**, e62510 (2013).
52. M. A. Umbarger, E. Toro, M. A. Wright, G. J. Porreca, D. Baù, S.-H. Hong, M. J. Fero, L. J. Zhu, M. A. Marti-Renom, H. H. McAdams, L. Shapiro, J. Dekker, G. M. Church, The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell* **44**, 252–264 (2011).
53. M. Marbouty, A. Le Gall, D. I. Cattoni, A. Cournac, A. Koh, J.-B. Fiche, J. Mozziconacci, H. Murray, R. Koszul, M. Nollmann, Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. *Mol. Cell* **59**, 588–602 (2015).
54. M.-E. Val, M. Marbouty, F. de Lemos Martins, S. P. Kennedy, H. Kemble, M. J. Bland, C. Possoz, R. Koszul, O. Skovgaard, D. Mazel, A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. *Sci. Adv.* **2**, e1501914 (2016).
55. X. Wang, T. B. K. Le, B. R. Lajoie, J. Dekker, M. T. Laub, D. Z. Rudner, Condensin promotes the juxtaposition of DNA flanking its loading site in *Bacillus subtilis*. *Genes Dev.* **29**, 1661–1675 (2015).
56. D. Arndt, J. R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang, D. S. Wishart, PHASTER: A better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
57. D. M. Kristensen, X. Cai, A. Mushegian, Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *J. Bacteriol.* **193**, 1806–1814 (2011).
58. D. M. Kristensen, A. S. Waller, T. Yamada, P. Bork, A. R. Mushegian, E. V. Koonin, Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.* **195**, 941–950 (2013).
59. S. Akhter, R. K. Aziz, R. A. Edwards, *PhiSpy*: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **40**, e126 (2012).
60. C. O. Flores, J. R. Meyer, S. Valverde, L. Farr, J. S. Weitz, Statistical structure of host–phage interactions. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E288–E297 (2011).
61. N. A. Bokulich, S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, J. G. Caporaso, Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10**, 57–59 (2013).
62. S. R. Eddy, A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
63. K. Rippe, Making contacts on a nucleic acid polymer. *Trends Biochem. Sci.* **26**, 733–740 (2001).
64. R. Overbeek, R. Olson, G. D. Pusch, G. J. Olsen, J. J. Davis, T. Disz, R. A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, A. R. Wattam, F. Xia, R. Stevens, The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2014).

Acknowledgments: We thank M. de Paepe and E. Rocha for fruitful discussions. We also thank J. Mozziconacci for helpful suggestions, J.-F. Flot for performing the original IDBA-UD assembly and Kraken analysis, and T. Pedron for providing us the mouse samples. **Funding:** This research was supported by funding from the European Research Council (ERC) under the 7th Framework Program (FP7/2007-2013)/ERC grant agreement 260822 (to R.K.). **Author contributions:** M.M. and R.K. conceived the analysis. M.M. performed the experiments. M.M., L.B., and A.C. performed the analysis, and M.M. and R.K. interpreted the results. M.M. and R.K. wrote the manuscript. **Competing interests:** The GRAAL program is owned by Institut Pasteur. The entire program and its full source code are freely available online for noncommercial purposes, but commercial usage requires a specific license. R.K., M.M., L.B., and A.C. have a patent application, PCT/EP2015/064286, submitted 12/30/2015, related to the described work through Institut Pasteur; the publication number is WO2015197711 A1. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. Raw sequences are accessible on Sequence Read Archive database through the following accession number: SRX1434905.

Submitted 2 September 2016
Accepted 9 January 2017
Published 17 February 2017
10.1126/sciadv.1602105

Citation: M. Marbouty, L. Baudry, A. Cournac, R. Koszul, Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci. Adv.* **3**, e1602105 (2017).

Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/3/2/e1602105/DC1

Supplementary Materials for

Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay

Martial Marbouty, Lyam Baudry, Axel Cournac, Romain Koszul

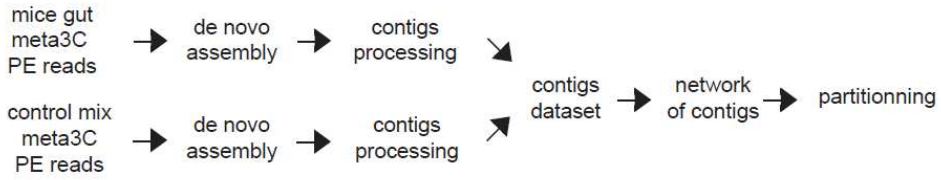
Published 17 February 2017, *Sci. Adv.* **3**, e1602105 (2017)
DOI: 10.1126/sciadv.1602105

The PDF file includes:

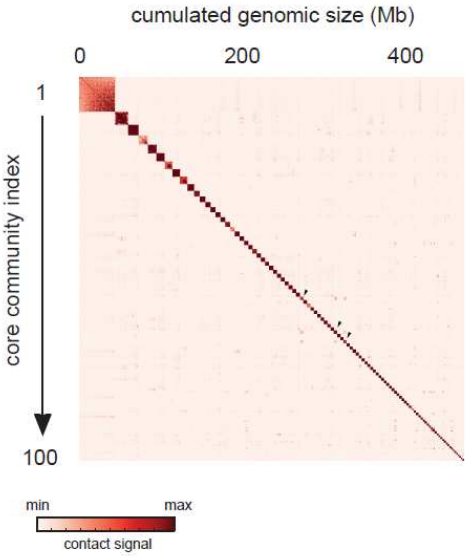
Other Supplementary Material for this manuscript includes the following:
(available at advances.sciencemag.org/cgi/content/full/3/2/e1602105/DC1)

- table S1 (Microsoft Excel format). Description of the 140 largest genomic structures (>500 kb) detected in the mice gut microbiome and their assembly/scaffolding statistics.
- table S2 (Microsoft Excel format). Description of the 59 contigs corresponding to candidate phages hailing from the unscaffolded output of the GRAAL software.
- table S3 (Microsoft Excel format). Description of the 43 contigs hailing from the reassembly of small CCs and corresponding to candidate phages.
- table S4 (Microsoft Excel format). CRISPR spacers' blast output (format #6).

A


mice gut
meta3C
PE reads
de novo
assembly
contigs
processing
control mix
meta3C
PE reads
de novo
assembly
contigs
processing
contigs
dataset
network
of contigs
partitionning

B


cumulated genomic size (Mb)
0
200
400
1
core community index
100
min
max
contact signal

C

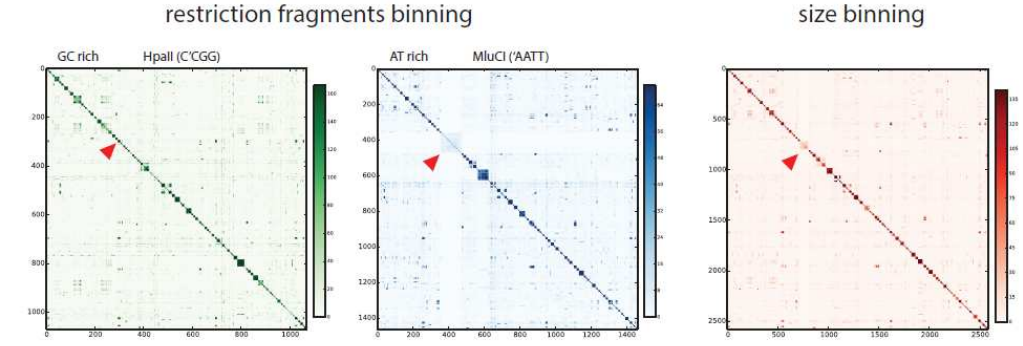

restriction fragments binning
size binning
GC rich
HpaII (C'CGG)
AT rich
MluCI ('AATT)

fig. S1. Generation of raw CCs. (**A**) Workflow of the original assembly process. The meta3C reads from an *in vitro* mixture of 3 species are processed concurrently with the meta3C gut library. Contigs from both libraries are pooled together before the segmentation of the network. (**B**) Contact map of the 100 largest CCs obtained after 100 Louvain iterations (1 vector = 200kb). The x and y axis are labeled with the cumulated DNA size and the index of the community, respectively. Black triangles point to the three control species (1 vector = 50 kb). (**C**) Illustration of the "visibility" of the communities with respect to different restriction enzymes. The biggest meta3C CCs carrying more than 1,000 contigs after one Louvain iteration on the pooled datasets are represented using *HpaII* (left) and *MluCI* (middle) restriction patterns and contact data. Contact maps are binned at the kb scale and combining both datasets (right) .Red triangles: extreme example of a community carrying AT-rich contigs. This community is being split into multiple small pieces by the MluCI enzyme that recognizes AATT sites.

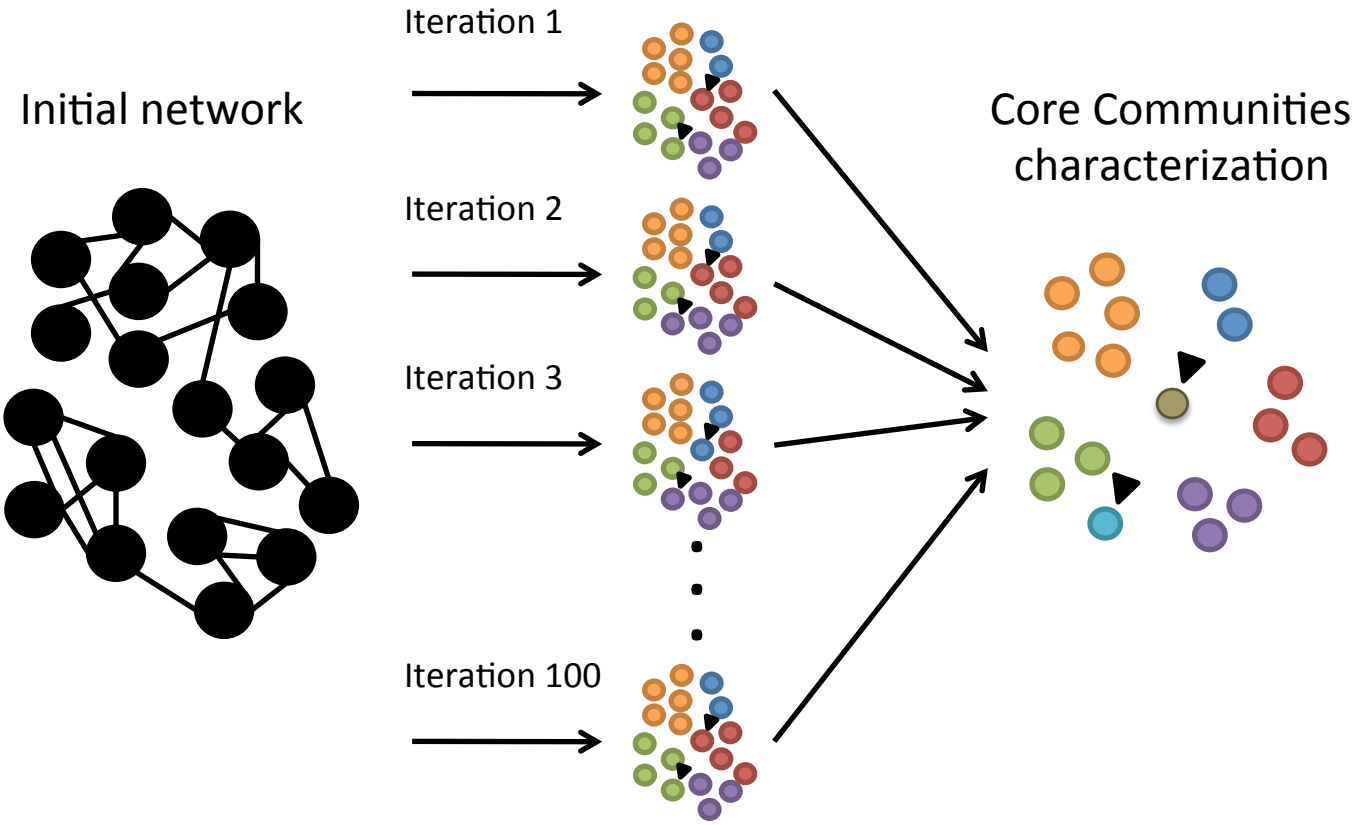

Initial network
Iteration 1
Iteration 2
Iteration 3
Iteration 100
Core Communities characterization

fig. S2. Iterative Louvain procedure and characterization of CCs. Starting from an invariant network of contig, Louvain segmentation was independently performed 100 times (contigs are represented as colored circles). The distribution of colors in the middle panel represent the groups (a.k.a. communities) of contigs characterized at each the iterations. Results from these independent iterations were then combined to characterize core communities, *i.e.* contigs that always cluster together during the 100 iterations (right panel). Dark triangles point at contigs that jump from one community to another over the iteration. As a result, these contigs cluster into small and isolated CCs after compilation of the 100 iterations.

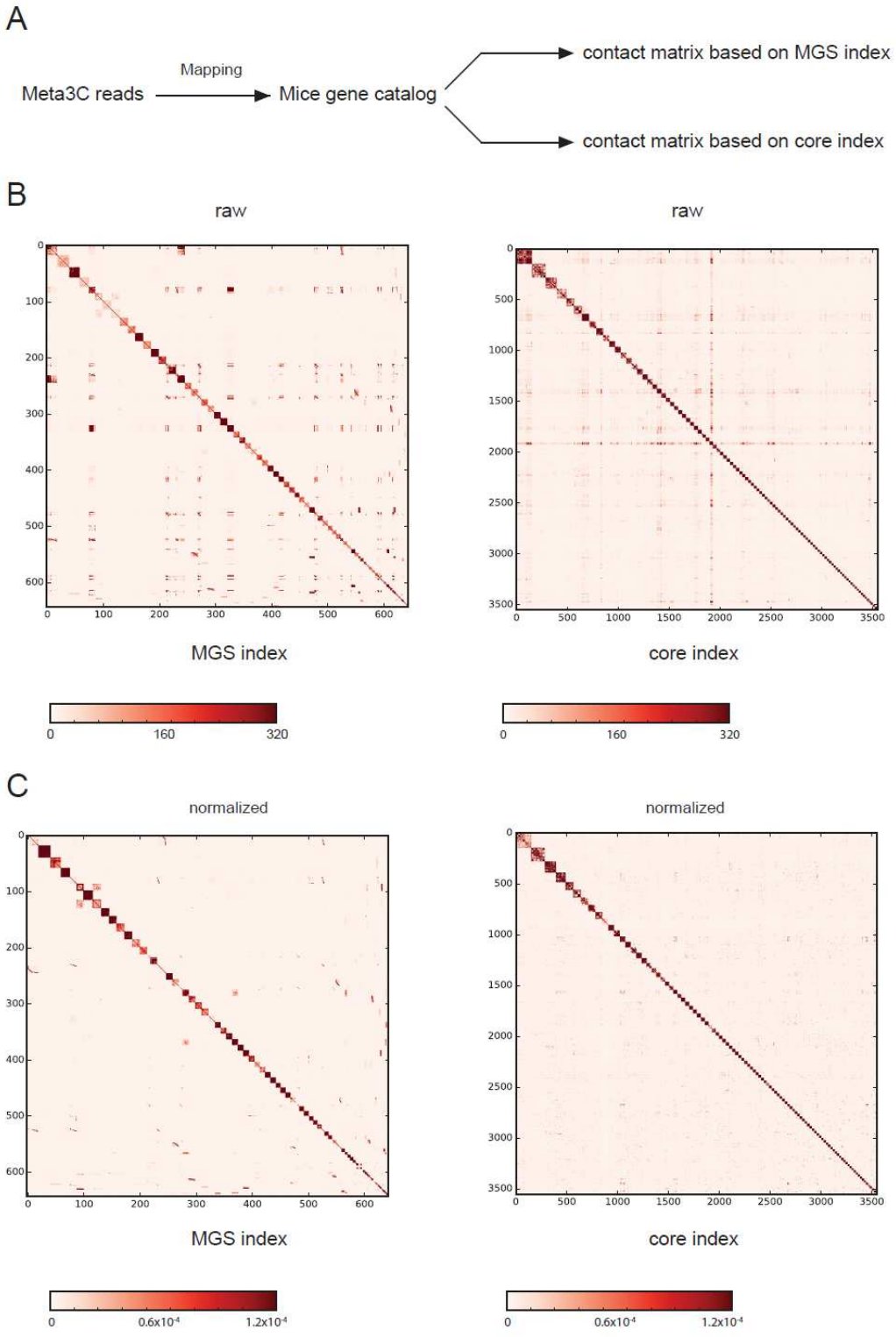

A
Meta3C reads
Mapping
Mice gene catalog
contact matrix based on MGS index
contact matrix based on core index
B
raw
raw
MGS index
core index
0
160
320
C
normalized
normalized
MGS index
core index
0
0.6x10-4
1.2x10-4

fig. S3. Comparison of CAGs and meta3C approaches. (**A**) Workflow used to compare both methods and generate the contact matrices. First, DNA sequences from the mice microbiome genes catalog (http://www.cbs.dtu.dk/databases/CAG/mouse/) were used as an index to map all meta3C reads. Genes were then pooled according to their CAGs or CCs indexes. (**B**) Left: raw contact map of the 100 most covered CAGs described in Xiao et al. 2015 (*43*). Right: 100 largest CCs obtained through the Louvain iterative procedure (right). (**C**) Same contact maps as above, after normalization by the reads coverage.

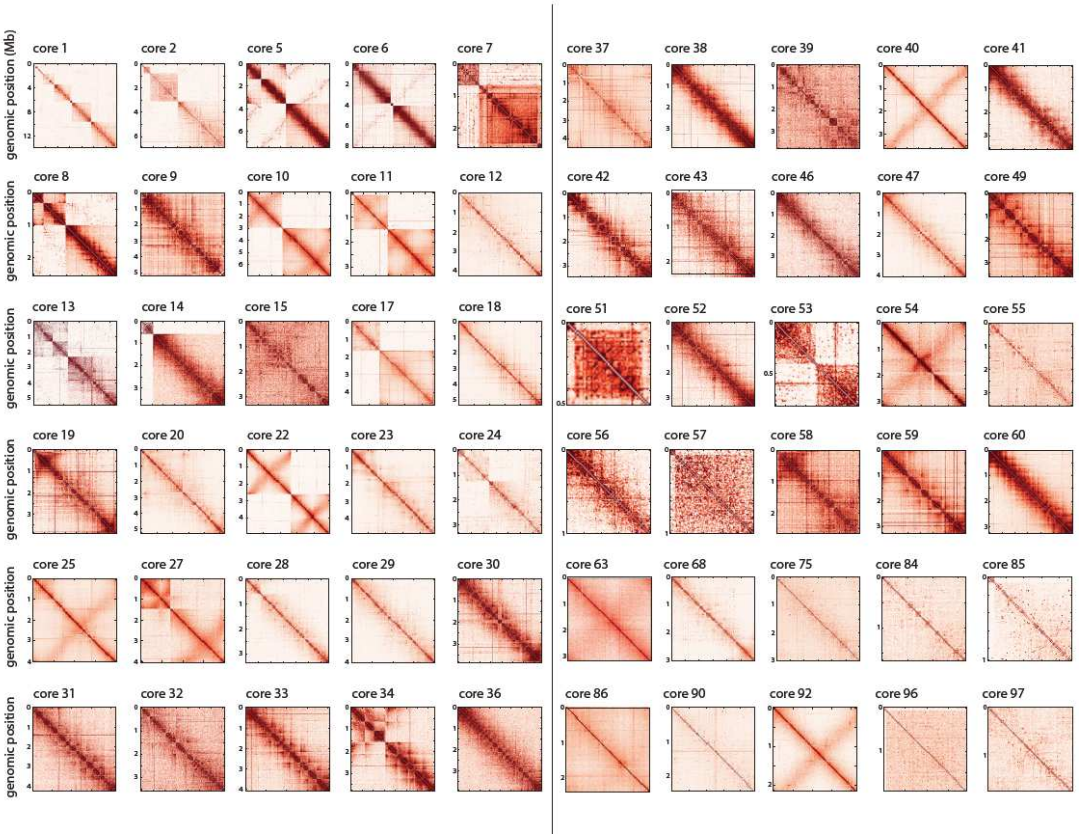

genomic position (Mb)
core 1
core 2
core 5
core 6
core 7
core 37
core 38
core 39
core 40
core 41
genomic position
core 8
core 9
core 10
core 11
core 12
core 42
core 43
core 46
core 47
core 49
genomic position
core 13
core 14
core 15
core 17
core 18
core 51
core 52
core 53
core 54
core 55
genomic position
core 19
core 20
core 22
core 23
core 24
core 56
core 57
core 58
core 59
core 60
genomic position
core 25
core 27
core 28
core 29
core 30
core 63
core 68
core 75
core 84
core 85
genomic position
core 31
core 32
core 33
core 34
core 36
core 86
core 90
core 92
core 96
core 97

fig. S4. Scaffolding of dozens of bacterial chromosomes. 60 contact maps displaying the scaffolds recovered after GRAAL processing of large CCs. These maps correspond to assemblies where the largest scaffold was larger than 300kb (10 kb bins; all scaffolds > 50 kb are represented). CCs indices are indicated above each map. Y-axis: cumulated DNA size.

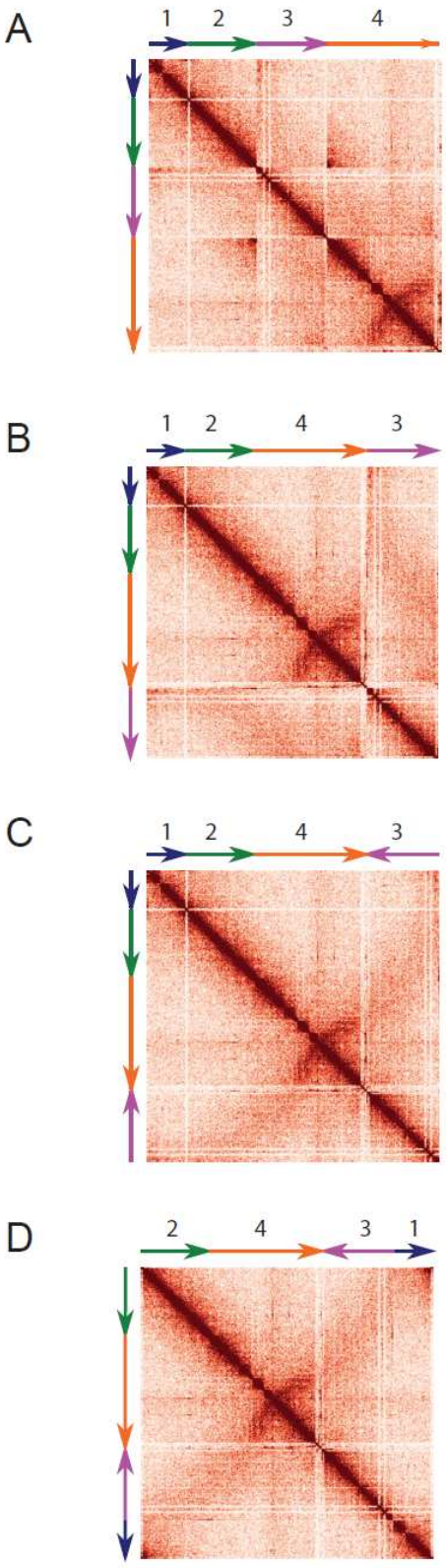

A
1 2 3 4
B
1 2 4 3
C
1 2 4 3
D
2 4 3 1

fig. S5. Example of post-GRAAL scaffold correction. (**A**) Contact map of a scaffold retrieved from CC #54 after GRAAL scaffolding. The inconsistencies in the signal delimit four regions in the heatmap (oriented colored arrows numbered from 1 to 4). (**B**) Swapping regions 3 and 4 (purple and orange arrows) eliminates the long-range incongruities between the two regions. (**C**) Inverting region 3 (purple arrow) eliminates all incongruities in the map, unveiling the secondary diagonal. (**D**) The resulting scaffold is centered on the crossing between the principal and secondary diagonal (circular permutation). As a result, a strong signal appears clearly in the map corner, indicating that the scaffold is circular.

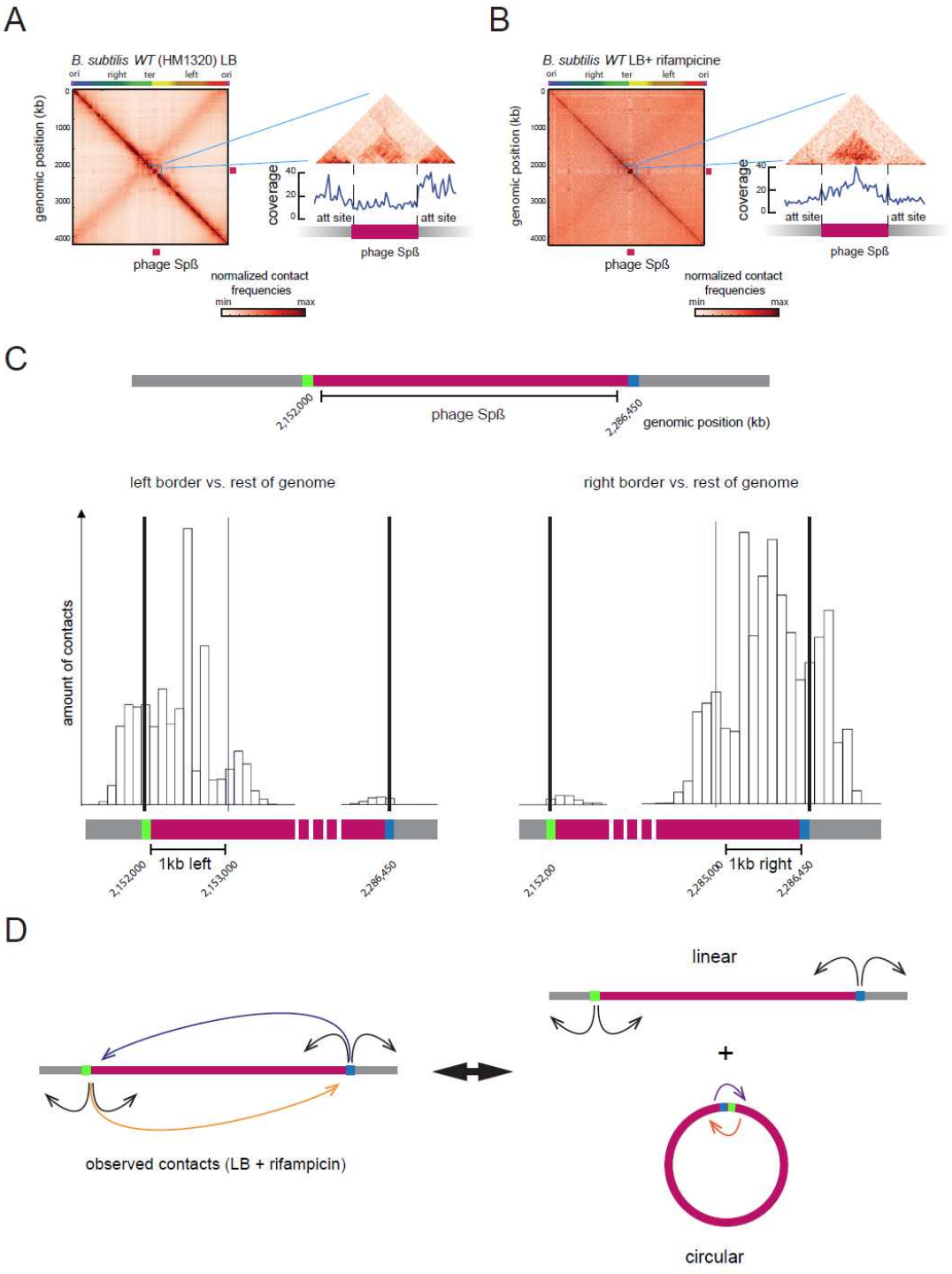
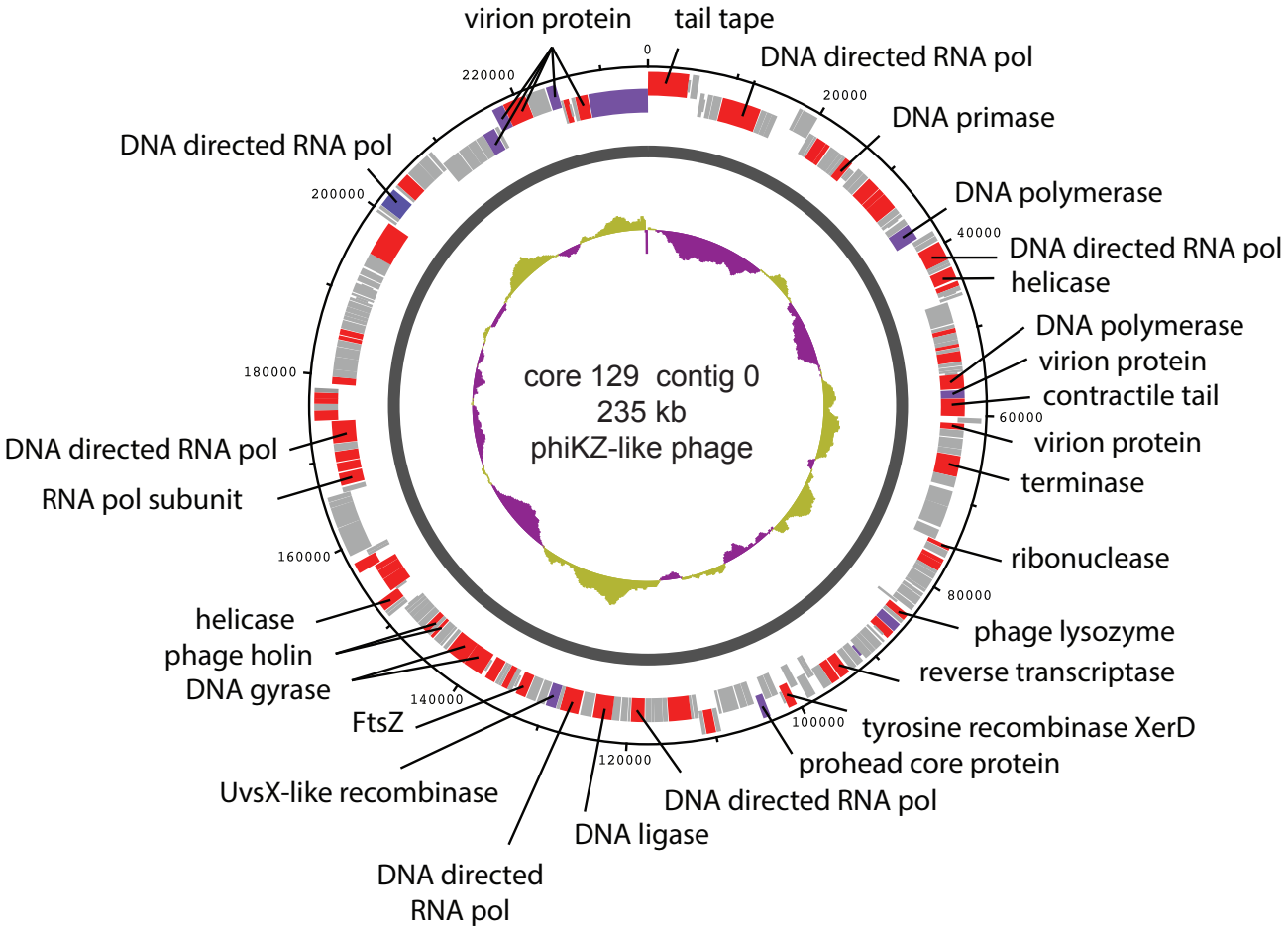

A
B. subtilis WT (HM1320) LB
ori
right
ter
left
ori
genomic position (kb)
coverage
att site
att site
phage Spß
normalized contact frequencies
min
max
B
B. subtilis WT LB+ rifampicine
C
2,152,000
phage Spß
2,286,450
genomic position (kb)
left border vs. rest of genome
right border vs. rest of genome
amount of contacts
1kb left
2,153,000
1kb right
2,285,000
2,286,450
D
linear
+
circular
observed contacts (LB + rifampicin)

fig. S6. Structural behavior of phage SPβ in *B. subtilis* genome. (**A**, **B**) Normalized contact maps of the chromosome of *Bacillus subtilis* cultures (strain HM1320) in the absence (A) or presence (B) of rifampicin (data from Marbouty et al., 2015; *53*). Purple bars: prophage Spß. Aside each contact map, a magnification of the region containing the Spß sequence in purple and its flanking sequences is presented with a 90 CCW rotation. Dashed lines delimit the borders of the phage sequence. In both conditions, the sequence of the Spß is clearly recognizable on the contact map and well-separated from the rest of the *B. subtilis* genome. Whereas the differential GC% content between prophage sequences and the *B. subtilis* genome may disturb the regular pattern of interaction of this region, treatment with rifampicin clearly induces a different response. The prophage SPβ exhibits a striking self-interaction pattern in the presence of rifampicin. The read coverage is plotted under the magnification panel, illustrating how the rifampicin treatment induces a multiplication of the phage locus, suggesting activation of replication of the phage is occuring. (**C**) Distributions of contacts made by 100bp bins over a 1kb window within the left and right extremities (black lines) of the Spß sequence show a enrichments in long-range contacts between the two extremities, near the site-specific recombination *att* loci (green and blue marks). This enrichment shows that these extremities are in a close vicinity to each other for in a subpopulation of the phage sequences, most likely present as circular molecules (alternatively, the formation of a stable loop bridging the two extremities of the phage at its basis could also explain, to some extent, this signal, but the uneven distribution towards the phage sequence suggests this is not the case). Because these long-range contacts stop precisely at the opposite *att* site, one can assess that they result most likely from a recombination event between *att* sequences and excision of the prophage. (**D**) Illustration of the contact patterns observed (left; blue and orange arrows illustrate the long-range contacts between the phage extremities) and the molecules likely to generate them (right).

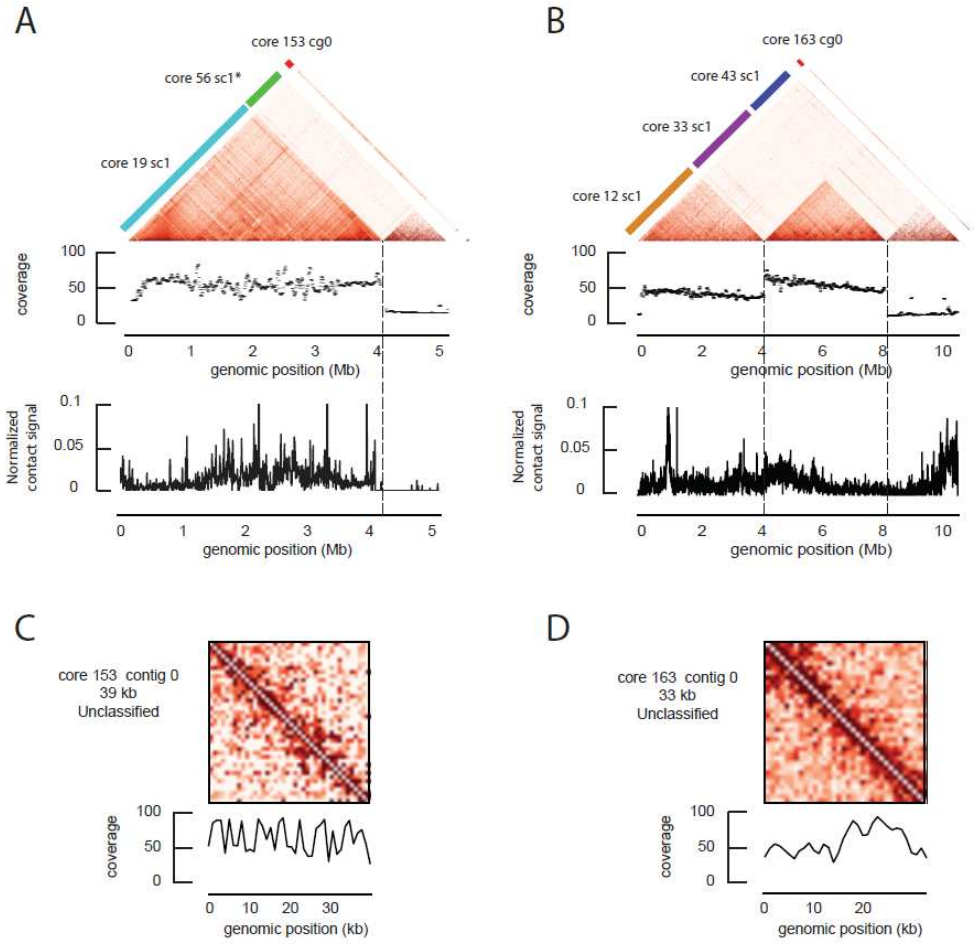
A



Hit against NCBI viral database

Hit against high quality POG database

fig. S7. Schematic representation of the phiKZ-like genome. (**A**) Circular representation of the phiKZ-like genome. The 218 putative open reading frames predicted with Metagenemark (see Materials and Methods) are indicated as squares. Depending on their orientation they are indicated in the upper part (forward strand) or the bottom part (reverse strand). Purple squares indicates hit against the high quality Phage Orthologous Groups (POG) database. Red squares indicate hits against the viral database from NBCI. Putative functions of the encoded proteins were determined using blastp against the refseq database from NCBI with an E-value threshold of 1.10^{-4}. GC content (2kb bins) is indicated on a diagram at the center of the circle (green: average GC content above 50%; purple: average GC content under 50%). Genomic coordinates are also indicated in the periphery of the circle representation. Annotation points to a relation of these sequences with the phiKZ-like phage family.


A
core 153 cg0
core 56 sc1*
core 19 sc1
coverage
100
50
0
0 1 2 3 4 5
genomic position (Mb)
Normalized contact signal
0.1
0.05
0
0 1 2 3 4 5
genomic position (Mb)
B
core 163 cg0
core 43 sc1
core 33 sc1
core 12 sc1
coverage
100
50
0
0 2 4 6 8 10
genomic position (Mb)
Normalized contact signal
0.1
0.05
0
0 2 4 6 8 10
genomic position (Mb)

C
core 153 contig 0
39 kb
Unclassified
coverage
100
50
0
0 10 20 30
genomic position (kb)
D
core 163 contig 0
33 kb
Unclassified
coverage
100
50
0
0 10 20
genomic position (kb)

fig. S8. Interactions of phages with bacterial genomes. (**A**–**D**) Analyses of additional candidate phage contigs (core 153 contig 0 and core 163 contig 0) interactions with bacterial scaffolds, similar to the data described in Fig. 4.

A

core129 contig0

```
> core0085_sc1
Length=41

 Score = 66.6 bits (41),  Expect = 6e-12
 Identities = 41/41 (100%), Gaps = 0/41 (0%)
 Strand=Plus/Minus

Query  7772  TTTTGTAATACTCACACCGTTTCACCTCCTAGCTTAGCTAG  7812
              |||||||||||||||||||||||||||||||||||||||||
Sbjct  41    TTTTGTAATACTCACACCGTTTCACCTCCTAGCTTAGCTAG  1
```

core151 contig0

```
> core0095_contigfused
Length=36

 Score = 47.5 bits (29),  Expect = 1e-06
 Identities = 31/33 (94%), Gaps = 0/33 (0%)
 Strand=Plus/Minus

Query  30035  GTAGTCCGGGATGCTCCGCTCCAGCATATTTGC  30067
               ||| || |||||||||||||||||||||||||
Sbjct  33     GTAATCGGGGATGCTCCGCTCCAGCATATTTGC  1
```

core153 contig0

```
> core0056_sc1
Length=35

 Score = 42.8 bits (26),  Expect = 2e-05
 Identities = 30/34 (88%), Gaps = 0/34 (0%)
 Strand=Plus/Plus

Query  21021  TTTGTAAAGATAGCTTTCGGACGCTTATAGTTTG  21054
               |||||||| || ||||||||||||||| || ||||
Sbjct  1      TTTGTAAAAATTGCTTTCGGACGCTTGTAATTTG  34
```

fig. S9. CRISPR spacers' blast output. Screenshot of the blast output for CRISPR spacer searches against core129 contig0, core151 contig0 and core153 contig0.

table S1. Description of the 140 largest genomic structures (>500 kb) detected in the mice gut microbiome and their assembly/scaffolding statistics.

table S2. Description of the 59 contigs corresponding to candidate phages hailing from the unscaffolded output of the GRAAL software.

table S3. Description of the 43 contigs hailing from the reassembly of small CCs and corresponding to candidate phages. 3 of these contigs do not show enriched contacts with any of the 140 large genomic structures described in table S1 and are not indexed (First column – xxx).

table S4. CRISPR spacers' blast output (format #6).

data set S1. Contig data (contigs_id, contig_name, GC content, coverage, core_community_index, core_size).

data set S2. Normalized contig network (contig_1, contig_2, normalized interaction).

data set S3. This file contains all the GRAAL scaffolds larger than 300 kb (FASTA format).

data set S4. This file, in complement of data set S3, contains all the contigs not included in the scaffolds larger than 300 kb (FASTA format).

data set S5. This file contains all the CC assemblies (contigs >5 kb, FASTA format) that were not scaffolded by GRAAL because of their small size (cumulated size, <500 kb; see steps 4 and 5 in fig. S2).

4.2 MetaTOR: recovering high-quality bins and dynamics insights

[Original Research]

MetaTOR recovers high-quality metagenome-assembled genomes (MAGs) from mammalian gut proximity-ligation (meta3C) libraries, regardless the number of samples

Lyam Baudry[1,2,3†], Théo Foutel-Rodier[1,2,3†], Agnès Thierry[1,2], Romain Koszul[1,2,*] and Martial Marbouty[1,2,*]

[1]Institut Pasteur, Unité Régulation Spatiale des Génomes, UMR3525, CNRS, 75015 Paris, France

[2]Institut Pasteur, Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), Paris, F-75015

[3] Sorbonne Université, Collège Doctoral, F-75005, Paris, France

†These authors have contributed equally to this work.

*Corresponding authors

Contacts: romain.koszul@pasteur.fr or martial.marbouty@pasteur.fr

Keywords: metagenomics, gut microbiome, proximity ligation, meta3C, metagenomics binning, metagenomics analysis, proximity ligation assay, binning algorithm,

4 Figures

4 Table

4,300 Words

Abstract

Characterizing the full genomic structure of complex microbial communities is a key step towards the understanding of their diversity, dynamics and evolution. These investigations are typically done through the analysis of millions of short DNA sequences directly extracted from the environment. Computational tools exploiting these metagenomics data display intrinsic limitations or constraints, such as assumptions regarding the genomic content of the genomes being investigated, and/or the need for multiple samples to accurately bin the interleaved metagenomic sequences according to their covariant characteristics. Here we present MetaTOR, an open-source and transparent computational solution that exploits meta3C, i.e. proximity ligation experiments (3C, Hi-C) performed on metagenomic samples, to bin the resulting sequencing reads into individual genomes according to their 3D contact frequencies. MetaTOR was applied on a combination of 20 newly generated meta3C libraries of mice gut microbiote sampled over time. We quantified the ability of the program to recover high-quality metagenomics-assembled genomes (MAGs) from metagenomics assemblies generated directly from the meta3C libraries. Whereas 16 MAGs are identified in the 148Mb assembly generated using a single meta3C library, MetaTOR identifies 122 MAGs in the 763Mb assembly generated from the merged 20 meta3C libraries, corresponding to a ~40% increase compared to MAGs recovered using current, state-of-the-art hybrid binning programs. Overall, the completion and contamination of meta3C bins were also improved. These results underline the potential of meta3C (and 3C based approaches) in metagenomics projects.

1. Introduction

Microbial communities hold important roles in the maintenance of multiple ecosystems (Philippot et al. 2013), including the human gut (Cho and Blaser 2012). Understanding the complexity of these ecosystems is a complex task, and recovering complete gene set for each microorganism present in these ecosystems represents an important if not essential step towards this objective (Quince et al. 2017). Supported by dropping costs of high-throughput sequencing technologies and backed by increasingly powerful computational resources, the field of metagenomics aims at exploring ecosystems through the analysis of DNA sequences extracted directly from the environment, to gain insights on the diversity of microbial population and their dynamics (Alberti et al. 2017; Hug et al. 2016). Characterizing complete or near complete genomes remain however difficult to tackle and dependent on the popularity of the ecosystem studied and the amount of data generated (Olson et al. 2017). The development of new metagenome binning techniques, that aims at solving this limitation,

has therefore accompanied the development of metagenomics studies in recent years (Albertsen et al. 2013; Alneberg et al. 2014; Frank et al. 2016).

Most computational approaches rely on the composition and/or co-abundance of sequences recovered from multiple samples to pool (bin) them together (Wu et al. 2014; Laczny et al. 2017; Nielsen et al. 2014; Alneberg et al. 2014; Kang et al. 2015). Composition based methods group together sequences that display similar metrics, such as GC content and/or tetra- and/or penta-nucleotides frequencies. Co-abundance approaches trace the relative amount of sequences over multiple samples and group together those with similar coverage variation. Co-abundance is very effective when multiple samples of the same ecosystem are available under different conditions. Nowadays, however, most metagenomics binning pipeline consists in hybrid approaches that combine these two strategies to improve the confidence of the resulting sequences bins (Kang et al. 2015; Alneberg et al. 2014; Wu et al. 2014). Some caveats and limitations remain. First, grouping sequences based on their composition implies a strong assumption regarding the genomes themselves, namely that they are relatively uniform with respect to the metric used to bin their constituent sequences. This hypothesis, though often reliable, is not valid when horizontal transfer or introgression of genetic material take place between species with (highly) divergent sequence compositions. The GC content of prophages and of their host bacterial genomes can differ, impairing the efficiency of sequence composition based binning approaches (Edwards et al. 2016; Arndt et al. 2016). In addition, co-abundance based methods require multiple samples to be fully effective, which can be impractical and/or costly. Moreover, these methods generally encounter problems with small contigs (<1,000 bp) limiting the exploration of genomic diversity whereas reaching at a comprehensive characterization of this diversity is a prerequisite to understand the dynamics underlying the network of interactions found within communities.

Novel technologies such as single-cell sequencing (Ji et al. 2017), long reads (Frank et al. 2016) or proximity ligation/chromosome conformation capture (3C) (reviewed in Marbouty and Koszul 2015; Flot, Marie-Nelly, and Koszul 2015), hold the potential to address some of these limitations. The latter approach, dubbed meta3C from the original 3C approach (Dekker et al. 2002), aims at quantifying and exploiting collisions between the DNA loci over a population of species to identify those that share the same cellular compartment. Sequences belonging to the same genome display enriched contact frequencies than those belonging to different genomes, as demonstrated by applying meta3C on controlled mixes of species (Burton 2014; Beitel 2014; Marbouty et al. 2014). Besides controlled mixes, meta3C successfully reconstructed genomes from truly unknown complex ecosystems as well (Marbouty et al. 2014; Stewart et al. 2018). Not only near complete genomes from microorganisms can be recovered from a single experiment, but additional information about the

genomic structure of these microbial populations can be recovered as well, including plasmid (Marbouty et al. 2014; Press et al. 2017) and phage-host infection spectrum (Marbouty et al. 2017). These studies suggest that meta3C (and derivate approaches) hold the potential to 1) accurately bin genomes and episomal DNA molecules and 2) assign episomal DNA molecules to their respective hosts. However, comprehensive, end-to-end computational pipelines to process raw meta3C datasets remain sparse (DeMaere and Darling 2019; Marbouty et al. 2017). Most analyses so far have focused on single mock communities, and quantifiable metrics are lacking to see how meta3C-like approaches truly compare – and possibly complement – traditional binning methods, notably regarding the quality, completeness and accuracy.

To address this need we developed MetaTOR (Metagenomic Tridimensional Organisation-based Reassembly), a lean and scalable tool to investigate single or multiple proximity ligation metagenomics experiments, from raw 3C reads to bins. MetaTOR was applied on meta3C libraries of mouse gut samples collected over time. This first dynamic meta3C study allowed us to reconstruct a high number of complete genome sequences, and to compare the genomic bins recovered using MetaTOR with bins generated by the state-of-the-art binning software MetaBat (Kang et al. 2015) and CONCOCT ((Alneberg et al. 2014). In each test case, MetaTOR compared favourably with the two aforementioned programs, both regarding the number of nearly complete genomes recovered and the amount of sequences binned. Therefore, 3C/Hi-C based metagenomic binning is a robust solution when seeking to reconstruct a comprehensive picture of a whole microbial community found in various microbial populations, regardless the number of samples processed.

2. Materials and methods

2.1. Faeces sampling and meta3C libraries generation

The faeces of three groups of two mice were sampled during twenty days as follow: day 2, 5 and 9 for cage n°1; day 2, 4, 5, 6, 7, 9, 10, 12, 16 for cage n°2; day 2, 5, 6, 7, 9, 11, 12, 16 for cage n°3. The samples were immediately suspended after collection in 30 mL of 1X tris-EDTA buffer supplemented with formaldehyde at a final concentration of 3%, then fixated for one hour under shaking. 10 mL of glycine 2.5 M was added to the mix for quenching during 20 min. A centrifugation recovered the resulting pellet for −80°C storage and awaiting further use. The libraries were then prepared and sequenced using pair-end (PE) Illumina sequencing (2 × 75 bp NextSeq) as described (Marbouty et al. 2014).

The Institut Pasteur ethics committee (CETEA) approved all the experiments performed on mice (Project dha170005).

2.2. Read processing and assembly

The first 10 bp of each read, corresponding to custom-made amplification primers, were removed, and the resulting 65 pb sequences were filtered and trimmed using cutadapt (Martin 2011). Quality was controlled with fastqc and a total of 813 million PE reads were kept in total (over the 20 samples). Reads were then used to perform three independent assemblies using MEGAHIT (Li et al. 2015) with default parameters. Contigs under 500 bp were discarded from further analyses.

2.3. Alignment step and network generation

Reads were aligned independently in single-end mode using bowtie2 (option –very-sensitive-local). For each sample, both alignment files were sorted and merged using the samtools and pysam libraries. Alignments with mapping quality under 20 were discarded. All pairs of reads for which both reads of the pair aligned on two different contigs were kept to generate the network. Contigs were considered as nodes, and the values of the edges (*i.e.* the weight) of the network were determined by counting the number of non-ambiguous alignments bridging two different contigs. Normalization is computed by dividing the edge value by the geometric mean of the nodes' coverage (*i.e.* contigs' coverage). Contig coverage was calculated using MetaBat script: jgi_summarize_bam_contig_depths with a contig size limit of 500 bp for every set of reads.

2.4. Louvain clustering

We showed before that the updated implementation of the Louvain community method provided in (Blondel et al. 2008) was a promising approach to identify subnetworks of contigs in the meta3C network that display enriched contacts between themselves (Marbouty et al. 2014). The Louvain algorithm was run 400 times on each network, using the classical Newman-Girvan criterion. Nodes that systematically clustered together for each of the first 100 iterations were pooled together in Core Communities (CCs), as described previously (Marbouty et al. 2017).

2.5. Bin and genome validation

CCs above 500 kb were evaluated for completeness and contamination using CheckM (Parks et al. 2015). CheckM was also used to assign taxa to these sequences. A CC was validated as a bin if its contamination rate range under 20%. Among those validated bins, a bin is said to be highly complete if it is at least 95% complete and no more than 5% contaminated, nearly complete if it is 90% complete

and less than 10% contaminated and partially complete if it is between 70% and 90% complete and below 10% of contamination.

2.6. Recursive Louvain Clustering

Partially complete CCs (> 70% completion) with high levels of contamination (> 20% contamination) were selected for recursive binning. Briefly, the partition step was re-run 10 times on these contaminated CCs (*i.e.* on their corresponding sub-network), yielding groups of smaller core communities (*i.e.* sub-CCs) that are then re-processed in the binning step to assess for their quality.

2.7. Pipeline comparison

CONCOCT and MetaBat 1 were run on the same set of reads and assemblies, using the different time samples for differential coverage. The resulting bins above 500 kb were retrieved and compared with MetaTOR's for completeness and contamination using CheckM. CONCOCT was run with the following parameters --r 65 -s 100 -k 6. Metabat 1 was run with default parameters.

3. Results

3.1. Algorithmic principles underneath the MetaTOR pipeline

MetaTOR (https://github.com/koszullab/metaTOR) aims at providing the most accurate overview of genome content of a population, starting from as little as one meta3C library, while taking full advantage as the availability of more libraries if possible. It is structured around four main steps: alignment, partition, annotation and binning (Figure 1). MetaTOR was purposely designed to maintain a high level of modularity and flexibility, so that users can supply their own intermediary inputs and tweak parameters to their liking at every step, to save time and resources. If starting from the raw data, all needed is the meta3C pair-end (PE) files and an assembly of the microbial community obtained either directly from the meta3C reads (as described in Marbouty et al. 2017, 2014) or from a DNA library generated independently (Figure 1A).

- **[align]** (Figure 1B): first, meta3C reads are aligned independently along the contigs of the metagenome assembly using bowtie2 (as aligners tend to leave out far-off alignments when run in pair-end mode). Contigs are then sorted, filtered for mapping quality and merged into a global alignment file. The alignment is converted into a contact network stored in a plain

text file [network.txt: column 1 – node 1 / column 2 – node 2 / column 3 - weight] to facilitate further third-party analyse. In the network, each node represents one contig and each edge (a.k.a. weight) represents the contacts score found between two contigs. This step integrate modifiable parameters such as enforcing a lower size limits for contigs or a normalization step. A normalization of the network typically use the coverage of the contigs, but other normalization can be implemented as well.

- **[partition]** (Figure 1C): an iterative Louvain procedure is applied on the network file to partition the network into groups of contigs that consistently cluster together, *i.e.* "see" each other's in space more often than their neighbors' (Marbouty et al. 2015, 2017; DeMaere and Darling 2016; Blondel et al. 2008). These clusters, or "core communities" (CC) constitute the matrix of the metagenomic binning. The number of iterations is a free parameter of the pipeline, but the number of groups stabilizes after a while with small oscillations around a fixed value.
- **[binning]** (Figure 1D): CCs are then extracted (Fasta file) and their gene content is assessed for completeness and contamination using CheckM (Parks et al. 2015). In parallel, the pipeline also extracts sub-networks for each CCs (*i.e.* network between the contigs that composed each CC) Extraction of each sub-network allows the user to perform, if needed, a recursive procedure at this step on the defined groups of contigs (*i.e.* CCs) (see Figure 1 – "*recursive procedure*"). Indeed, some CCs exhibit a high rate of completion but also a high degree of contamination suggesting that they may contain several genomes. By applying, the partition step only on their corresponding sub-network, it becomes possible to **re**-partitionnate this CCs into smaller ones (*i.e.* sub-CCs) that could present better CheckM statistics. Indeed, the partition step using the Louvain algorithm can be applied on any network provided by the user. This step generally results in breaking down the most contaminated CCs into smaller, low-contaminated sub-CCs. The retrieved sub-CCs can also be evaluated using CheckM and validated as bins.
- **[annotation]** (Figure 1F): the metagenome assembly is annotated. Gene prediction is performed using Prodigal (Hyatt et al. 2010) and genes of interest are detected using various HMM models publicly available (Albertsen et al. 2013; Guglielmini et al. 2014; Grazziotin, Koonin, and Kristensen 2017). However, this step is independent from the other ones and allow users to introduce any type of annotation tool in the pipeline.

Overall, MetaTOR generates a set of metagenomic annotated bins and their corresponding fasta sequences (in addition to the contact network) (Figure 1E).

3.2. Construction of meta3C libraries and generation of metagenomes assemblies

To validate and compare the pipeline to classical metagenomics binning algorithm, we investigated the gut microbiota of various mice using meta3C libraries. Faeces were sampled from three group of two mice from the Institut Pasteur animal facility, over 20 days (Materials and Methods) (Figure 2). Twenty meta3C libraries (3 from cage n°1, 9 from cage n°2 and 8 from cage n°3 were then generated as described (Marbouty et al. 2017) (Materials and Methods) using HpaII as restriction enzyme. The libraries were sequenced using PE Illumina 2x75 bp kits (Table 1) (NCBI BioProject SUB5459608). After trimming and quality filtering, between 25 and 100 millions of PE reads were recovered for each samples (~813 million PE reads in total).

Meta3C sequences can be directly used to generate a *de novo* assembly without notable increase of false/chimeric contigs (Marbouty et al. 2014). Three assemblies (1, 2, and 3) using reads collected from cage 3/day 2, cage 3/all samples and all cages/all samples, respectively, were generated using MEGAHIT (Li et al. 2015) (Materials and Methods). After discarding contigs below 500 bp, these three assemblies generated 61,600, 167,810, and 237,868 contigs for a cumulated sizes of 146, 475 and 763 Mb, respectively (Table 2). These three assemblies and their corresponding set of reads were then used to test various binning pipeline (MetaTOR, MetaBAT and CONCOCT) and their output (Material and Methods).

3.3. Binning of metagenomes using MetaTOR

PE meta3C reads from each of the libraries used to generate the three assemblies were aligned independently on their respective assembly to retrieve pairs of reads with each end aligning on a different contig (parameters: MQT = 20; contigs size limit = 500 bp). Contact scores between contigs where then normalized by dividing the weight of each contact by the root square of the product of the coverage of each contigs involved in the interaction. This step generates networks of weighted connections bridging contigs of the different assemblies (Table 3). These three datasets were used for further analysis. The next steps of the MetaTOR pipeline are illustrated for the assembly n°3 and its corresponding network in the Figure 3. Each network was partitioned into Core Communities (CCs) through iterative Louvain partitioning. After ~100 cycles the number of large CCs (> 500Kb) reached a plateau for the three networks and this number of iterations was retained to recover the CCs (Figure 3A). The resulting “reordering” of contacts matrix (Figure 3B) showed a low level of noise between the different CCs, suggesting Louvain successfully clustered together contigs displaying preferential contacts with each other’s.

The gene content of CCs containing more than 500 kb of sequences, corresponding to 17, 33 and 125 CCs for the 3 different datasets, was analysed using CheckM (Parks et al. 2015). The analysis showed that most of these CCs display completion and contamination levels above 80% and under 10 %, respectively (Figure 3C), suggesting that they contain nearly complete bacterial genomes and could therefore be annotated as validated bins or MAGs (metagenome-assembled genomes). However, a subset of CCs displayed more than 80% of completion but also more than 20% of contamination (respectively 4, 24 and 25 for assemblies 1, 2 and 3, respectively), with some exhibiting a contamination rate as high as 1,000% (Figure 3C). These CCs were processed through a recursive procedure: 10 extra Louvain clustering steps performed on their respective sub-network partitioned them into sub-CCs (Figure 3D). A CheckM analysis of these sub-CCs showed that they often display high quality signatures of bacterial genomes, suggesting that the large, contaminated CCs correspond to mixes of near complete bacterial genomes (Figure 3F). Regarding the assembly n°3, the overall procedure generated 1,001 bins (bins > 10 kb – 724 Mb in total) with 269 bins containing more than 500 kb of sequences each and representing 687 Mb of sequences (90% of the filtered assembly) (Figure 3E). Results obtained after the recursive procedure show a clear decrease in term of contamination (mean value decrease from 61.4 % to 1.9 %). This improvement was accompanied by a slight, but acceptable, loss of completion compared to results obtained without (mean value decreases from 88.4 % to 61.1 %) and validates the application of an iterative procedure on the largest, contaminated CCs. Among the characterized bins for the assembly n°3 92 represent highly complete MAGs (< 5% contamination and >= 95% complete, 31 near-complete MAGs (=< 10% contamination and >= 90% complete) and 33 were substantially complete MAGs (=< 10% contamination and >= 70% complete). The final results for the three datasets are presented in Table 3.

3.4. Comparison of MetaTOR with hybrid binning algorithm

To evaluate how MetaTOR compares to established and popular binning approaches, we also ran MetaBAT (versions 1) (Kang et al. 2015) and CONCOCT (Alneberg et al. 2014) on assemblies 1, 2 and 3 using the same filtered PE reads, allowing each pipeline to take advantage of the information from differential coverage across the independent experiments when several samples were used. The results were then compared with the ones generated by MetaTOR (Figure 4 and Table 3). For assemblies 1, 2 and 3, MetaTOR retrieved 16, 61 and 123 nearly complete MAGs, respectively (>90% completion, <10% contamination), compared to 11, 43, and 87 with MetaBAT and 5, 37 and 85 with CONCOCT. Overall, MetaTOR resulted in more high-completeness, low-contamination bins than all tested pipelines. In each case, the number of highly complete genomes (>95% completion, <5% contamination) recovered was equal or higher when using 3C data and our clustering approach. The

difference was even more pronounced when using the 20 libraries as MetaTOR was able to retrieved 123 near complete genomes representing 426 Mb and corresponding to 55% of the total assembly. The mean contamination rate of bins characterized by MetaTOR was also lower than the two other approaches regardless the number of libraries used. Finally, MetaTOR also allowed to recover more bins and assigned bigger amount of sequences than the two other approaches.

4. Discussion

MetaTOR is a lean and scalable pipeline that exploits metagenomic proximity ligation experiments (meta3C) to partition the resulting contigs into individual genomes according to their 3D contact frequencies. We and others showed previously that meta3C reads can be use on a single sample with very good results compared to other binning methods (Press et al. 2017; Marbouty et al. 2017; DeMaere and Darling 2019). In the present study, we extend our original analysis algorithm on multiple datasets to evaluate its efficiency and compare it to classical binning methods. Compared to state-of-the-art binning methods, MetaTOR retrieves more complete MAGs, with significantly lower contamination rates. Therefore, physical collisions between DNA sequences represent an objective, quantitative way, to cluster these molecules together, compared to indirect, commonly used approaches involving correlations between sequence composition or abundancy co-variation. This was true even when 20 independent experiments were used, highlighting the interest to include at least some meta3C experiments in planned future metagenomics projects, and this regardless of the number of planed libraries.

The large networks derived from our different meta3C experiments contain a certain number of highly connected sub-networks poorly connected to each other. These kind of highly modular networks are known to be well-suited to community detection algorithm like Louvain (Blondel et al. 2008). Moreover, the ‘iterative Louvain’ procedure allows us to identify sets of sequences that contact each other’. However, there are limits to the current iterative Louvain implementation. First, all modularity optimization algorithms tend to over-cluster nodes when the network reaches a certain size threshold, regardless of the underlying patterns. This well-documented property is known as the ‘resolution limit’ (Fortunato and Barthélemy 2007). However, it can be sidestepped by running the partitioning process recursively on the network corresponding to the studied sub-network. Since it should be comparatively small and under the scale at which the aforementioned limit becomes visible, the clusters found inside will separate again and yield bins as normal. The recursive procedure appears

as highly effective with a clear increase in the number of nearly complete MAGs retrieved. The second limit comes from our stringent nature of bin definition, as we only retain sequences that always, systematically cluster together. As a result, a unique 'jump' of a contig outside the main cluster during one of the iterations has for consequence to exclude this contig from the final bin. While this ensures a reduction in bin contamination, a number of meaningful sequences are likely to be excluded from the bin. For instance, mobile elements (e.g. phage or plasmids) shared by different genomes will most likely be excluded from the corresponding bins. However, using MetaTOR and annotation pipelines such as VirSorter or PlasFlow, this limitation can be overcome to investigate and infer *a posteriori* the hosts of these elements using the contact network (Marbouty et al. 2017) or the Louvain clustering score (computed from the iterative procedure, and corresponding to the number of times two CCs are grouped together). A detailed analysis of so-called overlapping communities (Wang et al., 2012) would be very useful in the future to study such associations and bring a new tool in the study of interactions between genomic entities in microbial communities.

Our pipeline is flexible and, though we developed it taking advantage of the Louvain algorithm (Blondel et al. 2008), other clustering algorithms yielding to nondeterministic community identifiers (e.g. a community detection algorithm with a different modularity) can be used instead with no side effects on the rest of the pipeline.

Proximity ligation assays were originally developed to capture the 3D folding of microbial or metazoan chromosomes (Dekker et al. 2002; Lieberman-Aiden et al. 2009). Proximity ligation approaches were developed as a side derivative of this original purposes, and applied to various genomics limitations including chromosome scaffolding (Marie-Nelly, Marbouty, Cournac, Flot, et al. 2014; Burton 2014; Kaplan and Dekker 2013), haplotype reconstruction (Selvaraj et al. 2013), and centromere prediction/genome annotation (Marie-Nelly, Marbouty, Cournac, Liti, et al. 2014), besides metagenomic binning. Haplotype phasing is an especially interesting application since strains of the same species are remain challenging to characterize. Improving the resolving power of 3C-based methods by combining metagenomics with haplotype approaches could help address this limitation. Future work should therefore involve back-and-forth interaction between wet and *in silico* experiments.

5. Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6. Data accession

The datasets generated for this study can be found on SRA database : BioProject SUB5459608.

7. Author Contributions

MM and RK conceived the study. LB, TFR and MM wrote the pipeline MetaTOR. MM, TFR and AT performed the experiments. LB, TFR, MM and RK analyzed and interpreted the data. LB, TFR, MM and RK wrote the manuscript.

Funding

Lyam Baudry is supported by an AMX fellowship from the French Ministry of Higher Education, Research and Innovation. Théo Foutel-Rodier is supported by an ENS fellowship by the French Ministry of Higher Education, Research and Innovation. This research was supported by funding to R.K. from the European Research Council under the Horizon 2020 Program (ERC grant agreement 260822) and from the Agence Nationale pour la Recherche (JPI-EC-AMR STARCS ANR-16-JPEC-0003-05).

8. Acknowledgments

We thank Corinne Fayolle and Xavier Montagutelli for their help in the sampling process. We thank our colleagues from the lab for discussions, feedback and comments on MetaTOR.

9. References

Alberti, Adriana, Julie Poulain, Stefan Engelen, Karine Labadie, Sarah Romac, Isabel Ferrera, Guillaume Albini, et al. 2017. "Viral to Metazoan Marine Plankton Nucleotide Sequences from the Tara Oceans Expedition." *Scientific Data* 4: 170093. https://doi.org/10.1038/sdata.2017.93.

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L. Nielsen, Gene W. Tyson, and Per H. Nielsen. 2013. "Genome Sequences of Rare, Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes." *Nature Biotechnology* 31 (6): 533–38. https://doi.org/10.1038/nbt.2579.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. "Binning Metagenomic Contigs by Coverage and Composition." *Nature Methods* 11 (11): 1144–46. https://doi.org/10.1038/nmeth.3103.

Arndt, David, Jason R. Grant, Ana Marcu, Tanvir Sajed, Allison Pon, Yongjie Liang, and David S. Wishart. 2016. "PHASTER: A Better, Faster Version of the PHAST Phage Search Tool." *Nucleic Acids Research* 44 (W1): W16-21. https://doi.org/10.1093/nar/gkw387.

Beitel, C. W.; Froenicke, L.; Lang, J. M.; Korf, I. F.; Michelmore, R. W.; Eisen, J. A.; Darling, A. E. 2014. "Strain- and Plasmid-Level Deconvolution of a Synthetic Metagenome by Sequencing Proximity Ligation Products." *PeerJ* 2: e415. https://doi.org/10.7717/peerj.415.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Burton, J. N.; Liachko, I.; Dunham, M. J.; Shendure, J. 2014. "Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps." In *G3 (Bethesda)*, 4:1339–46. 7. https://doi.org/10.1534/g3.114.011825.

Cho, Ilseung, and Martin J. Blaser. 2012. "The Human Microbiome: At the Interface of Health and Disease." *Nature Reviews. Genetics* 13 (4): 260–70. https://doi.org/10.1038/nrg3182.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. "Capturing Chromosome Conformation." *Science (New York, N.Y.)* 295 (5558): 1306–11. https://doi.org/10.1126/science.1067799.

DeMaere, Matthew Z., and Aaron E. Darling. 2019. "Bin3C: Exploiting Hi-C Sequencing Data to Accurately Resolve Metagenome-Assembled Genomes." *Genome Biology* 20 (1): 46. https://doi.org/10.1186/s13059-019-1643-1.

Edwards, Robert A., Katelyn McNair, Karoline Faust, Jeroen Raes, and Bas E. Dutilh. 2016. "Computational Approaches to Predict Bacteriophage-Host Relationships." *FEMS Microbiology Reviews* 40 (2): 258–72. https://doi.org/10.1093/femsre/fuv048.

Flot, Jean-François, Hervé Marie-Nelly, and Romain Koszul. 2015. "Contact Genomics: Scaffolding and Phasing (Meta)Genomes Using Chromosome 3D Physical Signatures." *FEBS Letters* 589 (20 Pt A): 2966–74. https://doi.org/10.1016/j.febslet.2015.04.034.

Fortunato, Santo, and Marc Barthélemy. 2007. "Resolution Limit in Community Detection." *Proceedings of the National Academy of Sciences* 104 (1): 36–41. https://doi.org/10.1073/pnas.0605965104.

Frank, J. A., Y. Pan, A. Tooming-Klunderud, V. G. H. Eijsink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope. 2016. "Improved Metagenome Assemblies and Taxonomic Binning Using Long-Read Circular Consensus Sequence Data." *Scientific Reports* 6 (May): 25373. https://doi.org/10.1038/srep25373.

Grazziotin, Ana Laura, Eugene V. Koonin, and David M. Kristensen. 2017. "Prokaryotic Virus Orthologous Groups (PVOGs): A Resource for Comparative Genomics and Protein Family Annotation." *Nucleic Acids Research* 45 (D1): D491–98. https://doi.org/10.1093/nar/gkw975.

Guglielmini, Julien, Bertrand Néron, Sophie S. Abby, María Pilar Garcillán-Barcia, Fernando de la

Cruz, and Eduardo P. C. Rocha. 2014. “Key Components of the Eight Classes of Type IV Secretion Systems Involved in Bacterial Conjugation or Protein Secretion.” *Nucleic Acids Research* 42 (9): 5715–27. https://doi.org/10.1093/nar/gku194.

Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, et al. 2016. “A New View of the Tree of Life.” *Nature Microbiology* 1 (5): 16048. https://doi.org/10.1038/nmicrobiol.2016.48.

Hyatt, Doug, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (March): 119. https://doi.org/10.1186/1471-2105-11-119.

Ji, Peifeng, Yanming Zhang, Jinfeng Wang, and Fangqing Zhao. 2017. “MetaSort Untangles Metagenome Assembly by Reducing Microbial Community Complexity.” *Nature Communications* 8: 14306. https://doi.org/10.1038/ncomms14306.

Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. “MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities.” *PeerJ* 3: e1165. https://doi.org/10.7717/peerj.1165.

Kaplan, Noam, and Job Dekker. 2013. “High-Throughput Genome Scaffolding from in-Vivo DNA Interaction Frequency.” *Nature Biotechnology* 31 (12): 1143–47. https://doi.org/10.1038/nbt.2768.

Laczny, Cedric C., Christina Kiefer, Valentina Galata, Tobias Fehlmann, Christina Backes, and Andreas Keller. 2017. “BusyBee Web: Metagenomic Data Analysis by Bootstrapped Supervised Binning and Annotation.” *Nucleic Acids Research* 45 (W1): W171–79. https://doi.org/10.1093/nar/gkx348.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics (Oxford, England)* 31 (10): 1674–76. https://doi.org/10.1093/bioinformatics/btv033.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science (New York, N.Y.)* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Marbouty, Martial, Lyam Baudry, Axel Cournac, and Romain Koszul. 2017. “Scaffolding Bacterial Genomes and Probing Host-Virus Interactions in Gut Microbiome by Proximity Ligation (Chromosome Capture) Assay.” *Science Advances* 3 (2). https://doi.org/10.1126/sciadv.1602105.

Marbouty, Martial, Axel Cournac, Jean-François Flot, Hervé Marie-Nelly, Julien Mozziconacci, and Romain Koszul. 2014. “Metagenomic Chromosome Conformation Capture (Meta3C) Unveils the Diversity of Chromosome Organization in Microorganisms.” *ELife* 3 (December): e03318. https://doi.org/10.7554/eLife.03318.

Marbouty, Martial, and Romain Koszul. 2015. “Metagenome Analysis Exploiting High-Throughput Chromosome Conformation Capture (3C) Data.” *Trends in Genetics: TIG* 31 (12): 673–82. https://doi.org/10.1016/j.tig.2015.10.003.

Marie-Nelly, Hervé, Martial Marbouty, Axel Cournac, Jean-François Flot, Gianni Liti, Dante Poggi Parodi, Sylvie Syan, et al. 2014. “High-Quality Genome (Re)Assembly Using Chromosomal Contact Data.” *Nature Communications* 5 (December). https://doi.org/10.1038/ncomms6695.

Marie-Nelly, Hervé, Martial Marbouty, Axel Cournac, Gianni Liti, Gilles Fischer, Christophe Zimmer, and Romain Koszul. 2014. “Filling Annotation Gaps in Yeast Genomes Using Genome-Wide Contact Maps.” *Bioinformatics (Oxford, England)* 30 (15): 2105–13. https://doi.org/10.1093/bioinformatics/btu162.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.

Nielsen, H. Bjørn, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, et al. 2014. “Identification and Assembly of Genomes and Genetic Elements in Complex Metagenomic Samples without Using Reference Genomes.” *Nature*

Biotechnology 32 (8): 822–28. https://doi.org/10.1038/nbt.2939.

Olson, Nathan D., Todd J. Treangen, Christopher M. Hill, Victoria Cepeda-Espinoza, Jay Ghurye, Sergey Koren, and Mihai Pop. 2017. "Metagenomic Assembly through the Lens of Validation: Recent Advances in Assessing and Improving the Quality of Genomes Assembled from Metagenomes." *Briefings in Bioinformatics*, August. https://doi.org/10.1093/bib/bbx098.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. "CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes." *Genome Research* 25 (7): 1043–55. https://doi.org/10.1101/gr.186072.114.

Philippot, Laurent, Jos M. Raaijmakers, Philippe Lemanceau, and Wim H. van der Putten. 2013. "Going Back to the Roots: The Microbial Ecology of the Rhizosphere." *Nature Reviews. Microbiology* 11 (11): 789–99. https://doi.org/10.1038/nrmicro3109.

Press, Maximilian O., Andrew H. Wiser, Zev N. Kronenberg, Kyle W. Langford, Migun Shakya, Chien-Chi Lo, Kathryn A. Mueller, Shawn T. Sullivan, Patrick S. G. Chain, and Ivan Liachko. 2017. "Hi-C Deconvolution of a Human Gut Microbiome Yields High-Quality Draft Genomes and Reveals Plasmid-Genome Interactions." *BioRxiv*, October, 198713. https://doi.org/10.1101/198713.

Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. "Shotgun Metagenomics, from Sampling to Analysis." *Nature Biotechnology* 35 (9): 833–44. https://doi.org/10.1038/nbt.3935.

Selvaraj, Siddarth, Jesse R Dixon, Vikas Bansal, and Bing Ren. 2013. "Whole-Genome Haplotype Reconstruction Using Proximity-Ligation and Shotgun Sequencing." *Nature Biotechnology* 31 (12): 1111–18. https://doi.org/10.1038/nbt.2728.

Stewart, Robert D., Marc D. Auffret, Amanda Warr, Andrew H. Wiser, Maximilian O. Press, Kyle W. Langford, Ivan Liachko, et al. 2018. "Assembly of 913 Microbial Genomes from Metagenomic Sequencing of the Cow Rumen." *Nature Communications* 9 (1): 870. https://doi.org/10.1038/s41467-018-03317-6.

Wu, Yu-Wei, Yung-Hsu Tang, Susannah G. Tringe, Blake A. Simmons, and Steven W. Singer. 2014. "MaxBin: An Automated Binning Method to Recover Individual Genomes from Metagenomes Using an Expectation-Maximization Algorithm." *Microbiome* 2 (1): 26. https://doi.org/10.1186/2049-2618-2-26.

10. Figures legends

Figure 1: MetaTOR pipeline

Schematic representation of the main steps of the MetaTOR pipeline. **A**. MetaTOR is initialized with an assembly and a set of 3C/Hi-C PE reads. **B**. The first [align] step aligns, sorts and merges reads and deliver a network of contigs interactions. **C**. Then, the [partition] step deconvolves the defined network using a Louvain iterative procedure and **D**. [binning] allows to retrieve CCs (Fasta file and corresponding sub-network) of selected partition in order to evaluate them using CheckM. At this step, it is possible to perform a recursive procedure on selected CCs in order to partitionate them further into sub-CCs. **F**. [annotation] is an optional step that use HMM models to provide final annotations. **E**. The final output of the pipeline is a set of annotated bins.

Figure 2: Experimental design

Three groups of two mice were sampled during twenty days as follow: day 2, 5 and 9 for the cage n°1; day 2, 4, 5, 6, 7, 9, 10, 12, 16 for the cage n°2; day 2, 5, 6, 7, 9, 11, 12, 16 for the cage n°3. Samples were then processed for meta3C sequencing. The resulting sequences were used to generate *de novo* assemblies and test the different binning methods.

Figure 3: MetaTOR partitioning of a complex microbial community

A. Evolution of the number of CCs, ordered by size categories, during 400 Louvain iterations for the assembly n°3 (20 samples). Blue: CCs encompassing between 10 kb and 100 kb of sequences. Red: CCs encompassing between 100 kb and 500 kb of sequences. Green: CCs encompassing more than 500 kb of sequences. **B**. Contact matrix encompassing the 224 largest CCs ordered by size, after 100 Louvain iterations (1 pixel = 200 kb). Y-axis: cumulated DNA size. **C.** Completion (red) and contamination (blue) of the 129 CCs containing more than 500 kb of sequences after 100 Louvain iterations. Dashed lines: thresholds used to process the CCs through a recursive procedure (completion threshold: upper 70%; contamination threshold: upper 20%). **D**. Contact map of a highly contaminated CC (CC #3 – 100% complete – 1400% contaminated) before (left) and after (right) the recursive procedure (10 iterations; 1 bin: 20kb). Left map: contigs are ordered by size. Right map: sub-CCs are ordered by size. **E**. Completion and contamination of the 269 bins larger than 500 kb defined after the whole procedure. Red: completion. Blue: contamination. **F**. Completion (red) and

contamination (blue) levels of the sub-CCs retrieved from the original CC #3 treated with 10 iterations of recursive process.

Figure 4: Comparison of MetaTOR, MetaBAT and CONCOCT.

Comparison of the three methods' outputs applied on the 3 datasets tested in this work. **A**. dataset #1 (1 library). **B**. dataset #2 (8 libraries). **C**. dataset #3 (20 libraries). Box plot of completion (left), box plot of contamination (middle) and histogram of retrieved MAGs (right) are presented for the three methods. Only MAGs over 500 kb are analyzed (thresholds used to draw the histogram: dark red: 95% completion – 5% contamination; red : 90% completion – 10% contamination; orange: 70% completion – 10% contamination; yellow: other MAGs).

11. Tables

Table 1: Meta3C libraries constructed and sequenced

sample	raw paired-end reads
cage1-day 1	79 868 626
cage1-day2	38 728 350
cage1-day3	33 173 429
cage2-day1	40 380 356
cage2-day2	62 424 123
cage2-day3	31 436 086
cage2-day4	34 124 320
cage2-day5	48 472 570
cage2-day6	36 129 310
cage2-day7	32 608 370
cage2-day8	43 473 731
cage2-day9	67 768 796
cage3-day1	108 114 353
cage3-day2	39 719 377
cage3-day3	37 792 067
cage3-day4	36 805 550
cage3-day5	34 529 306
cage3-day6	59 092 136
cage3-day7	28 833 461
cage3-day8	30 521 091

Table 2: assembly metrics

Only the metrics concerning assemblies filtered for the contigs above 500bp are shown.

	PE reads (filtered)	total size (contigs > 500 bp)	contigs > 500 bp	N50 (contigs > 500 bp)
assembly #1 (cage 3 – day 2)	100,258,683	146,319,508 bp	61,666	6,176 bp
assembly #2 (cage 3 – samples x 8)	330,324,521	475,681,220 bp	167,810	7,578 bp
assembly #3 (samples x 20)	813,376,239	763,455,888 bp	237,868	12,339 bp

Table 3: Networks features

	PE reads (filtered)	mapped PE reads	intercontigs interactions	weighted interactions
assembly #1	100,258,683	67,994,798	6,457,842	1,322,003
assembly #2	330,324,521	215,768,714	30,206,795	8,505,609
assembly #3	813,376,239	541,384,131	96,546,376	77,577,924

Table 4: comparison of MetaTOR, CONCOCT and MetaBAT results.

(*near completes MAGs correspond to bins with a completion higher than 90% and a contamination lower than 10%)

		assembly #1 (148 Mb)		assembly #2 (483 Mb)		assembly #3 (763 Mb)	
		nb	size (bp)	nb	size (bp)	nb	size (bp)
MetaTOR	10 Kb < bins < 100 Kb	284	7,537,821	807	21,139,528	617	15,175,457
	100 Kb < bins < 500 Kb	43	11,319,827	144	30,749,287	106	22,963,515
	bins > 500 Kb	56	119,111,306	183	399,972,204	269	685,955,810
	near complete MAGs*	16	61,643,887	61	222,857,936	122	426,281,987
MetaBat	10 Kb < bins < 100 Kb	0	0	0	0	0	0
	100 Kb < bins < 500 Kb	18	5,703,905	55	17,583,986	65	24,087,225
	bins > 500 Kb	36	82,290,484	126	284,973,235	172	420,081,339
	near complete MAGs*	11	36,209,901	43	129,221,658	87	262,912,014
CONCOCT	10 Kb < bins < 100 Kb	11	432,808	25	1,040,872	24	1,122,733
	100 Kb < bins < 500 Kb	7	1,351,308	23	6,275,583	6	5,193,580
	bins > 500 Kb	29	120,778,514	126	412,598,588	195	673,338,423
	near complete MAGs*	5	13,959,215	37	122,970,516	85	304,517,832

Figure 1

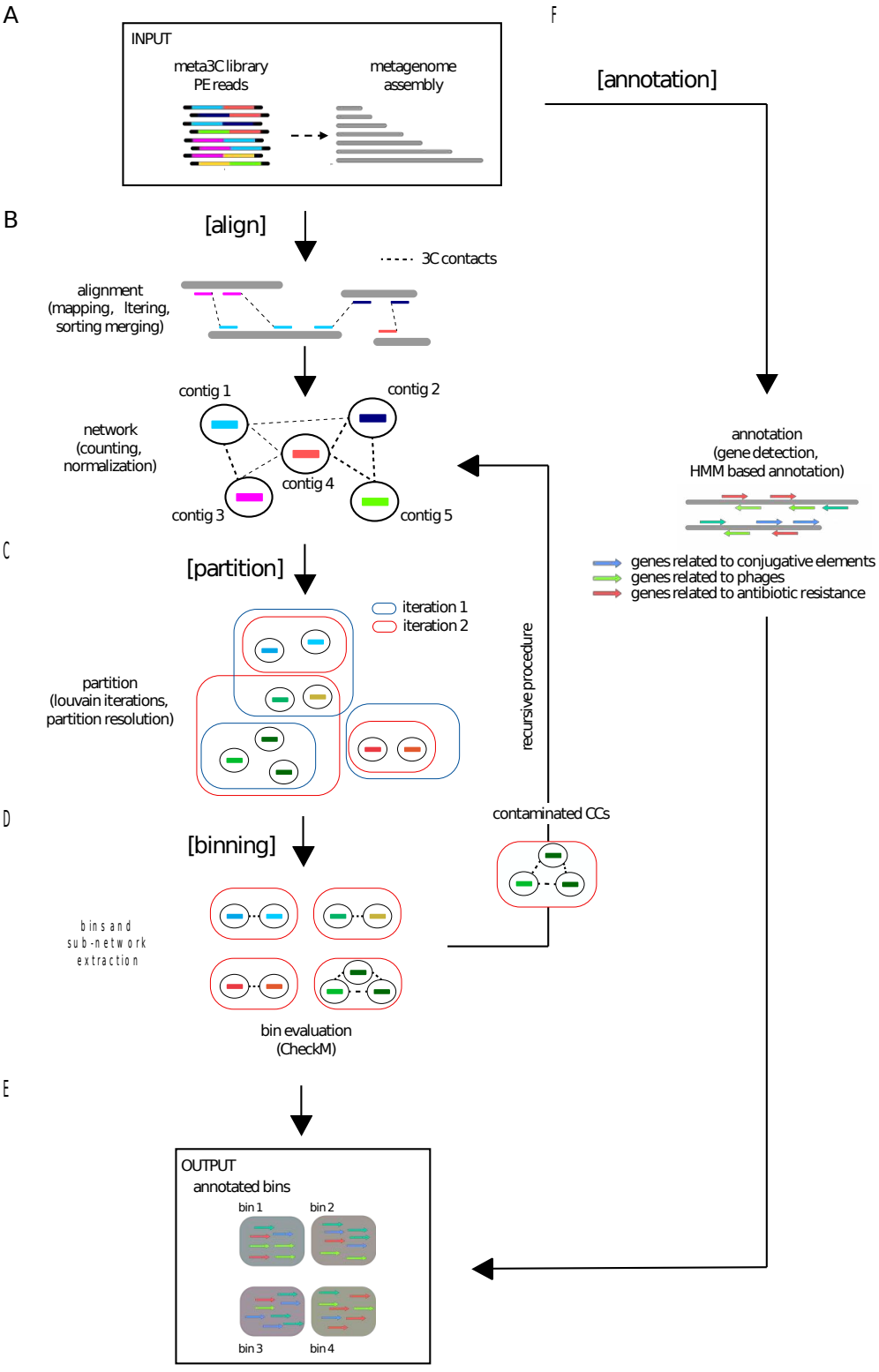

A
INPUT
meta3C library
PE reads
metagenome
assembly
B
[align]
3C contacts
alignment
(mapping, filtering,
sorting merging)
contig 1
contig 2
network
(counting,
normalization)
contig 4
contig 3
contig 5
C
[partition]
iteration 1
iteration 2
partition
(louvain iterations,
partition resolution)
recursive procedure
D
[binning]
contaminated CCs
bins and
sub-network
extraction
bin evaluation
(CheckM)
E
OUTPUT
annotated bins
bin 1
bin 2
bin 3
bin 4
F
[annotation]
annotation
(gene detection,
HMM based annotation)
genes related to conjugative elements
genes related to phages
genes related to antibiotic resistance

Figure 2

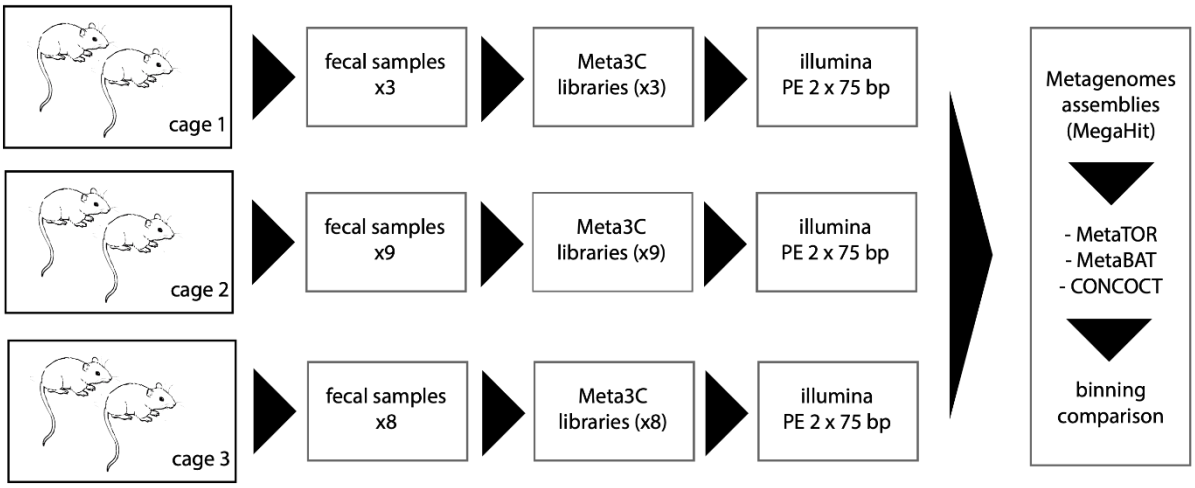

cage 1
cage 2
cage 3
fecal samples x3
fecal samples x9
fecal samples x8
Meta3C libraries (x3)
Meta3C libraries (x9)
Meta3C libraries (x8)
illumina PE 2 x 75 bp
illumina PE 2 x 75 bp
illumina PE 2 x 75 bp
Metagenomes assemblies (MegaHit)
- MetaTOR
- MetaBAT
- CONCOCT
binning comparison

Figure 3

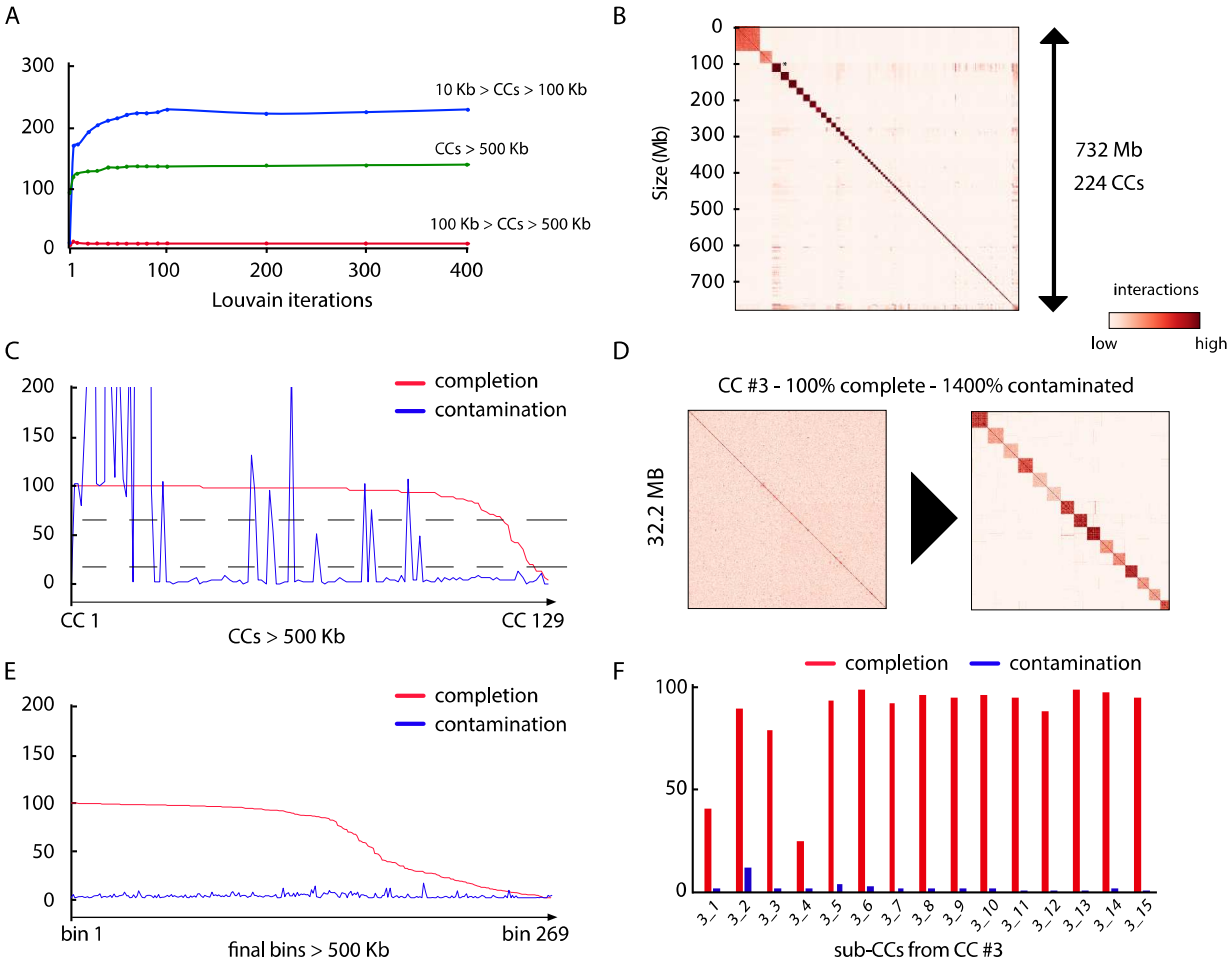
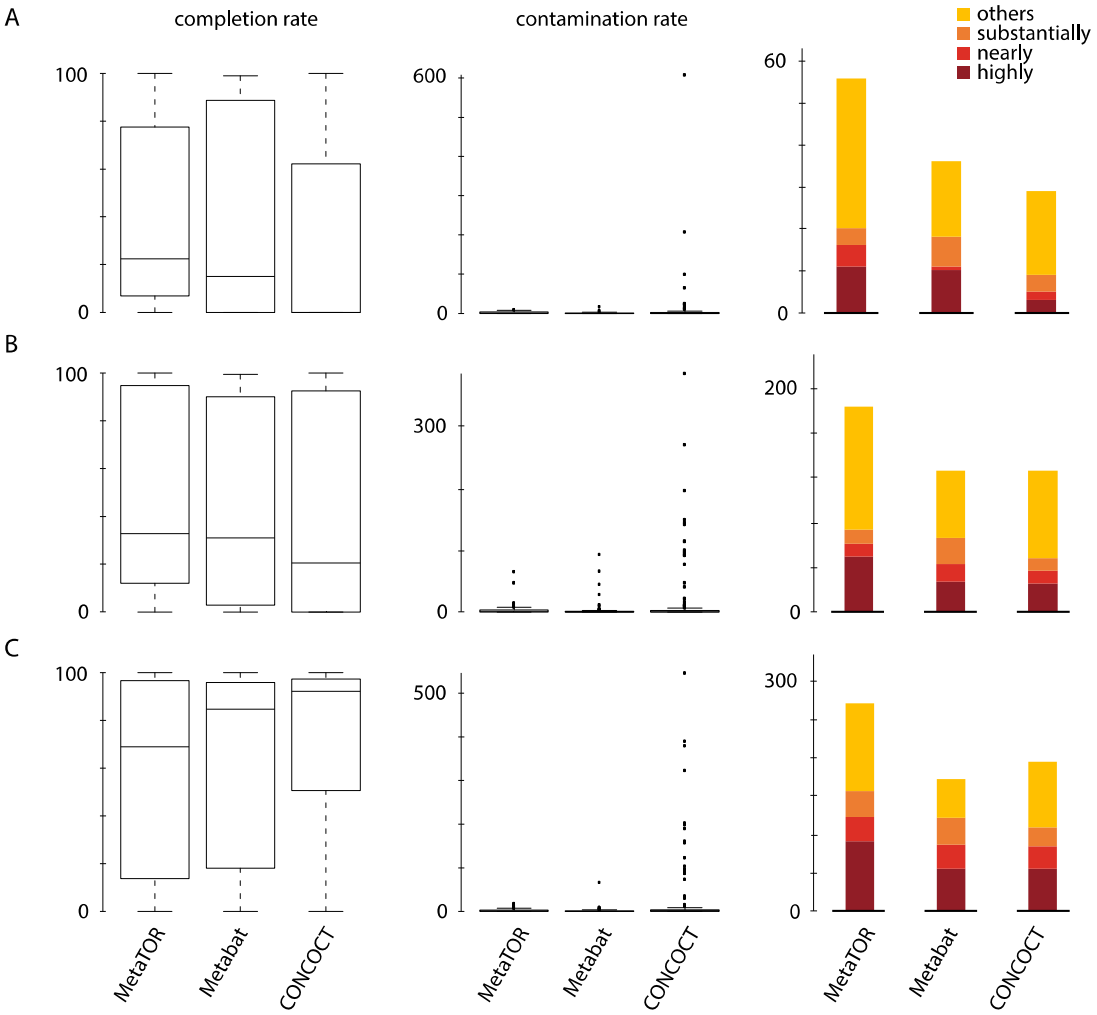

A
300
200
100
0
1
100
200
300
400
Louvain iterations
10 Kb > CCs > 100 Kb
CCs > 500 Kb
100 Kb > CCs > 500 Kb
B
0
100
200
300
400
500
600
700
Size (Mb)
732 Mb
224 CCs
interactions
low
high
C
200
150
100
50
0
completion
contamination
CC 1
CC 129
CCs > 500 Kb
D
CC #3 - 100% complete - 1400% contaminated
32.2 MB
E
200
150
100
50
0
completion
contamination
bin 1
bin 269
final bins > 500 Kb
F
completion
contamination
100
50
0
3_1
3_2
3_3
3_4
3_5
3_6
3_7
3_8
3_9
3_10
3_11
3_12
3_13
3_14
3_15
sub-CCs from CC #3

Figure 4


A
completion rate
contamination rate
others
substantially
nearly
highly
100
0
600
0
60
0
B
100
0
300
0
200
0
C
100
0
500
0
300
0
MetaTOR
Metabat
CONCOCT
MetaTOR
Metabat
CONCOCT
MetaTOR
Metabat
CONCOCT

5 Discussion and conclusion

In this work we have presented a range of results on (meta-)genome assembly and unveiled deep implications in terms of chromosome rearrangements, DNA transfers and general chromosome dynamics. In this chapter we will discuss the main results and limits of our work, the perspectives that should guide further attempts at developing the underlying biological questions, as well as the future of assembly projects.

5.1 Limits and improvements on our framework

Hi-C based genome assembly and dynamics analysis involves a number of concepts and methods borrowed from diverse fields ranging from cell biology to polymer physics to graph theory. While we cannot possibly hope to examine each individual aspect in its entirety, in this section we will briefly summarize the strengths, limits and potential improvements in each possible direction.

Hi-C based protocols The 3C, Hi-C and meta3C protocols are in constant evolution and refinement. However, as we have seen, a number of novel protocols exploiting the 3D conformation of the genome are being explored. The most prominent limitations to overcome are the following:

- Working with single cells instead of whole populations. As the ergodicity of chromatin behavior (or indeed if this property would change at all from one species or condition to another) remains an open question, it is crucial to verify that models and interpretations derived from data observed at the population level are consistent with the individual level. Current attempts at single-cell Hi-C have shown the inherent variability in chromosome dynamics and organization from cell to cell [250] [251].
- Resolution limits are set by the restriction enzymes being used and prevent a fully homogeneous, fine-grained analysis of an entire genome at every scale. Experimental efforts have been undertaken to understand what Hi-C data drawn from a fully 'unbiased' genome could look like, such as re-engineering large stretches from a yeast genome so that restriction sites are evenly and shortly spaced [263].

These theoretical constraints are compounded by practical ones, such as the cost of sequencing or expertise needed to perform the protocol. These bottlenecks could be removed as costs keep decreasing and optimized commercial kits make the technique popular and accessible to all. Expectations of results from Hi-C data are thus going to rise.

Chromosome modeling The models we have used throughout this work to represent chromosome dynamics were relatively simple and straightforward. We essentially relied on the fact that:

- Inter-chromosomal interactions are always lower than intra-chromosomal interactions.
- Intra-chromosomal interactions can be approximated with a power law function of the genomic distance.

As we have hinted at in the introduction, more sophisticated models have been introduced in order to account for the higher-order organization of chromatin folding. We have shown that our assembly framework is quite robust to deviations from the initial model, as long as the above two assumptions were satisfied.

However, a better understanding of chromosome dynamics should eventually be able to successfully integrate these mechanisms into the base model. There are a number of difficulties, as the biological features such as TADs, loops, and compartments are not easily represented by simple analytical functions; moreover, there tends to be different polymer behavior depending on the species of interest [309]. This indicates that these features are specific to each case study and can't be extrapolated for other species that are not model organisms, which most of our work focused on. Hopefully a multi-scale, unified representation of all higher-order chromosome folding levels will be elucidated, possibly through the use of alternative technology such as single-cell imaging. These techniques should sidestep the limitations that are inherent to most 3C-based protocols.

Assembly method As we have seen, there exists an number of other Hi-C based assembly methods. Most of them are based on an 'error-based' point of view, seeking to 'correct' misjoins as a human might do. On the other hand, the MCMC method we used assesses the entirety of the contact map and is able to fully explore genome space for an heuristically optimal family of solutions. It has been formally shown to eventually converge to the 'correct' genome [341], and was successfully demonstrated in practice. The combination of scaffolding and polishing from instaGRAAL lets us narrow down a whole range of high-likelihood families to the most correct one with the re-injection of initial assembly data. Since the qualities of assemblies and Hi-C libraries can be highly variable, our implementation allows for flexibility, but in the long term our approach should let us scaffold and polish an assembly from beginning to end with little to no manual intervention. This is an important draw in current assembly projects that become increasingly complex.

However, other promising approaches have been undertaken. SALSA2 [337] directly integrates Hi-C weights into the initial assembly graph, so that one ideally avoids a two-pass method (shotgun assembly followed Hi-C scaffolding). Not only does this help alleviate biases in the initial assembly graph (which Hi-C based scaffolding cannot correct), it also makes independent data integration easier, as there is no longer any need to reconcile Hi-C based scaffolding with other scaffoldings obtained from independent

sources. Instead, a fully integrated Hi-C/shotgun based assembly serves as the baseline for further scaffolding.

Integrating independent data sources Assembly projects have recently grown enormous in size, complexity and costs. When resources allow, scaffolding often needs independent validation from multiple sources, whether it be Hi-C, optical mapping, genetic maps or linked reads. To that effect, it is crucial to design a strategy reconciling these sources. Moreover, the implementation of such a method in a scalable and seamless way cannot be understated when dealing with large genomes with many chromosomes. Currently most polishing and conflict handling is done manually, with various detrimental consequences.

In our work, we have implemented a simple strategy with respect to genetic map data, which is given precedence over Hi-C data. In doing so, we acknowledge the shortcomings of Hi-C based techniques and the probabilistic nature of our framework: while there may be a number of reasons Hi-C scaffolding conflicts with pseudochromosomal structures from genetic maps, it is much more likely that any error is incurred from the limitations of Hi-C data. We note, on the other hand, that no such conflict was found in the case of *Ectocarpus sp.*, and thus the Hi-C scaffolding was in fact further validated by existing genetic maps. More work is needed to refine this strategy and attempt to integrate other types of data that is commonly used for genome scaffolding.

The advent and popularizing of long-read technology has introduced an additional data channel into assembly project pipelines. We have covered in the introduction the various ways short and long reads can be reconciled to yield high-quality hybrid assemblies, and our Hi-C framework seamlessly integrates into such pipelines. Indeed, we have successfully demonstrated its use on both short-read based reference assemblies (*Ectocarpus sp.* and *Trichoderma reesei*) and long-read based ones (*Cataglyphis hispanica*). On the other hand, interesting avenues for improvements could involve the design of long-read and short-read specific algorithms. For instance, any reference genome based polishing is going to be more prone to error if it involves a long read based reference (as opposed to a short read based one). The construction of *super-reads* mentioned in the introduction was also shown to be fruitful and presumably Hi-C contacts could be integrated into such graphs.

Assessing the correctness of rearrangements A crucial question is whether the rearrangements we uncovered through genome scaffolding can be trusted, notably in *C. hispanica* where the chromosomal fusion was relatively unexpected. The first step was to extensively validate each lineage genome, but a number of artifacts are expected to remain, among which the fusion could figure. Certainly the Hi-C data remains remarkably consistent with the fusion, and more independent data is needed to confirm it with absolute certainty. On the other hand, if a fusion did *not* occur, the abnormal, intra-like levels of contacts between both chromomsomes warrants further investigation.

The mechanism under which a fusion could occur remains an open question as well. Our scaffolding has confirmed that all chromosomes of *C. hispanica* were acrocentric,

and presumably two centromeres located on far ends of chromosomes could have merged in a Robertsonian translocation. More work is needed in order to understand the precise mechanism under which two chromosomes could become linked, and whether this structural change could underlie the reproduction strategy of *C. hispanica*.

Metagenomics The metagenome assembly and binning method we have put forward has allowed the reconstruction of hundreds of genomes, but so did traditional binning approaches on the same datasets we used. However, these do not necessarily overlap; in that respect, 3C based binning is complementary with other binning methods, and does not intend to supersede them. In order to get a comprehensive picture of a whole metagenome, several independent approaches are presumably necessary and 3C acts as an additional tool in the available range of options.

On the other hand, phage-host relationship predictions with 3C contacts are relatively unique. While many other methods exists, our approach lets us identify new relationships without any prior bias about either the phage or the bacterial host. On the other hand, successfully identifying these relies on relatively high coverage and successful scaffolding of all genomes involved, a result that can be difficult to achieve in practice for all species of interest. Likewise, a combination of existing methods are necessary to fully understand the dynamics between phages and bacteria in complex communities.

5.2 Future perspectives

Assembly projects are thriving in the community. As the low-hanging fruit gets solved, the complexity of genomes being tackled in the coming years is expected to increase:

- They will be larger, requiring more and more efficient methods and implementations to process the relevant data. A recent landmark was achieved with the chromosome-level assembly of the 32 Gb axolotl genome [355], and we can expect future assembly projects to reach comparable sizes.
- The amount of repeated sequences and other such problematic regions will increase; as we have mentioned, Hi-C based methods tend to struggle when not coupled with other data such as long reads or linked reads. We should expect such data to be more and more prominent in future Hi-C based assembly projects.
- Issues of ploidy will arise: we have seen that so-called *homology patterns* are easily discernible in Hi-C contact maps, but the problem will be compounded in the case of polyploid species. These are very common among plants, including staple crops; the recent chromosome-level characterization of the wheat genome [202] was a crucial landmark in that respect.

Alongside complexities, ambitions will grow as well:

- The advent of single-cell technologies will facilitate the study of many cell lines in a single species, or even a single individual.

- Large-scale projects such as the Vertebrate Genomes Project or the i10k Genomes Project will involve chromosome-level assemblies from dozens of species, and multiple lineages from each species. Our joint reassembly studies could be expected to generalize to many such lineages to be scaffolded *de novo* and investigated for rearrangements.
- Expectations for quality will further increase. With so many data sources to draw from, and long reads becoming cheaper, many projects will focus on the telomere-to-telomere reassembly of every single chromosome in a genome, thus giving access to unparalleled resolution for the purpose of chromosome dynamics.
- More refined structural rearrangements could be detected. The implications are crucial for comparative studies involving multiple cell lines, as such structural variants could be cancer-inducing and their study could unveil the potential mechanisms underlying cancer formation.

In summary, genome assembly and chromosome dynamics are expected to grow more and more complex with ever more ambitious scopes. While our framework is fit for tackling today's problems, more sophisticated methods should be necessary to keep up with the rising expectations and technological progress as more discoveries further expand the field.

Bibliography

[1] J. D. Watson and F. H. Crick. "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". In: *Nature* 171.4356 (1953), pp. 737–738.

[2] C. A. Hutchison. "DNA sequencing: bench to bedside and beyond". In: *Nucleic Acids Res.* 35.18 (2007), pp. 6227–6237.

[3] Robert W. Holley et al. "Nucleotide and oligonucleotide composition of the alanine-, valine-, and tyrosine-acceptor "soluble" ribonucleic acids of yeast". In: *Journal of the American Chemical Society* 83.23 (1961), pp. 4861–4862. DOI: 10.1021/ja01484a040. URL: https://doi.org/10.1021/ja01484a040.

[4] Robert W Holley, James T Madison, and Ada Zamir. "A new method for sequence determination of large oligonucleotides". In: *Biochemical and Biophysical Research Communications* 17.4 (1964), pp. 389–394.

[5] R. W. Holley et al. "Structure of ribonucleic acid". In: *Science* 147.3664 (1965), pp. 1462–1465.

[6] W. Min Jou et al. "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein". In: *Nature* 237.5350 (1972), pp. 82–88.

[7] W. Fiers et al. "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene". In: *Nature* 260.5551 (1976), pp. 500–507.

[8] R. Wu and A. D. Kaiser. "Structure and base sequence in the cohesive ends of bacteriophage lambda DNA". In: *J. Mol. Biol.* 35.3 (1968), pp. 523–537.

[9] F. Sanger et al. "Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA". In: *Proc. Natl. Acad. Sci. U.S.A.* 70.4 (1973), pp. 1209–1213.

[10] R. Wu. "Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage lambda and 186 DNA". In: *J. Mol. Biol.* 51.3 (1970), pp. 501–521.

[11] R. Padmanabhan and R. Wu. "Nucleotide sequence analysis of DNA. IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis". In: *Biochem. Biophys. Res. Commun.* 48.5 (1972), pp. 1295–1302.

[12] R. Padmanabhan, E. Jay, and R. Wu. "Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage T4". In: *Proc. Natl. Acad. Sci. U.S.A.* 71.6 (1974), pp. 2510–2514.

[13] James M. Heather and Benjamin Chain. "The sequence of sequencers: The history of sequencing DNA". In: *Genomics* 107.1 (Jan. 2016), pp. 1–8. DOI: 10.1016/j.ygeno.2015.11.003. URL: https://doi.org/10.1016%2Fj.ygeno.2015.11.003.

[14] F. Sanger and A. R. Coulson. "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *J. Mol. Biol.* 94.3 (1975), pp. 441–448.

[15] Frederick Sanger. "The Croonian Lecture, 1975 Nucleotide sequences in DNA". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 191.1104 (1975), pp. 317–333.

[16] F. Sanger et al. "Nucleotide sequence of bacteriophage phi X174 DNA". In: *Nature* 265.5596 (1977), pp. 687–695.

[17] A. M. Maxam and W. Gilbert. "A new method for sequencing DNA". In: *Proc. Natl. Acad. Sci. U.S.A.* 74.2 (1977), pp. 560–564.

[18] C. Saccone and G. Pesole. *Handbook of Comparative Genomics: Principles and Methodology*. Wiley, 2005. ISBN: 9780471326410. URL: https://books.google.fr/books?id=dXk0JvN2Y-IC.

[19] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proc. Natl. Acad. Sci. U.S.A.* 74.12 (1977), pp. 5463–5467.

[20] Z. G. Chidgeavadze et al. "2',3'-Dideoxy-3' aminonucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases". In: *Nucleic Acids Res.* 12.3 (1984), pp. 1671–1686.

[21] Wonyong Kim. "Application of Metagenomic Techniques: Understanding the Unrevealed Human Microbiota and Explaining the in Clinical Infectious Diseases". In: *Journal of Bacteriology and Virology* 42 (Jan. 2012), p. 263. DOI: 10.4167/jbv.2012.42.4.263.

[22] F. Sanger et al. "Nucleotide sequence of bacteriophage lambda DNA". In: *J. Mol. Biol.* 162.4 (1982), pp. 729–773.

[23] R. Baer et al. "DNA sequence and expression of the B95-8 Epstein-Barr virus genome". In: *Nature* 310.5974 (1984), pp. 207–211.

[24] S. Beck and F. M. Pohl. "DNA sequencing with direct blotting electrophoresis". In: *EMBO J.* 3.12 (1984), pp. 2905–2909.

[25] L. M. Smith et al. "The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis". In: *Nucleic Acids Res.* 13.7 (1985), pp. 2399–2412.

[26] J. M. Prober et al. "A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides". In: *Science* 238.4825 (1987), pp. 336–341.

[27] W. Ansorge et al. "Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis". In: *Nucleic Acids Res.* 15.11 (1987), pp. 4593–4602.

[28] W. Ansorge et al. “A non-radioactive automated method for DNA sequence determination”. In: *J. Biochem. Biophys. Methods* 13.6 (1986), pp. 315–323.

[29] L. M. Smith et al. “Fluorescence detection in automated DNA sequence analysis”. In: *Nature* 321.6071 (1986), pp. 674–679.

[30] H. Swerdlow and R. Gesteland. “Capillary gel electrophoresis for rapid, high resolution DNA sequencing”. In: *Nucleic Acids Res.* 18.6 (1990), pp. 1415–1419.

[31] J. A. Luckey et al. “High speed DNA sequencing by capillary electrophoresis”. In: *Nucleic Acids Res.* 18.15 (1990), pp. 4417–4421.

[32] C. Y. Chen. “DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present”. In: *Front Microbiol* 5 (2014), p. 305.

[33] T. Hunkapiller et al. “Large-scale and automated DNA sequence determination”. In: *Science* 254.5028 (1991), pp. 59–67.

[34] J. Gocayne et al. “Primary structure of rat cardiac beta-adrenergic and muscarinic cholinergic receptors obtained by automated DNA sequence analysis: further evidence for a multigene family”. In: *Proc. Natl. Acad. Sci. U.S.A.* 84.23 (1987), pp. 8296–8300.

[35] M. D. Adams et al. “A model for high-throughput automated DNA sequencing and analysis core facilities”. In: *Nature* 368.6470 (1994), pp. 474–475.

[36] C. Alexander Valencia et al. “Sanger Sequencing Principles, History, and Landmarks”. In: *Next Generation Sequencing Technologies in Medical Genetics*. New York, NY: Springer New York, 2013, pp. 3–11. ISBN: 978-1-4614-9032-6. DOI: 10.1007/978-1-4614-9032-6_1. URL: https://doi.org/10.1007/978-1-4614-9032-6_1.

[37] R. Staden. “A strategy of DNA sequencing employing computer programs”. In: *Nucleic Acids Res.* 6.7 (1979), pp. 2601–2610.

[38] S. Anderson. “Shotgun DNA sequencing using cloned DNase I-generated fragments”. In: *Nucleic Acids Res.* 9.13 (1981), pp. 3015–3027.

[39] R. K. Saiki et al. “Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia”. In: *Science* 230.4732 (1985), pp. 1350–1354.

[40] R. K. Saiki et al. “Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase”. In: *Science* 239.4839 (1988), pp. 487–491.

[41] D. A. Jackson, R. H. Symons, and P. Berg. “Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli”. In: *Proc. Natl. Acad. Sci. U.S.A.* 69.10 (1972), pp. 2904–2909.

[42] S. N. Cohen et al. “Construction of biologically functional bacterial plasmids in vitro”. In: *Proc. Natl. Acad. Sci. U.S.A.* 70.11 (1973), pp. 3240–3244.

[43] S. G. Oliver et al. "The complete DNA sequence of yeast chromosome III". eng. In: *Nature* 357.6373 (May 1992), pp. 38–46. ISSN: 0028-0836. DOI: 10.1038/357038a0.

[44] A. Goffeau et al. "Life with 6000 genes". In: *Science* 274.5287 (1996), pp. 563–567.

[45] R. D. Fleischmann et al. "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". In: *Science* 269.5223 (1995), pp. 496–512.

[46] F. Kunst et al. "The complete genome sequence of the gram-positive bacterium Bacillus subtilis". In: *Nature* 390.6657 (1997), pp. 249–256.

[47] N. T. Perna et al. "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7". In: *Nature* 409.6819 (2001), pp. 529–533.

[48] C. elegans Sequencing Consortium. "Genome sequence of the nematode C. elegans: a platform for investigating biology". In: *Science* 282.5396 (1998), pp. 2012–2018.

[49] Mark D Adams et al. "The genome sequence of Drosophila melanogaster". In: *Science* 287.5461 (2000), pp. 2185–2195.

[50] F. Collins and D. Galas. "A new five-year plan for the U.S. Human Genome Project". In: *Science* 262.5130 (1993), pp. 43–46.

[51] R. H. Waterston, E. S. Lander, and J. E. Sulston. "On the sequencing of the human genome". In: *Proc. Natl. Acad. Sci. U.S.A.* 99.6 (2002), pp. 3712–3716.

[52] W. J. Ansorge. "Next-generation DNA sequencing techniques". In: *N Biotechnol* 25.4 (2009), pp. 195–203.

[53] J. C. Venter et al. "The sequence of the human genome". In: *Science* 291.5507 (2001), pp. 1304–1351.

[54] E. S. Lander et al. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (2001), pp. 860–921.

[55] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (2004), pp. 931–945.

[56] E. Dolgin. "Human genomics: The genome finishers". In: *Nature* 462.7275 (2009), pp. 843–845.

[57] E. D. Hyman. "A new method of sequencing DNA". In: *Anal. Biochem.* 174.2 (1988), pp. 423–436.

[58] P. Nyren and A. Lundin. "Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis". In: *Anal. Biochem.* 151.2 (1985), pp. 504–509.

[59] M. Ronaghi et al. "Real-time DNA sequencing using detection of pyrophosphate release". In: *Anal. Biochem.* 242.1 (1996), pp. 84–89.

[60] P. Nyren. "Enzymatic method for continuous monitoring of DNA polymerase activity". In: *Anal. Biochem.* 167.2 (1987), pp. 235–238.

[61] M. Ronaghi, M. Uhlen, and P. Nyren. "A sequencing method based on real-time pyrophosphate". In: *Science* 281.5375 (1998), pp. 363, 365.

[62] Jan Berka et al. "Bead emulsion nucleic acid amplification". en. Pat. US8748102B2. June 2014. URL: `https://patents.google.com/patent/US8748102/en` (visited on 04/01/2019).

[63] M. Margulies et al. "Genome sequencing in microfabricated high-density picolitre reactors". In: *Nature* 437.7057 (2005), pp. 376–380.

[64] D. S. Tawfik and A. D. Griffiths. "Man-made cell-like compartments for molecular evolution". In: *Nat. Biotechnol.* 16.7 (1998), pp. 652–656.

[65] J. Shendure and H. Ji. "Next-generation DNA sequencing". In: *Nat. Biotechnol.* 26.10 (2008), pp. 1135–1145.

[66] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi. "Next-generation sequencing: from basic research to diagnostics". In: *Clin. Chem.* 55.4 (2009), pp. 641–658.

[67] M. Fedurco et al. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies". In: *Nucleic Acids Res.* 34.3 (2006), e22.

[68] D. R. Bentley et al. "Accurate whole human genome sequencing using reversible terminator chemistry". In: *Nature* 456.7218 (2008), pp. 53–59.

[69] G. Turcatti et al. "A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis". In: *Nucleic Acids Res.* 36.4 (2008), e25.

[70] M. A. Quail et al. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers". In: *BMC Genomics* 13 (2012), p. 341.

[71] J. M. Rothberg et al. "An integrated semiconductor device enabling non-optical genome sequencing". In: *Nature* 475.7356 (2011), pp. 348–352.

[72] K. J. McKernan et al. "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding". In: *Genome Res.* 19.9 (2009), pp. 1527–1541.

[73] T. C. Glenn. "Field guide to next-generation DNA sequencers". In: *Mol Ecol Resour* 11.5 (2011), pp. 759–769.

[74] R. Drmanac et al. "Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays". In: *Science* 327.5961 (2010), pp. 78–81.

[75] Christoph Lippert and David Heckerman. "Computational and statistical issues in personalized medicine". In: *XRDS: Crossroads, The ACM Magazine for Students* 21 (July 2015), pp. 24–27. DOI: `10.1145/2788502`.

[76] W. J. Greenleaf and A. Sidow. "The future of sequencing: convergence of intelligent design and market Darwinism". In: *Genome Biol.* 15.3 (2014), p. 303.

[77] J. Handelsman. "Metagenomics: application of genomics to uncultured microorganisms". In: *Microbiol. Mol. Biol. Rev.* 68.4 (2004), pp. 669–685.

[78] Phillip E. C. Compeau, Pavel A. Pevzner, and Glenn Tesler. "Why are de Bruijn graphs useful for genome assembly?" In: *Nature biotechnology* 29.11 (Nov. 2011), pp. 987–991. ISSN: 1087-0156. DOI: 10.1038/nbt.2023. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5531759/ (visited on 04/02/2019).

[79] E. E. Schadt, S. Turner, and A. Kasarskis. "A window into third-generation sequencing". In: *Hum. Mol. Genet.* 19.R2 (2010), R227–240.

[80] T. P. Niedringhaus et al. "Landscape of next-generation sequencing technologies". In: *Anal. Chem.* 83.12 (2011), pp. 4327–4341.

[81] C. S. Pareek, R. Smoczynski, and A. Tretyn. "Sequencing technologies and genome sequencing". In: *J. Appl. Genet.* 52.4 (2011), pp. 413–435.

[82] I. Braslavsky et al. "Sequence information can be obtained from single DNA molecules". In: *Proc. Natl. Acad. Sci. U.S.A.* 100.7 (2003), pp. 3960–3964.

[83] A. Rhoads and K. F. Au. "PacBio Sequencing and Its Applications". In: *Genomics Proteomics Bioinformatics* 13.5 (2015), pp. 278–289.

[84] I. G. Gut. "New sequencing technologies". In: *Clin Transl Oncol* 15.11 (2013), pp. 879–881.

[85] M. J. Levene et al. "Zero-mode waveguides for single-molecule analysis at high concentrations". In: *Science* 299.5607 (2003), pp. 682–686.

[86] B. A. Flusberg et al. "Direct detection of DNA methylation during single-molecule, real-time sequencing". In: *Nat. Methods* 7.6 (2010), pp. 461–465.

[87] M. Wanunu. "Nanopores: A journey towards DNA sequencing". In: *Phys Life Rev* 9.2 (2012), pp. 125–158.

[88] Y. Wang, Q. Yang, and Z. Wang. "The evolution of nanopore sequencing". In: *Front Genet* 5 (2014), p. 449.

[89] Rongqin Ke et al. "Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences". In: *Human Mutation* 37.12 (Aug. 2016), pp. 1363–1367. DOI: 10.1002/humu.23051. URL: https://doi.org/10.1002%2Fhumu.23051.

[90] J. C. Roach et al. "Pairwise end sequencing: a unified approach to genomic mapping and sequencing". In: *Genomics* 26.2 (1995), pp. 345–353.

[91] A. Edwards et al. "Automated DNA sequencing of the human HPRT locus". In: *Genomics* 6.4 (1990), pp. 593–608.

[92] Zelin Chen et al. *De Novo assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole genome duplication.* July 2018. bioRxiv: 373431. URL: https://doi.org/10.1101%2F373431.

[93] Rebecca E O'Connor et al. "Chromosome-level assembly reveals extensive rearrangement in saker falcon and budgerigar, but not ostrich, genomes". In: *Genome Biology* 19.1 (Oct. 2018). DOI: 10.1186/s13059-018-1550-x. URL: https://doi.org/10.1186%2Fs13059-018-1550-x.

[94] Marten Boetzer et al. "Scaffolding pre-assembled contigs using SSPACE". en. In: *Bioinformatics* 27.4 (Feb. 2011), pp. 578–579. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq683. URL: https://academic.oup.com/bioinformatics/article/27/4/578/197626 (visited on 04/03/2019).

[95] David Sherman et al. "Génolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts". In: *Nucleic Acids Research* 32.Database issue (Jan. 2004), pp. D315–D318. ISSN: 0305-1048. DOI: 10.1093/nar/gkh091. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308825/ (visited on 05/07/2019).

[96] David Sherman et al. "Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts". eng. In: *Nucleic Acids Research* 34.Database issue (Jan. 2006), pp. D432–435. ISSN: 1362-4962. DOI: 10.1093/nar/gkj160.

[97] David J. Sherman et al. "Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes". eng. In: *Nucleic Acids Research* 37.Database issue (Jan. 2009), pp. D550–554. ISSN: 1362-4962. DOI: 10.1093/nar/gkn859.

[98] Tiphaine Martin, David J. Sherman, and Pascal Durrens. "The Génolevures database". eng. In: *Comptes Rendus Biologies* 334.8-9 (Sept. 2011), pp. 585–589. ISSN: 1768-3238. DOI: 10.1016/j.crvi.2011.05.004.

[99] S. S. Vembar et al. "Complete telomere-to-telomere de novo assembly of the Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing". In: *DNA Res.* 23.4 (2016), pp. 339–351.

[100] Gregory Dick. "Metagenomic Binning". en. In: *Genomic Approaches in Earth and Environmental Sciences*. John Wiley & Sons, Ltd, 2018, pp. 89–99. ISBN: 978-1-118-70823-1. DOI: 10.1002/9781118708231.ch7. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118708231.ch7 (visited on 04/03/2019).

[101] Sergey Koren, Todd J. Treangen, and Mihai Pop. "Bambus 2: scaffolding metagenomes". In: *Bioinformatics* 27.21 (Nov. 2011), pp. 2964–2971. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr520. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198580/ (visited on 04/03/2019).

[102] Toshiaki Namiki et al. "MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads". eng. In: *Nucleic Acids Research* 40.20 (Nov. 2012), e155. ISSN: 1362-4962. DOI: 10.1093/nar/gks678.

[103] Yu Peng et al. "Meta-IDBA: a de Novo assembler for metagenomic data". In: *Bioinformatics* 27.13 (July 2011), pp. i94–i101. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr216. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117360/ (visited on 04/03/2019).

[104] Benjamin J. Tully, Elaina D. Graham, and John F. Heidelberg. "The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans". en. In: *Scientific Data* 5 (Jan. 2018), p. 170203. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.203. URL: https://www.nature.com/articles/sdata2017203 (visited on 04/03/2019).

[105] Donovan H. Parks et al. "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life". En. In: *Nature Microbiology* 2.11 (Nov. 2017), p. 1533. ISSN: 2058-5276. DOI: 10.1038/s41564-017-0012-7. URL: https://www.nature.com/articles/s41564-017-0012-7 (visited on 04/03/2019).

[106] Kari-Jouko Räihä and Esko Ukkonen. "The shortest common supersequence problem over binary alphabet is NP-complete". In: *Theoretical Computer Science* 16.2 (1981), pp. 187 –198. ISSN: 0304-3975. DOI: https://doi.org/10.1016/0304-3975(81)90075-X. URL: http://www.sciencedirect.com/science/article/pii/030439758190075X.

[107] Jason R. Miller, Sergey Koren, and Granger Sutton. "Assembly Algorithms for Next-Generation Sequencing Data". In: *Genomics* 95.6 (June 2010), pp. 315–327. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2010.03.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2874646/ (visited on 04/02/2019).

[108] Abdul Rafay Khan et al. "A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective". In: *Evolutionary Bioinformatics Online* 14 (Feb. 2018). ISSN: 1176-9343. DOI: 10.1177/1176934318758650. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5826002/ (visited on 04/02/2019).

[109] E. S. Lander and M. S. Waterman. "Genomic mapping by fingerprinting random clones: a mathematical analysis". In: *Genomics* 2.3 (1988), pp. 231–239.

[110] Z. Li et al. "Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph". In: *Briefings in Functional Genomics* 11.1 (Dec. 2011), pp. 25–37. DOI: 10.1093/bfgp/elr035. URL: https://doi.org/10.1093%2Fbfgp%2Felr035.

[111] René L. Warren et al. "Assembling millions of short DNA sequences using SSAKE". eng. In: *Bioinformatics (Oxford, England)* 23.4 (Feb. 2007), pp. 500–501. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btl629.

[112] William R. Jeck et al. "Extending assembly of short DNA sequences to handle error". eng. In: *Bioinformatics (Oxford, England)* 23.21 (Nov. 2007), pp. 2942–2944. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm451.

[113] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. en. Google-Books-ID: fjxGAQAAIAAJ. W. H. Freeman, 1979, pp. 199–200. ISBN: 978-0-7167-1044-8.

[114] Sara El-Metwally et al. "Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges". In: *PLoS Computational Biology* 9.12 (Dec. 2013). ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1003345. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3861042/ (visited on 04/02/2019).

[115] Melissa de la Bastide and W. Richard McCombie. "Assembling genomic DNA sequences with PHRAP". eng. In: *Current Protocols in Bioinformatics* Chapter 11 (Mar. 2007), Unit11.4. ISSN: 1934-340X. DOI: 10.1002/0471250953.bi1104s17.

[116] Granger G. Sutton et al. "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects". In: *Genome Science and Technology* 1.1 (1995), pp. 9–19. DOI: 10.1089/gst.1995.1.9. eprint: https://doi.org/10.1089/gst.1995.1.9. URL: https://doi.org/10.1089/gst.1995.1.9.

[117] Serafim Batzoglou et al. "ARACHNE: A Whole-Genome Shotgun Assembler". In: *Genome Research* 12.1 (Jan. 2002), pp. 177–189. ISSN: 1088-9051. DOI: 10.1101/gr.208902. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC155255/ (visited on 04/03/2019).

[118] P. A. Pevzner. "DNA physical mapping and alternating Eulerian cycles in colored graphs". In: *Algorithmica* 13.1 (1995), pp. 77–105. ISSN: 1432-0541. DOI: 10.1007/BF01188582. URL: https://doi.org/10.1007/BF01188582.

[119] J. T. Simpson and M. Pop. "The Theory and Practice of Genome Sequence Assembly". In: *Annu Rev Genomics Hum Genet* 16 (2015), pp. 153–172.

[120] *How To Pronounce "De Bruijn"?* URL: https://www.biostars.org/p/7186/.

[121] Wei-Chun Kao, Andrew H. Chan, and Yun S. Song. "ECHO: A reference-free short-read error correction algorithm". In: *Genome Research* 21.7 (July 2011), pp. 1181–1192. ISSN: 1088-9051. DOI: 10.1101/gr.111351.110. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129260/ (visited on 04/03/2019).

[122] Sergey I. Nikolenko, Anton I. Korobeynikov, and Max A. Alekseyev. "BayesHammer: Bayesian clustering for error correction in single-cell sequencing". In: *BMC Genomics* 14.1 (Jan. 2013), S7. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-S1-S7. URL: https://doi.org/10.1186/1471-2164-14-S1-S7 (visited on 04/03/2019).

[123] Jan Schröder et al. "SHREC: a short-read error correction method". eng. In: *Bioinformatics (Oxford, England)* 25.17 (Sept. 2009), pp. 2157–2163. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp379.

[124] Anton Bankevich et al. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing". In: *Journal of Computational Biology* 19.5 (May 2012), pp. 455–477. DOI: 10.1089/cmb.2012.0021. URL: https://doi.org/10.1089%2Fcmb.2012.0021.

[125] Mark J. Chaisson, Dumitru Brinza, and Pavel A. Pevzner. "De novo fragment assembly with short mate-paired reads: Does the read length matter?" In: *Genome Research* 19.2 (Feb. 2009), pp. 336–346. ISSN: 1088-9051. DOI: 10.1101/gr.079053.108. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652199/ (visited on 04/04/2019).

[126] Antoine Limasset. "Nouvelles approches pour l'exploitation des données de séquences génomique haut débit". thesis. Rennes 1, July 2017. URL: http://www.theses.fr/2017REN1S049 (visited on 04/03/2019).

[127] Daniel R. Zerbino and Ewan Birney. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". eng. In: *Genome Research* 18.5 (May 2008), pp. 821–829. ISSN: 1088-9051. DOI: 10.1101/gr.074492.107.

[128] Ruiqiang Li et al. "De novo assembly of human genomes with massively parallel short read sequencing". eng. In: *Genome Research* 20.2 (Feb. 2010), pp. 265–272. ISSN: 1549-5469. DOI: 10.1101/gr.097261.109.

[129] "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler". In: 1 (Dec. 2012), p. 18. ISSN: 2047-217X. DOI: 10.1186/2047-217X-1-18. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3626529/.

[130] Jared T. Simpson et al. "ABySS: A parallel assembler for short read sequence data". In: *Genome Research* 19.6 (June 2009), pp. 1117–1123. ISSN: 1088-9051. DOI: 10.1101/gr.089532.108. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694472/ (visited on 04/04/2019).

[131] Marten Boetzer and Walter Pirovano. "Toward almost closed genomes with GapFiller". In: *Genome Biology* 13.6 (2012), R56. ISSN: 1465-6906. DOI: 10.1186/gb-2012-13-6-r56. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3446322/ (visited on 04/03/2019).

[132] Huilong Du and Chengzhi Liang. *Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads*. June 2018. bioRxiv: 345983. URL: https://doi.org/10.1101%2F345983.

[133] Haowen Zhang, Chirag Jain, and Srinivas Aluru. "A comprehensive evaluation of long read error correction methods". en. In: *bioRxiv* (Jan. 2019), p. 519330. DOI: 10.1101/519330. URL: https://www.biorxiv.org/content/10.1101/519330v1 (visited on 04/03/2019).

[134] Jeremy R. Wang et al. "FMLRC: Hybrid long read error correction using an FM-index". In: *BMC Bioinformatics* 19 (Feb. 2018). ISSN: 1471-2105. DOI: 10.1186/s12859-018-2051-3. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5807796/ (visited on 04/03/2019).

[135] Mari Miyamoto et al. "Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes". eng. In: *BMC genomics* 15 (Aug. 2014), p. 699. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-699.

[136] Sergey Koren et al. "Hybrid error correction and de novo assembly of single-molecule sequencing reads". eng. In: *Nature Biotechnology* 30.7 (July 2012), pp. 693–700. ISSN: 1546-1696. DOI: 10.1038/nbt.2280.

[137] Leena Salmela and Eric Rivals. "LoRDEC: accurate and efficient long read error correction". eng. In: *Bioinformatics (Oxford, England)* 30.24 (Dec. 2014), pp. 3506–3514. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu538.

[138] Leena Salmela et al. "Accurate self-correction of errors in long reads using de Bruijn graphs". eng. In: *Bioinformatics (Oxford, England)* 33.6 (2017), pp. 799–806. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw321.

[139] Chen-Shan Chin et al. "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data". eng. In: *Nature Methods* 10.6 (June 2013), pp. 563–569. ISSN: 1548-7105. DOI: 10.1038/nmeth.2474.

[140] Sergey Koren et al. "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". eng. In: *Genome Research* 27.5 (2017), pp. 722–736. ISSN: 1549-5469. DOI: 10.1101/gr.215087.116.

[141] Dmitry Antipov et al. "hybridSPAdes: an algorithm for hybrid assembly of short and long reads". eng. In: *Bioinformatics (Oxford, England)* 32.7 (2016), pp. 1009–1015. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv688.

[142] Aleksey V. Zimin et al. "The MaSuRCA genome assembler". In: *Bioinformatics* 29.21 (Nov. 2013), pp. 2669–2677. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt476. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3799473/ (visited on 04/03/2019).

[143] Mun Hua Tan et al. "Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly". en. In: *GigaScience* 7.3 (Mar. 2018). DOI: 10.1093/gigascience/gix137. URL: https://academic.oup.com/gigascience/article/7/3/gix137/4803946 (visited on 04/03/2019).

[144] Dongyan Zhao et al. "A chromosomal-scale genome assembly of Tectona grandis reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways". In: *GigaScience* 8.3 (Jan. 2019). DOI: 10.1093/gigascience/giz005. URL: https://doi.org/10.1093%2Fgigascience%2Fgiz005.

[145] In: ().

[146] Norman R. Pace. "A Molecular View of Microbial Diversity and the Biosphere". en. In: *Science* 276.5313 (May 1997), pp. 734–740. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.276.5313.734. URL: http://science.sciencemag.org/content/276/5313/734 (visited on 04/04/2019).

[147] Felix Sommer and Fredrik Bäckhed. "The gut microbiota–masters of host development and physiology". eng. In: *Nature Reviews. Microbiology* 11.4 (Apr. 2013), pp. 227–238. ISSN: 1740-1534. DOI: 10.1038/nrmicro2974.

[148] Jop de Vrieze. "The littlest farmhands". en. In: *Science* 349.6249 (Aug. 2015), pp. 680–683. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.349.6249.680. URL: http://science.sciencemag.org/content/349/6249/680 (visited on 04/04/2019).

[149] Sean M Gibbons and Jack A Gilbert. "Microbial diversity — exploration of natural ecosystems and microbiomes". In: *Current opinion in genetics & development* 35 (Dec. 2015), pp. 66–72. ISSN: 0959-437X. DOI: 10.1016/j.gde.2015.10.003. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4852739/ (visited on 04/04/2019).

[150] Jungwook Yang, Joseph W. Kloepper, and Choong-Min Ryu. "Rhizosphere bacteria help plants tolerate abiotic stress". eng. In: *Trends in Plant Science* 14.1 (Jan. 2009), pp. 1–4. ISSN: 1360-1385. DOI: 10.1016/j.tplants.2008.10.004.

[151] Laurent Philippot et al. "Going back to the roots: the microbial ecology of the rhizosphere". eng. In: *Nature Reviews. Microbiology* 11.11 (Nov. 2013), pp. 789–799. ISSN: 1740-1534. DOI: 10.1038/nrmicro3109.

[152] Andrew L. Kau et al. "Human nutrition, the gut microbiome and the immune system". eng. In: *Nature* 474.7351 (June 2011), pp. 327–336. ISSN: 1476-4687. DOI: 10.1038/nature10213.

[153] John F. Cryan and Timothy G. Dinan. "Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour". eng. In: *Nature Reviews. Neuroscience* 13.10 (Oct. 2012), pp. 701–712. ISSN: 1471-0048. DOI: 10.1038/nrn3346.

[154] Kim Lewis. "Persisters, Biofilms, and the Problem of Cultivability". In: Mar. 2009, pp. 181–194. DOI: 10.1007/7171_2008_7.

[155] Sergey Nurk et al. "metaSPAdes: a new versatile metagenomic assembler". In: *Genome Research* 27.5 (Mar. 2017), pp. 824–834. DOI: 10.1101/gr.213959.116. URL: https://doi.org/10.1101%2Fgr.213959.116.

[156] Dinghua Li et al. "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph". eng. In: *Bioinformatics (Oxford, England)* 31.10 (May 2015), pp. 1674–1676. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv033.

[157] Dinghua Li et al. "MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices". eng. In: *Methods (San Diego, Calif.)* 102 (2016), pp. 3–11. ISSN: 1095-9130. DOI: 10.1016/j.ymeth.2016.02.020.

[158] Yu Peng et al. "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth". eng. In: *Bioinformatics (Oxford, England)* 28.11 (June 2012), pp. 1420–1428. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts174.

[159] Sharmila S. Mande, Monzoorul Haque Mohammed, and Tarini Shankar Ghosh. "Classification of metagenomic sequences: methods and challenges". en. In: *Briefings in Bioinformatics* 13.6 (Nov. 2012), pp. 669–681. ISSN: 1467-5463. DOI: 10.1093/bib/bbs054. URL: https://academic.oup.com/bib/article/13/6/669/193900 (visited on 04/03/2019).

[160] Mads Albertsen et al. "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes". eng. In: *Nature Biotechnology* 31.6 (June 2013), pp. 533–538. ISSN: 1546-1696. DOI: 10.1038/nbt.2579.

[161] Henrik J. Nielsen et al. "Dynamics of Escherichia coli Chromosome Segregation during Multifork Replication". en. In: *Journal of Bacteriology* 189.23 (Dec. 2007), pp. 8660–8666. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.01212-07. URL: https://jb.asm.org/content/189/23/8660 (visited on 04/04/2019).

[162] Michael Imelfort et al. "GroopM: an automated tool for the recovery of population genomes from related metagenomes". In: *PeerJ* 2 (Sept. 2014), e603. DOI: 10.7717/peerj.603. URL: https://doi.org/10.7717%2Fpeerj.603.

[163] Elaina D. Graham, John F. Heidelberg, and Benjamin J. Tully. "BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation". In: *PeerJ* 5 (Mar. 2017), e3035. DOI: 10.7717/peerj.3035. URL: https://doi.org/10.7717%2Fpeerj.3035.

[164] Cedric C. Laczny et al. "BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation". In: *Nucleic Acids Research* 45.W1 (May 2017), W171–W179. DOI: 10.1093/nar/gkx348. URL: https://doi.org/10.1093%2Fnar%2Fgkx348.

[165] Dongwan D. Kang et al. "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities". In: *PeerJ* 3 (Aug. 2015), e1165. DOI: 10.7717/peerj.1165. URL: https://doi.org/10.7717%2Fpeerj.1165.

[166] Johannes Alneberg et al. "Binning metagenomic contigs by coverage and composition". In: *Nature Methods* 11.11 (Sept. 2014), pp. 1144–1146. DOI: 10.1038/nmeth.3103. URL: https://doi.org/10.1038%2Fnmeth.3103.

[167] Yu-Wei Wu et al. "MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm". In: *Microbiome* 2.1 (Aug. 2014). DOI: 10.1186/2049-2618-2-26. URL: https://doi.org/10.1186%2F2049-2618-2-26.

[168] Yu-Wei Wu, Blake A. Simmons, and Steven W. Singer. "MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets". In: *Bioinformatics* 32.4 (Oct. 2015), pp. 605–607. DOI: 10.1093/bioinformatics/btv638. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtv638.

[169] Damayanthi Herath et al. "CoMet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision". In: *BMC Bioinformatics* 18.Suppl 16 (Dec. 2017). ISSN: 1471-2105. DOI: 10.1186/s12859-017-1967-3. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5751405/ (visited on 04/03/2019).

[170] Yang Young Lu et al. "COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge". In: *Bioinformatics* (June 2016), btw290. DOI: 10.1093/bioinformatics/btw290. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtw290.

[171] Christian M. K. Sieber et al. "Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy". In: *Nature Microbiology* 3.7 (May 2018), pp. 836–843. DOI: 10.1038/s41564-018-0171-1. URL: https://doi.org/10.1038%2Fs41564-018-0171-1.

[172] Hsin-Hung Lin and Yu-Chieh Liao. "Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes". In: *Scientific Reports* 6 (Apr. 2016). ISSN: 2045-2322. DOI: 10.1038/srep24175. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4828714/ (visited on 04/03/2019).

[173] Brian Cleary et al. "Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning". In: *Nature Biotechnology* 33.10 (Sept. 2015), pp. 1053–1060. DOI: 10.1038/nbt.3329. URL: https://doi.org/10.1038%2Fnbt.3329.

[174] Jakob Nybo Nissen et al. "Binning microbial genomes using deep learning". en. In: *bioRxiv* (Dec. 2018), p. 490078. DOI: 10.1101/490078. URL: https://www.biorxiv.org/content/10.1101/490078v2 (visited on 04/03/2019).

[175] S. L. Salzberg et al. "GAGE: A critical evaluation of genome assemblies and assembly algorithms". In: *Genome Research* 22.3 (Jan. 2012), pp. 557–567. DOI: 10.1101/gr.131383.111. URL: https://doi.org/10.1101%2Fgr.131383.111.

[176] "Assemblathon 1: A competitive assessment of de novo short read assembly methods". In: 21 (Dec. 2011), pp. 2224–2241. ISSN: 1088-9051. DOI: 10.1101/gr.126599.111. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227110/.

[177] "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species". In: 2 (July 2013), p. 10. ISSN: 2047-217X. DOI: 10.1186/2047-217X-2-10.

[178] Olivier Devuyst. "The 1000 Genomes Project: Welcome to a New World". In: *Peritoneal Dialysis International : Journal of the International Society for Peritoneal Dialysis* 35.7 (Dec. 2015), pp. 676–677. ISSN: 0896-8608. DOI: 10.3747/pdi.2015.00261. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4690620/ (visited on 04/03/2019).

[179] Laura Clarke et al. "The 1000 Genomes Project: Data Management and Community Access". In: *Nature methods* 9.5 (Apr. 2012), pp. 459–462. ISSN: 1548-7091. DOI: 10.1038/nmeth.1974. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3340611/ (visited on 04/03/2019).

[180] Genome 10K Community of Scientists. "Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species". eng. In: *The Journal of Heredity* 100.6 (Dec. 2009), pp. 659–674. ISSN: 1465-7333. DOI: 10.1093/jhered/esp086.

[181] Elizabeth Pennisi. "DNA sequencing. No genome left behind". eng. In: *Science (New York, N.Y.)* 326.5954 (Nov. 2009), pp. 794–795. ISSN: 1095-9203. DOI: 10.1126/science.326_794.

[182] Erika Check Hayden. "10,000 genomes to come". eng. In: *Nature* 462.7269 (Nov. 2009), p. 21. ISSN: 1476-4687. DOI: 10.1038/462021a.

[183] Klaus-Peter Koepfli, Benedict Paten, and Stephen J. O'Brien. "The Genome 10K Project: A Way Forward". In: *Annual review of animal biosciences* 3 (2015), pp. 57–111. ISSN: 2165-8102. DOI: 10.1146/annurev-animal-090414-014900. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5837290/ (visited on 04/03/2019).

[184] GIGA Community of Scientists et al. "The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes". eng. In: *The Journal of Heredity* 105.1 (Feb. 2014), pp. 1–18. ISSN: 1465-7333. DOI: 10.1093/jhered/est084.

[185] Christian R. Voolstra, Gert Wörheide, and Jose V. Lopez. "Advancing Genomics through the Global Invertebrate Genomics Alliance (GIGA)". In: *Invertebrate systematics* 31.1 (Mar. 2017), pp. 1–7. ISSN: 1445-5226. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5464758/ (visited on 04/04/2019).

[186] Gene E. Robinson et al. "Creating a buzz about insect genomes". eng. In: *Science (New York, N.Y.)* 331.6023 (Mar. 2011), p. 1386. ISSN: 1095-9203. DOI: 10.1126/science.331.6023.1386.

[187] Nicole M. Davis et al. "Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data". In: *Microbiome* 6.1 (Dec. 2018), p. 226. ISSN: 2049-2618. DOI: 10.1186/s40168-018-0605-2. URL: https://doi.org/10.1186/s40168-018-0605-2 (visited on 04/03/2019).

[188] Jennifer Lu and Steven L. Salzberg. "Removing contaminants from databases of draft genomes". en. In: *PLOS Computational Biology* 14.6 (June 2018), e1006277. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006277. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006277 (visited on 04/03/2019).

[189] Ludovic Mallet et al. "PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies". en. In: *Bioinformatics* 33.20 (Oct. 2017), pp. 3283–3285. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx396. URL: https://academic.oup.com/bioinformatics/article/33/20/3283/3868725 (visited on 04/03/2019).

[190] Hao Zhao et al. "CrossMap: a versatile tool for coordinate conversion between genome assemblies". eng. In: *Bioinformatics (Oxford, England)* 30.7 (Apr. 2014), pp. 1006–1007. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt730.

[191] Victoria Dominguez Del Angel et al. "Ten steps to get started in Genome Assembly and Annotation". en. In: *F1000Research* 7 (Feb. 2018), p. 148. ISSN: 2046-1402. DOI: 10.12688/f1000research.13598.1. URL: https://f1000research.com/articles/7-148/v1 (visited on 04/03/2019).

[192] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (May 2018). Ed. by Inanc Birol, pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbty191.

[193] Alexey Gurevich et al. "QUAST: quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8 (Feb. 2013), pp. 1072–1075. DOI: 10.1093/bioinformatics/btt086. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtt086.

[194] Alla Mikheenko et al. "Versatile genome assembly evaluation with QUAST-LG". In: *Bioinformatics* 34.13 (June 2018), pp. i142–i150. DOI: 10.1093/bioinformatics/bty266. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbty266.

[195] Martin Hunt et al. "REAPR: a universal tool for genome assembly evaluation". In: *Genome Biology* 14.5 (2013), R47. ISSN: 1465-6906. DOI: 10.1186/gb-2013-14-5-r47. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3798757/ (visited on 04/03/2019).

[196] Janna L. Fierst. "Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools". In: *Frontiers in Genetics* 6 (June 2015). ISSN: 1664-8021. DOI: 10.3389/fgene.2015.00220. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4473057/ (visited on 04/04/2019).

[197] Reuben J Pengelly and Andrew Collins. "Linkage disequilibrium maps to guide contig ordering for genome assembly". In: *Bioinformatics* 35.4 (Aug. 2018). Ed. by Inanc Birol, pp. 541–545. DOI: 10.1093/bioinformatics/bty687. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbty687.

[198] Haibao Tang et al. "ALLMAPS: robust scaffold ordering based on multiple maps". eng. In: *Genome Biology* 16 (Jan. 2015), p. 3. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0573-1.

[199] Caroline Belser et al. "Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps". In: *Nature Plants* 4.11 (Nov. 2018), pp. 879–887. DOI: 10.1038/s41477-018-0289-4. URL: https://doi.org/10.1038%2Fs41477-018-0289-4.

[200] Changsheng Li et al. "Genome Sequencing and Assembly by Long Reads in Plants". In: *Genes* 9.1 (Dec. 2017). ISSN: 2073-4425. DOI: 10.3390/genes9010006. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5793159/ (visited on 04/03/2019).

[201] Maximilian H.-W. Schmidt et al. "De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing". en. In: *The Plant Cell* 29.10 (Oct. 2017), pp. 2336–2348. ISSN: 1040-4651, 1532-298X. DOI: 10.1105/tpc.17.00521. URL: http://www.plantcell.org/content/29/10/2336 (visited on 04/04/2019).

[202] The International Wheat Genome Sequencing Consortium (IWGSC). "Shifting the limits in wheat research and breeding using a fully annotated reference genome". In: *Science* 361.6403 (Aug. 2018), eaar7191. DOI: 10.1126/science.aar7191. URL: https://doi.org/10.1126%2Fscience.aar7191.

[203] Mahul Chakraborty et al. "Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage". In: *Nucleic Acids Research* 44.19 (Nov. 2016), e147. ISSN: 0305-1048. DOI: 10.1093/nar/gkw654. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5100563/ (visited on 04/03/2019).

[204] Alejandro Hernandez Wences and Michael C. Schatz. "Metassembler: merging and optimizing de novo genome assemblies". In: *Genome Biology* 16.1 (Sept. 2015), p. 207. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0764-4. URL: https://doi.org/10.1186/s13059-015-0764-4 (visited on 04/03/2019).

[205] Sergey S. Aganezov and Max A. Alekseyev. "CAMSA: a tool for comparative analysis and merging of scaffold assemblies". In: *BMC Bioinformatics* 18.Suppl 15 (Dec. 2017). ISSN: 1471-2105. DOI: 10.1186/s12859-017-1919-y. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5731503/ (visited on 04/03/2019).

[206] Hind Alhakami, Hamid Mirebrahim, and Stefano Lonardi. "A comparative evaluation of genome assembly reconciliation tools". In: *Genome Biology* 18.1 (May 2017), p. 93. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1213-3. URL: https://doi.org/10.1186/s13059-017-1213-3 (visited on 04/03/2019).

[207] Adam C. English et al. "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology". In: *PLoS ONE* 7.11 (Nov. 2012). Ed. by Zhanjiang Liu, e47768. DOI: 10.1371/journal.pone.0047768. URL: https://doi.org/10.1371%2Fjournal.pone.0047768.

[208] Andreas Wallberg et al. *A hybrid de novo genome assembly of the honeybee, Apis mellifera , with chromosome-length scaffolds*. July 2018. bioRxiv: 361469. URL: https://doi.org/10.1101%2F361469.

[209] Robert VanBuren et al. "A near complete, chromosome-scale assembly of the black raspberry (Rubus occidentalis) genome". In: *GigaScience* 7.8 (Aug. 2018). DOI: 10.1093/gigascience/giy094. URL: https://doi.org/10.1093%2Fgigascience%2Fgiy094.

[210] Alina Ott et al. "Linked read technology for assembling large complex and polyploid genomes". In: *BMC Genomics* 19 (Sept. 2018). ISSN: 1471-2164. DOI: 10.1186/s12864-018-5040-z. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6122573/ (visited on 04/03/2019).

[211] Guangyi Fan et al. "The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution". en. In: *Molecular Ecology Resources* 0.ja (). ISSN: 1755-0998. DOI: 10.1111/1755-0998.13003. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13003 (visited on 04/04/2019).

[212] Lauren Coombe et al. "Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics' GemCode Sequencing Data". In: *PLoS ONE* 11.9 (Sept. 2016). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0163059. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5025161/ (visited on 04/03/2019).

[213] David C. Danko et al. "Minerva: an alignment- and reference-free approach to deconvolve Linked-Reads for metagenomics". In: *Genome Research* 29.1 (Jan. 2019), pp. 116–124. ISSN: 1088-9051. DOI: 10.1101/gr.235499.118. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6314158/ (visited on 04/03/2019).

[214] Joshua A. Udall and R. Kelly Dawe. "Is It Ordered Correctly? Validating Genome Assemblies by Optical Mapping". In: *The Plant Cell* 30.1 (Dec. 2017), pp. 7–14. DOI: 10.1105/tpc.17.00514. URL: https://doi.org/10.1105%2Ftpc.17.00514.

[215] "A reference standard for genome biology". en. In: *Nature Biotechnology* 36.12 (Dec. 2018), p. 1121. ISSN: 1546-1696. DOI: 10.1038/nbt.4318. URL: https://www.nature.com/articles/nbt.4318 (visited on 04/04/2019).

[216] Doug Hyatt et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11 (Mar. 2010), p. 119. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-119. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2848648/ (visited on 04/04/2019).

[217] Doug Hyatt et al. "Gene and translation initiation site prediction in metagenomic sequences". eng. In: *Bioinformatics (Oxford, England)* 28.17 (Sept. 2012), pp. 2223–2230. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts429.

[218] Mario Stanke et al. "AUGUSTUS: a web server for gene finding in eukaryotes". eng. In: *Nucleic Acids Research* 32.Web Server issue (July 2004), W309–312. ISSN: 1362-4962. DOI: 10.1093/nar/gkh379.

[219] Mario Stanke et al. "AUGUSTUS: ab initio prediction of alternative transcripts". In: *Nucleic Acids Research* 34.Web Server issue (July 2006), W435–W439. ISSN: 0305-1048. DOI: 10.1093/nar/gkl200. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1538822/ (visited on 04/04/2019).

[220] Katharina J. Hoff and Mario Stanke. "Predicting Genes in Single Genomes with AUGUSTUS". eng. In: *Current Protocols in Bioinformatics* 65.1 (2019), e57. ISSN: 1934-340X. DOI: 10.1002/cpbi.57.

[221] Donovan H. Parks et al. "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes". In: *Genome Research* 25.7 (May 2015), pp. 1043–1055. DOI: 10.1101/gr.186072.114. URL: https://doi.org/10.1101%2Fgr.186072.114.

[222] Victor M. Markowitz et al. "IMG 4 version of the integrated microbial genomes comparative analysis system". eng. In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D560–567. ISSN: 1362-4962. DOI: 10.1093/nar/gkt963.

[223] Felipe A. Simão et al. "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". en. In: *Bioinformatics* 31.19 (Oct. 2015), pp. 3210–3212. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv351. URL: https://academic.oup.com/bioinformatics/article/31/19/3210/211866 (visited on 04/04/2019).

[224] Robert M. Waterhouse et al. "OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs". In: *Nucleic Acids Research* 41.Database issue (Jan. 2013), pp. D358–D365. ISSN: 0305-1048. DOI: 10.1093/nar/gks1116. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531149/ (visited on 04/04/2019).

[225] Evgenia V. Kriventseva et al. "OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs". eng. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D807–D811. ISSN: 1362-4962. DOI: 10.1093/nar/gky1053.

[226] Genis Parra, Keith Bradnam, and Ian Korf. "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes". eng. In: *Bioinformatics (Oxford, England)* 23.9 (May 2007), pp. 1061–1067. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm071.

[227] Genis Parra et al. "Assessing the gene space in draft genomes". eng. In: *Nucleic Acids Research* 37.1 (Jan. 2009), pp. 289–297. ISSN: 1362-4962. DOI: 10.1093/nar/gkn916.

[228] Elisabeth Veeckman, Tom Ruttink, and Klaas Vandepoele. "Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences". en. In: *The Plant Cell* 28.8 (Aug. 2016), pp. 1759–1768. ISSN: 1040-4651, 1532-298X. DOI: 10.1105/tpc.16.00349. URL: http://www.plantcell.org/content/28/8/1759 (visited on 04/04/2019).

[229] Michiel Van Bel et al. "Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform". en. In: *Plant Physiology* 158.2 (Feb. 2012), pp. 590–600. ISSN: 0032-0889, 1532-2548. DOI: 10.1104/pp.111.189514. URL: http://www.plantphysiol.org/content/158/2/590 (visited on 04/04/2019).

[230] Ousmane Hamadoun Cisse and Jason Eric Stajich. "FGMP: assessing fungal genome completeness and gene content". en. In: *bioRxiv* (Oct. 2018), p. 049619. DOI: 10.1101/049619. URL: https://www.biorxiv.org/content/10.1101/049619v2 (visited on 04/04/2019).

[231] David Koslicki and Daniel Falush. "MetaPalette: a k-mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation". en. In: *mSystems* 1.3 (June 2016), e00020–16. ISSN: 2379-5077. DOI: 10.1128/mSystems.00020-16. URL: https://msystems.asm.org/content/1/3/e00020-16 (visited on 04/04/2019).

[232] Sebastian Deorowicz et al. "KMC 2: fast and resource-frugal k-mer counting". eng. In: *Bioinformatics (Oxford, England)* 31.10 (May 2015), pp. 1569–1576. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv022.

[233] Marek Kokot, Maciej Dlugosz, and Sebastian Deorowicz. "KMC 3: counting and manipulating k-mer statistics". eng. In: *Bioinformatics (Oxford, England)* 33.17 (Sept. 2017), pp. 2759–2761. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx304.

[234] Daniel Mapleson et al. "KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies". In: *Bioinformatics* 33.4 (Feb. 2017), pp. 574–576. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw663. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408915/ (visited on 04/03/2019).

[235] J. Jurka et al. "Repbase Update, a database of eukaryotic repetitive elements". eng. In: *Cytogenetic and Genome Research* 110.1-4 (2005), pp. 462–467. ISSN: 1424-859X. DOI: 10.1159/000084979.

[236] Weidong Bao, Kenji K. Kojima, and Oleksiy Kohany. "Repbase Update, a database of repetitive elements in eukaryotic genomes". In: *Mobile DNA* 6 (June 2015). ISSN: 1759-8753. DOI: 10.1186/s13100-015-0041-9. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4455052/ (visited on 04/04/2019).

[237] Maja Tarailo-Graovac and Nansheng Chen. "Using RepeatMasker to identify repetitive elements in genomic sequences". eng. In: *Current Protocols in Bioinformatics* Chapter 4 (Mar. 2009), Unit 4.10. ISSN: 1934-340X. DOI: 10.1002/0471250953.bi0410s25.

[238] Hani Z. Girgis. "Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale". In: *BMC Bioinformatics* 16.1 (July 2015), p. 227. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0654-5. URL: https://doi.org/10.1186/s12859-015-0654-5 (visited on 04/04/2019).

[239] Mark Allen Mitchell. "Interhelical DNA-DNA Crosslinking of Bacteriophage Lambda: Bis(monoazidomethidium)octaoxahexacosanediamine and Bis(psoralen)nonaethyleneoxy ether, Probes of Packaged Nucleic Acid". phd. California Institute of Technology, 1983. URL: http://resolver.caltech.edu/CaltechETD:etd-09112006-153134 (visited on 04/02/2019).

[240] K. Rippe. "Making contacts on a nucleic acid polymer". In: *Trends Biochem. Sci.* 26.12 (2001), pp. 733–740.

[241] J. Dekker. "Capturing Chromosome Conformation". In: *Science* 295.5558 (Feb. 2002), pp. 1306–1311. DOI: 10.1126/science.1067799. URL: https://doi.org/10.1126%2Fscience.1067799.

[242] Elzo de Wit and Wouter de Laat. "A decade of 3C technologies: insights into nuclear organization". In: *Genes & Development* 26.1 (Jan. 2012), pp. 11–24. ISSN: 0890-9369. DOI: 10.1101/gad.179804.111. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3258961/ (visited on 05/08/2019).

[243] Na Xu et al. "Evidence that homologous X-chromosome pairing requires transcription and Ctcf protein". eng. In: *Nature Genetics* 39.11 (Nov. 2007), pp. 1390–1396. ISSN: 1546-1718. DOI: 10.1038/ng.2007.5.

[244] Marieke Simonis et al. "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)". eng. In: *Nature Genetics* 38.11 (Nov. 2006), pp. 1348–1354. ISSN: 1061-4036. DOI: 10.1038/ng1896.

[245] Daan Noordermeer and Wouter de Laat. "Joining the loops: -Globin gene regulation". en. In: *IUBMB Life* 60.12 (2008), pp. 824–833. ISSN: 1521-6551. DOI: 10.1002/iub.129. URL: https://iubmb.onlinelibrary.wiley.com/doi/abs/10.1002/iub.129 (visited on 05/08/2019).

[246] Josée Dostie et al. "Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements". en. In: *Genome Research* 16.10 (Oct. 2006), pp. 1299–1309. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.5571506. URL: http://genome.cshlp.org/content/16/10/1299 (visited on 04/04/2019).

[247] Elphège P. Nora et al. "Spatial partitioning of the regulatory landscape of the X-inactivation centre". eng. In: *Nature* 485.7398 (Apr. 2012), pp. 381–385. ISSN: 1476-4687. DOI: 10.1038/nature11049.

[248] Nynke L. van Berkum et al. "Hi-C: A Method to Study the Three-dimensional Architecture of Genomes." In: *Journal of Visualized Experiments : JoVE* 39 (May 2010). ISSN: 1940-087X. DOI: 10.3791/1869. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149993/ (visited on 04/04/2019).

[249] E. Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *Science* 326.5950 (2009), pp. 289–293.

[250] Takashi Nagano et al. "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure". eng. In: *Nature* 502.7469 (Oct. 2013), pp. 59–64. ISSN: 1476-4687. DOI: 10.1038/nature12593.

[251] Tim J. Stevens et al. "3D structure of individual mammalian genomes studied by single cell Hi-C". In: *Nature* 544.7648 (Apr. 2017), pp. 59–64. ISSN: 0028-0836. DOI: 10.1038/nature21429. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5385134/ (visited on 04/04/2019).

[252] Bryan R. Lajoie, Job Dekker, and Noam Kaplan. "The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines". In: *Methods (San Diego, Calif.)* 72 (Jan. 2015), pp. 65–75. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2014.10.031. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4347522/ (visited on 04/07/2019).

[253] Monika Sekelja, Jonas Paulsen, and Philippe Collas. "4D nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation?" In: *Genome Biology* 17 (Apr. 2016). ISSN: 1474-7596. DOI: 10.1186/s13059-016-0923-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4823877/ (visited on 04/04/2019).

[254] Longzhi Tan et al. "3D Genome Structures of Single Diploid Human Cells". In: *Science (New York, N.Y.)* 361.6405 (Aug. 2018), pp. 924–928. ISSN: 0036-8075. DOI: 10.1126/science.aat5641. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6360088/ (visited on 04/04/2019).

[255] Leonid A. Mirny. "The fractal globule as a model of chromatin architecture in the cell". In: *Chromosome Research* 19.1 (2011), pp. 37–51. ISSN: 0967-3849. DOI: 10.1007/s10577-010-9177-0. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3040307/ (visited on 04/04/2019).

[256] H. Muller et al. "Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for Hi-C". In: *Mol. Syst. Biol.* 14.7 (July 2018), e8293.

[257] Vittore F. Scolari et al. "Kinetic Signature of Cooperativity in the Irreversible Collapse of a Polymer". In: *Phys. Rev. Lett.* 121 (5 2018), p. 057801. DOI: 10.1103/PhysRevLett.121.057801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.057801.

[258] G. Tiana and L. Giorgetti. *Modeling the 3D Conformation of Genomes*. Series in Computational Biophysics. CRC Press, 2019. ISBN: 9781351387002. URL: https://books.google.fr/books?id=DzqDDwAAQBAJ.

[259] Luciana Lazar-Stefanita et al. "Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle". In: *The EMBO Journal* 36.18 (2017), pp. 2684–2697. ISSN: 0261-4189. DOI: 10.15252/embj.201797342. eprint: http://emboj.embopress.org/content/36/18/2684.full.pdf. URL: http://emboj.embopress.org/content/36/18/2684.

[260] Luciana Lazar-Stefanita. "Functional reorganization of the yeast genome during the cell cycle". thesis. Paris 6, Sept. 2017. URL: http://www.theses.fr/2017PA066400 (visited on 04/02/2019).

[261] Karen J. Meaburn and Tom Misteli. "Cell biology: chromosome territories". eng. In: *Nature* 445.7126 (Jan. 2007), pp. 379–781. ISSN: 1476-4687. DOI: 10.1038/445379a.

[262] Thomas Cremer and Marion Cremer. "Chromosome territories". eng. In: *Cold Spring Harbor Perspectives in Biology* 2.3 (Mar. 2010), a003889. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a003889.

[263] G. Mercy et al. "3D organization of synthetic and scrambled chromosomes". In: *Science* 355.6329 (Mar. 2017).

[264] Hervé Marie-Nelly et al. "Filling annotation gaps in yeast genomes using genome-wide contact maps". In: *Bioinformatics* 30.15 (Apr. 2014), pp. 2105–2113. DOI: 10.1093/bioinformatics/btu162. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtu162.

[265] Nelle Varoquaux et al. "Accurate identification of centromere locations in yeast genomes using Hi-C". In: *Nucleic Acids Research* 43.11 (June 2015), pp. 5331–5339. ISSN: 0305-1048. DOI: 10.1093/nar/gkv424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4477656/ (visited on 04/07/2019).

[266] Yuval Benjamini and Terence P. Speed. "Summarizing and correcting the GC content bias in high-throughput sequencing". eng. In: *Nucleic Acids Research* 40.10 (May 2012), e72. ISSN: 1362-4962. DOI: 10.1093/nar/gks001.

[267] Zhijun Han and Gang Wei. "Computational tools for Hi-C data analysis". In: *Quantitative Biology* 5.3 (2017), pp. 215–225. ISSN: 2095-4697. DOI: 10.1007/s40484-017-0113-6. URL: https://doi.org/10.1007/s40484-017-0113-6.

[268] F. Ay and W. S. Noble. "Analysis methods for studying the 3D architecture of the genome". In: *Genome Biol.* 16 (2015), p. 183.

[269] Nicolas Servant et al. "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing". eng. In: *Genome Biology* 16 (Dec. 2015), p. 259. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0831-x.

[270] Michael Eg Sauria et al. "HiFive: a tool suite for easy and efficient HiC and 5C data analysis". eng. In: *Genome Biology* 16 (Oct. 2015), p. 237. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0806-y.

[271] Marie-Eve Val et al. "A checkpoint control orchestrates the replication of the two chromosomes of Vibrio cholerae". In: *Science Advances* 2.4 (Apr. 2016), e1501914. DOI: 10.1126/sciadv.1501914. URL: https://doi.org/10.1126%2Fsciadv.1501914.

[272] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4 (Mar. 2012), pp. 357–359. DOI: 10.1038/nmeth.1923. URL: https://doi.org/10.1038%2Fnmeth.1923.

[273] E. Yaffe and A. Tanay. "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture". In: *Nat. Genet.* 43.11 (2011), pp. 1059–1065.

[274] Yannick G. Spill, David Castillo, and Marc A. Marti-Renom. *Binless normalization of Hi-C data provides significant interaction and difference detection independently of resolution*. Nov. 2017. bioRxiv: 214403. URL: https://doi.org/10.1101%2F214403.

[275] F. Ay, T. L. Bailey, and W. S. Noble. "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts". In: *Genome Res.* 24.6 (2014), pp. 999–1011.

[276] Ming Hu et al. "HiCNorm: removing biases in Hi-C data via Poisson regression". en. In: *Bioinformatics* 28.23 (Dec. 2012), pp. 3131–3133. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts570. URL: https://academic.oup.com/bioinformatics/article/28/23/3131/192582 (visited on 04/04/2019).

[277] Netta Mendelson Cohen et al. "SHAMAN: bin-free randomization, normalization and screening of Hi-C matrices". en. In: *bioRxiv* (Sept. 2017), p. 187203. DOI: 10.1101/187203. URL: https://www.biorxiv.org/content/10.1101/187203v1 (visited on 04/05/2019).

[278] M. Imakaev et al. "Iterative correction of Hi-C data reveals hallmarks of chromosome organization". In: *Nat. Methods* 9.10 (2012), pp. 999–1003.

[279] Axel Cournac et al. "Normalization of a chromosomal contact map". In: *BMC Genomics* 13.1 (2012), p. 436. DOI: 10.1186/1471-2164-13-436. URL: https://doi.org/10.1186%2F1471-2164-13-436.

[280] Philip A. Knight and Daniel Ruiz. "A fast algorithm for matrix balancing". en. In: *IMA Journal of Numerical Analysis* 33.3 (July 2013), pp. 1029–1047. ISSN: 0272-4979. DOI: 10.1093/imanum/drs019. URL: https://academic.oup.com/imajna/article/33/3/1029/659457 (visited on 04/02/2019).

[281] N. Servant et al. "Effective normalization for copy number variation in Hi-C data". In: *BMC Bioinformatics* 19.1 (2018), p. 313.

[282] Hua-Jun Wu and Franziska Michor. "A computational strategy to adjust for copy number in tumor Hi-C data". eng. In: *Bioinformatics (Oxford, England)* 32.24 (2016), pp. 3695–3701. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw540.

[283] Yan Zhang et al. "Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus". En. In: *Nature Communications* 9.1 (Feb. 2018), p. 750. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03113-2. URL: https://www.nature.com/articles/s41467-018-03113-2 (visited on 04/04/2019).

[284] L. Carron et al. "Boost-HiC: computational enhancement of long-range contacts in chromosomal contact maps". en. In: *Bioinformatics* (). DOI: 10.1093/bioinformatics/bty1059. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty1059/5273482 (visited on 04/04/2019).

[285] Ye Zheng, Ferhat Ay, and Sunduz Keles. "Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies". In: *eLife* 8 (Jan. 2019). Ed. by Bing Ren, e38070. ISSN: 2050-084X. DOI: 10.7554/eLife.38070. URL: https://doi.org/10.7554/eLife.38070 (visited on 04/04/2019).

[286] Mattia Forcato et al. "Comparison of computational methods for Hi-C data analysis". In: *Nature methods* 14.7 (July 2017), pp. 679–685. ISSN: 1548-7091. DOI: 10.1038/nmeth.4325. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5493985/ (visited on 04/04/2019).

[287] Tao Yang et al. "HiCRep: assessing the reproducibility of Hi-C data using a stratum- adjusted correlation coefficient". en. In: *Genome Research* (Aug. 2017), gr.220640.117. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.220640.117. URL: http://genome.cshlp.org/content/early/2017/08/30/gr.220640.117 (visited on 04/05/2019).

[288] Enrique Vidal et al. "OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes". In: *Nucleic Acids Research* 46.8 (May 2018), e49. ISSN: 0305-1048. DOI: 10.1093/nar/gky064. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5934634/ (visited on 04/05/2019).

[289] Koon-Kiu Yan et al. "HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps". en. In: *Bioinformatics* 33.14 (July 2017), pp. 2199–2201. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx152. URL: https://academic.oup.com/bioinformatics/article/33/14/2199/3078603 (visited on 04/04/2019).

[290] Oana Ursu et al. "GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs". en. In: *Bioinformatics* 34.16 (Aug. 2018), pp. 2701–2707. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty164. URL: https://academic.oup.com/bioinformatics/article/34/16/2701/4938489 (visited on 04/04/2019).

[291] Virginia S. Lioy et al. "Multiscale Structuring of the E. coli Chromosome by Nucleoid-Associated and Condensin Proteins". eng. In: *Cell* 172.4 (2018), 771–783.e18. ISSN: 1097-4172. DOI: 10.1016/j.cell.2017.12.027.

[292] 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. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0683-0. URL: https://doi.org/10.1186/s12859-015-0683-0 (visited on 04/05/2019).

[293] Shubiao Li et al. "Quasi-negative binomial distribution: Properties and applications". In: *Computational Statistics & Data Analysis* 55.7 (July 2011), pp. 2363–2371. ISSN: 0167-9473. DOI: 10.1016/j.csda.2011.02.003. URL: http://www.sciencedirect.com/science/article/pii/S016794731100048X (visited on 04/05/2019).

[294] John C. Stansfield et al. "HiCcompare: an R-package for joint normalization and comparison of HI-C datasets". In: *BMC Bioinformatics* 19.1 (July 2018), p. 279. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2288-x. URL: https://doi.org/10.1186/s12859-018-2288-x (visited on 04/05/2019).

[295] 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. ISSN: 1088-9051. DOI: 10.1101/gr.212241.116. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5848619/ (visited on 04/05/2019).

[296] Abbas Roayaei Ardakany, Ferhat Ay, and Stefano Lonardi. "Selfish: Discovery of Differential Chromatin Interactions via a Self-Similarity Measure". en. In: *bioRxiv* (Feb. 2019), p. 540708. DOI: 10.1101/540708. URL: https://www.biorxiv.org/content/10.1101/540708v1 (visited on 04/05/2019).

[297] Job Dekker, Marc A. Marti-Renom, and Leonid A. Mirny. "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data". In: *Nature reviews. Genetics* 14.6 (June 2013), pp. 390–403. ISSN: 1471-0056. DOI: 10.1038/nrg3454. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3874835/ (visited on 04/05/2019).

[298] Luca Giorgetti et al. "Structural organization of the inactive X chromosome in the mouse". eng. In: *Nature* 535.7613 (2016), pp. 575–579. ISSN: 1476-4687. DOI: 10.1038/nature18589.

[299] S. S. Rao et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". In: *Cell* 159.7 (2014), pp. 1665–1680.

[300] Tyrone Ryba et al. "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types". eng. In: *Genome Research* 20.6 (June 2010), pp. 761–770. ISSN: 1549-5469. DOI: 10.1101/gr.099655.109.

[301] Bas van Steensel and Andrew S. Belmont. "Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression". eng. In: *Cell* 169.5 (May 2017), pp. 780–791. ISSN: 1097-4172. DOI: 10.1016/j.cell.2017.04.022.

[302] Jesse R. Dixon et al. "Chromatin Architecture Reorganization during Stem Cell Differentiation". In: *Nature* 518.7539 (Feb. 2015), pp. 331–336. ISSN: 0028-0836. DOI: 10.1038/nature14222. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4515363/ (visited on 04/05/2019).

[303] Jean-Philippe Fortin and Kasper D. Hansen. "Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data". In: *Genome Biology* 16.1 (Aug. 2015), p. 180. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0741-y. URL: https://doi.org/10.1186/s13059-015-0741-y (visited on 04/05/2019).

[304] Xiaobin Zheng and Yixian Zheng. "CscoreTool: fast Hi-C compartment analysis at high resolution". en. In: *Bioinformatics* 34.9 (May 2018), pp. 1568–1570. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx802. URL: https://academic.oup.com/bioinformatics/article/34/9/1568/4735156 (visited on 04/05/2019).

[305] Ana Pombo and Niall Dillon. "Three-dimensional genome architecture: players and mechanisms". en. In: *Nature Reviews Molecular Cell Biology* 16.4 (Apr. 2015), pp. 245–257. ISSN: 1471-0080. DOI: 10.1038/nrm3965. URL: https://www.nature.com/articles/nrm3965 (visited on 04/05/2019).

[306] Benjamin D. Pope et al. "Topologically-associating domains are stable units of replication-timing regulation". In: *Nature* 515.7527 (Nov. 2014), pp. 402–405. ISSN: 0028-0836. DOI: 10.1038/nature13986. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4251741/ (visited on 04/05/2019).

[307] Darío G. Lupiáñez et al. "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions". In: *Cell* 161.5 (May 2015), pp. 1012–1025. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.04.004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4791538/ (visited on 04/05/2019).

[308] Tom Sexton et al. "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome". In: *Cell* 148.3 (Feb. 2012), pp. 458–472. ISSN: 0092-8674. DOI: 10.1016/j.cell.2012.01.010. URL: http://www.sciencedirect.com/science/article/pii/S0092867412000165 (visited on 05/09/2019).

[309] Vuthy Ea et al. "Distinct polymer physics principles govern chromatin dynamics in mouse and Drosophila topological domains". In: *BMC Genomics* 16.1 (Aug. 2015). ISSN: 1471-2164. DOI: 10.1186/s12864-015-1786-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4536789/ (visited on 05/09/2019).

[310] Yuri B. Schwartz and Giacomo Cavalli. "Three-Dimensional Genome Organization and Function in Drosophila". eng. In: *Genetics* 205.1 (2017), pp. 5–24. ISSN: 1943-2631. DOI: 10.1534/genetics.115.185132.

[311] Quentin Szabo et al. "TADs are 3D structural units of higher-order chromosome organization in Drosophila". In: *Science Advances* 4.2 (Feb. 2018). ISSN: 2375-2548. DOI: 10.1126/sciadv.aar8082. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5829972/ (visited on 05/09/2019).

[312] Jesse R. Dixon et al. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions". In: *Nature* 485.7398 (Apr. 2012), pp. 376–380. ISSN: 0028-0836. DOI: 10.1038/nature11082. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3356448/ (visited on 04/05/2019).

[313] Emily Crane et al. "Condensin-driven remodelling of X chromosome topology during dosage compensation". en. In: *Nature* 523.7559 (July 2015), pp. 240–244. ISSN: 1476-4687. DOI: 10.1038/nature14450. URL: https://www.nature.com/articles/nature14450 (visited on 04/05/2019).

[314] Nicolas Servant et al. "HiTC: exploration of high-throughput 'C' experiments". eng. In: *Bioinformatics (Oxford, England)* 28.21 (Nov. 2012), pp. 2843–2844. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts521.

[315] Kai Kruse et al. "TADtool: visual parameter identification for TAD-calling algorithms". en. In: *Bioinformatics* 32.20 (Oct. 2016), pp. 3190–3192. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw368. URL: https://academic.oup.com/bioinformatics/article/32/20/3190/2196485 (visited on 04/05/2019).

[316] François Serra et al. "Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors". en. In: *PLOS Computational Biology* 13.7 (July 2017), e1005665. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005665. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005665 (visited on 04/05/2019).

[317] Wenbao Yu, Bing He, and Kai Tan. "Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test". En. In: *Nature Communications* 8.1 (Sept. 2017), p. 535. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00478-8. URL: https://www.nature.com/articles/s41467-017-00478-8 (visited on 04/05/2019).

[318] Koon-Kiu Yan, Shaoke Lou, and Mark Gerstein. "MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions". en. In: *PLOS Computational Biology* 13.7 (July 2017), e1005647. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005647. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005647 (visited on 04/05/2019).

[319] Sjoerd Holwerda and Wouter de Laat. "Chromatin loops, gene positioning, and gene expression". In: *Frontiers in Genetics* 3 (Oct. 2012). ISSN: 1664-8021. DOI: 10.3389/fgene.2012.00217. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3473233/ (visited on 04/05/2019).

[320] K. E. Cullen, M. P. Kladde, and M. A. Seyfred. "Interaction between transcription regulatory regions of prolactin chromatin". eng. In: *Science (New York, N.Y.)* 261.5118 (July 1993), pp. 203–206. ISSN: 0036-8075.

[321] Bas Tolhuis et al. "Looping and interaction between hypersensitive sites in the active beta-globin locus". eng. In: *Molecular Cell* 10.6 (Dec. 2002), pp. 1453–1465. ISSN: 1097-2765.

[322] Stephan Kadauke and Gerd A. Blobel. "Chromatin loops in gene regulation". In: *Biochimica et biophysica acta* 1789.1 (Jan. 2009), pp. 17–25. ISSN: 0006-3002. DOI: 10.1016/j.bbagrm.2008.07.002. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638769/ (visited on 04/05/2019).

[323] Suhas S. P. Rao et al. "Cohesin loss eliminates all loop domains". In: *Cell* 171.2 (Oct. 2017), 305–320.e24. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.09.026. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5846482/ (visited on 04/05/2019).

[324] Anders S. Hansen et al. "Recent evidence that TADs and chromatin loops are dynamic structures". In: *Nucleus* 9.1 (Dec. 2017), pp. 20–32. ISSN: 1949-1034. DOI: 10.1080/19491034.2017.1389365. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5990973/ (visited on 04/05/2019).

[325] Kim Nasmyth. "Segregating Sister Genomes: The Molecular Biology of Chromosome Separation". en. In: *Science* 297.5581 (July 2002), pp. 559–565. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1074757. URL: https://science.sciencemag.org/content/297/5581/559 (visited on 05/08/2019).

[326] Elnaz Alipour and John F. Marko. "Self-organization of domain structures by DNA-loop-extruding enzymes". eng. In: *Nucleic Acids Research* 40.22 (Dec. 2012), pp. 11202–11212. ISSN: 1362-4962. DOI: 10.1093/nar/gks925.

[327] Johannes Nuebler et al. "Chromatin organization by an interplay of loop extrusion and compartmental segregation". en. In: *Proceedings of the National Academy of Sciences* 115.29 (July 2018), E6697–E6706. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1717730115. URL: https://www.pnas.org/content/115/29/E6697 (visited on 04/04/2019).

[328] Adrian L. Sanborn et al. "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.47 (Nov. 2015), E6456–E6465. ISSN: 0027-8424. DOI: 10.1073/pnas.1518552112. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4664323/ (visited on 04/05/2019).

[329] Anton Goloborodko, John F. Marko, and Leonid A. Mirny. "Chromosome Compaction by Active Loop Extrusion". eng. In: *Biophysical Journal* 110.10 (2016), pp. 2162–2168. ISSN: 1542-0086. DOI: 10.1016/j.bpj.2016.02.041.

[330] Geoffrey Fudenberg et al. "Formation of Chromosomal Domains by Loop Extrusion". In: *Cell reports* 15.9 (May 2016), pp. 2038–2049. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2016.04.085. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4889513/ (visited on 04/05/2019).

[331] Galip Gürkan Yardımcı and William Stafford Noble. "Predictive model of 3D domain formation via CTCF-mediated extrusion". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.47 (Nov. 2015), pp. 14404–14405. ISSN: 0027-8424. DOI: 10.1073/pnas.1519849112. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4664329/ (visited on 04/05/2019).

[332] Christopher Barrington, Ronald Finn, and Suzana Hadjur. "Cohesin biology meets the loop extrusion model". en. In: *Chromosome Research* 25.1 (Mar. 2017), pp. 51–60. ISSN: 1573-6849. DOI: 10.1007/s10577-017-9550-3. URL: https://doi.org/10.1007/s10577-017-9550-3 (visited on 04/05/2019).

[333] Geoffrey Fudenberg et al. "Emerging Evidence of Chromosome Folding by Loop Extrusion". en. In: *Cold Spring Harbor Symposia on Quantitative Biology* 82 (Jan. 2017), pp. 45–55. ISSN: 0091-7451, 1943-4456. DOI: 10.1101/sqb.2017.82.034710. URL: http://symposium.cshlp.org/content/82/45 (visited on 04/05/2019).

[334] Mahipal Ganji et al. "Real-time imaging of DNA loop extrusion by condensin". en. In: *Science* 360.6384 (Apr. 2018), pp. 102–105. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aar7831. URL: http://science.sciencemag.org/content/360/6384/102 (visited on 04/05/2019).

[335] Joshua N. Burton et al. "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions". In: *Nature biotechnology* 31.12 (Dec. 2013), pp. 1119–1125. ISSN: 1087-0156. DOI: 10.1038/nbt.2727. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4117202/ (visited on 04/02/2019).

[336] Olga Dudchenko et al. "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds". In: *Science (New York, N.Y.)* 356.6333 (Apr. 2017), pp. 92–95. ISSN: 0036-8075. DOI: 10.1126/science.aal3327. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5635820/ (visited on 04/02/2019).

[337] Jay Ghurye et al. "Integrating Hi-C links with assembly graphs for chromosome-scale assembly". en. In: *bioRxiv* (Jan. 2019), p. 261149. DOI: 10.1101/261149. URL: https://www.biorxiv.org/content/10.1101/261149v2 (visited on 04/05/2019).

[338] Hervé Marie-Nelly et al. "High-quality genome (re)assembly using chromosomal contact data". In: *Nature Communications* 5.1 (Dec. 2014). DOI: 10.1038/ncomms6695. URL: https://doi.org/10.1038%2Fncomms6695.

[339] Ferhat Ay, Timothy L. Bailey, and William Stafford Noble. "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts". In: *Genome Research* 24.6 (June 2014), pp. 999–1011. ISSN: 1088-9051. DOI: 10.1101/gr.160374.113. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4032863/ (visited on 04/07/2019).

[340] Zhijun Duan et al. "A Three-Dimensional Model of the Yeast Genome". In: *Nature* 465.7296 (May 2010), pp. 363–367. ISSN: 0028-0836. DOI: 10.1038/nature08973. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2874121/ (visited on 04/07/2019).

[341] Hervé Marie-Nelly. "A probabilistic approach for genome assembly from high-throughput chromosome conformation capture data". thesis. Paris 6, Jan. 2013. URL: http://www.theses.fr/2013PA066714 (visited on 04/02/2019).

[342] Pilar Manrique, Michael Dills, and Mark Young. "The Human Gut Phage Community and Its Implications for Health and Disease". In: *Viruses* 9.6 (June 2017), p. 141. DOI: 10.3390/v9060141. URL: https://doi.org/10.3390%2Fv9060141.

[343] Pilar Manrique et al. "Healthy human gut phageome". In: *Proceedings of the National Academy of Sciences* 113.37 (Aug. 2016), pp. 10400–10405. DOI: 10.1073/pnas.1601060113. URL: https://doi.org/10.1073%2Fpnas.1601060113.

[344] Martial Marbouty et al. "Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms". In: *eLife* 3 (Dec. 2014). DOI: 10.7554/elife.03318. URL: https://doi.org/10.7554%2Felife.03318.

[345] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. DOI: 10.1088/1742-5468/2008/10/p10008. URL: https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008.

[346] Romain Campigotto, Patricia Conde Céspedes, and Jean-Loup Guillaume. "A Generalized and Adaptive Method for Community Detection". In: *arXiv:1406.2518 [physics, stat]* (June 2014). arXiv: 1406.2518. URL: http://arxiv.org/abs/1406.2518 (visited on 04/08/2019).

[347] Matthew Z. DeMaere and Aaron E. Darling. "Deconvoluting simulated metagenomes: the performance of hard- and soft- clustering algorithms applied to metagenomic chromosome conformation capture (3C)". In: *PeerJ* 4 (Nov. 2016), e2676. DOI: 10.7717/peerj.2676. URL: https://doi.org/10.7717%2Fpeerj.2676.

[348] Santo Fortunato and Marc Barthélemy. "Resolution limit in community detection". en. In: *Proceedings of the National Academy of Sciences* 104.1 (Jan. 2007), pp. 36–41. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0605965104. URL: https://www.pnas.org/content/104/1/36 (visited on 04/09/2019).

[349] Luis F. Jover, Michael H. Cortez, and Joshua S. Weitz. "Mechanisms of multi-strain coexistence in host–phage systems with nested infection networks". In: *Journal of Theoretical Biology* 332 (Sept. 2013), pp. 65–77. DOI: 10.1016/j.jtbi.2013.04.011. URL: https://doi.org/10.1016%2Fj.jtbi.2013.04.011.

[350] Christian Peeters and Serge Aron. "Evolutionary reduction of female dispersal in Cataglyphis desert ants". en. In: *Biological Journal of the Linnean Society* 122.1 (Sept. 2017), pp. 58–70. ISSN: 0024-4066. DOI: 10.1093/biolinnean/blx052. URL: https://academic.oup.com/biolinnean/article/122/1/58/3883920 (visited on 05/07/2019).

[351] P. A. Eyer et al. "Hybridogenesis through thelytokous parthenogenesis in two Cataglyphis desert ants". en. In: *Molecular Ecology* 22.4 (Feb. 2013), pp. 947–955. ISSN: 1365-294X. DOI: 10.1111/mec.12141. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12141 (visited on 05/07/2019).

[352] Laurianne Leniaud et al. "Social Hybridogenesis in the Clonal Ant Cataglyphis hispanica". In: *Current Biology* 22.13 (July 2012), pp. 1188–1193. ISSN: 0960-9822. DOI: 10.1016/j.cub.2012.04.060. URL: http://www.sciencedirect.com/science/article/pii/S0960982212005167 (visited on 05/07/2019).

[353] Raphaël Boulay et al. "Social Life in Arid Environments: The Case Study of Cataglyphis Ants". In: *Annual Review of Entomology* 62.1 (2017), pp. 305–321. DOI: 10.1146/annurev-ento-031616-034941. URL: https://doi.org/10.1146/annurev-ento-031616-034941 (visited on 05/07/2019).

[354] Hugo Darras, Laurianne Leniaud, and Serge Aron. "Large-scale distribution of hybridogenetic lineages in a Spanish desert ant". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1774 (Jan. 2014). ISSN: 0962-8452. DOI: 10.1098/rspb.2013.2396. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3843834/ (visited on 05/07/2019).

[355] Jeramiah J. Smith et al. "A chromosome-scale assembly of the axolotl genome". In: *Genome Research* 29.2 (Jan. 2019), pp. 317–324. DOI: 10.1101/gr.241901.118. URL: https://doi.org/10.1101%2Fgr.241901.118.