Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures.

Jean-François Flot, Hervé Marie-Nelly, Romain Koszul

FEBS Letters 589 (2015) 2966–2974

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures

Jean-François Flot [a], Hervé Marie-Nelly [b,c,1], Romain Koszul [b,c,*]

[a] Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK
[b] Institut Pasteur, Department of Genomes and Genetics, Groupe Régulation Spatiale des Génomes, 75015 Paris, France
[c] CNRS, UMR 3525, 75015 Paris, France

ARTICLE INFO

Article history:
Received 19 March 2015
Revised 17 April 2015
Accepted 17 April 2015
Available online 29 April 2015

Edited by Wilhelm Just

Keywords:
Contact genomics
3C
Hi-C
Genome assembly
Haplotype phasing
Metagenomics
Scaffolding

ABSTRACT

High-throughput DNA sequencing technologies are fuelling an accelerating trend to assemble *de novo* or resequence the genomes of numerous species as well as to complete unfinished assemblies. While current DNA sequencing technologies remain limited to reading stretches of a few hundreds or thousands of base pairs, experimental and computational methods are continuously improving with the goal of assembling entire genomes from large numbers of short DNA sequences. However, the algorithms that piece together DNA strands face important limitations due, notably, to the presence of repeated sequences or of multiple haplotypes within one genome, thus leaving many assemblies incomplete. Recently, the realization that the physical contacts experienced by a portion of a DNA molecule could be used as a robust and quantitative assay to determine its genomic position has led to the emerging field of contact genomics, which promises to revolutionize current genome assembly approaches by exploiting the flexible polymer properties of chromosomes. Here we review the current applications of contact genomics to genome scaffolding, haplotyping and metagenomic assembly, then outline the future developments we envision.

1. Introduction

Advances in sequencing technologies have led to a tremendous increase in the catalog of sequenced species [1]. However, although it is now relatively easy and accessible to recover a massive amount of sequences from the genome of a given species, producing a fully assembled genome sequence remains a serious challenge. This is notably because the current DNA sequencing technologies remain limited to reading stretches of only a few hundreds/thousands of base pairs. These short sequences (called reads) have to be pieced together by sophisticated computer programs called assemblers into longer stretches of continuous DNA sequences called contigs [2,3]. In an ideal world, one would recover after assembly as many contigs as there are chromosomes in the species being sequenced. However, this is hardly ever the case: most of the genome sequences published are called "unfinished" as they are heavily fragmented, whereas the number of truly "finished" genomes remains remarkably low. Only the small, compact genomes of a few so-called "model" organisms have been fully assembled until now, mostly bacteria (such as *Haemophilus influenzae* and *Escherichia coli* [4,5]) and fungi (e.g., *Saccharomyces cerevisiae* [6]) along with a single metazoan to date, the nematode *Caenorhabditis elegans* [7]. All other sequences consist of "drafts" of varying quality, including the human genome that still contains numerous gaps but is nevertheless the most complete mammalian reference assembly available [8,9]. Assembly algorithms often lead to fragmented draft assemblies for several reasons, including heterozygosity, the presence of repeated sequences of various sizes/proportions, or strong sequence composition biases. In addition, there is so far no rigorous, quantitative metric to evaluate the quality of an assembly. As a result, genome assembly remains shrouded in magic and it is typical to try a variety of assembly algorithms on a given dataset and look for the "best" solution in a semi-empirical way [10].

Today's assembly algorithms fall roughly into two large categories that employ different paradigms [11,12]. Assemblers based on the Overlap-Layout-Consensus (OLC) paradigm look for overlaps between reads that allow the gradual construction of extended sequences. These assemblers (such as CELERA [13] and MIRA [14])

* Corresponding author at: Institut Pasteur, Department of Genomes and Genetics, Groupe Régulation Spatiale des Génomes, 75015 Paris, France.
E-mail addresses: j.flot@ucl.ac.uk (J.-F. Flot), herve.marienelly@gmail.com (H. Marie-Nelly), romain.koszul@pasteur.fr (R. Koszul).
[1] Current address: Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3370, USA.

http://dx.doi.org/10.1016/j.febslet.2015.04.034
0014-5793/

were initially developed for the long, high-quality reads produced by Sanger sequencers [15]. They were subsequently adapted to deal with the shorter, more error-prone reads produced by 454 pyrosequencers [16], but the computational cost to apply them to the huge number of tiny reads (typically smaller than 150 bp [17]) produced by the new generations of sequencing machines became unbearable at the turn of the millennium. New assembly approaches were therefore explored, leading to the development of assemblers using De Bruijn Graphs (DBGs) from the work of Nicolaas de Bruijn [18]. First, the reads are split into smaller sequences of k elements (k-mers). The aim of the program is then to find the superstring of nucleotide that best recapitulates the available k-mers. In a DBG this task is achieved by finding an Eulerian path through the graph [18]. VELVET [19], ABySS [20], SOAPdenovo [21], ALLPATHS-LG [22], IDBA-UD [23] and SPAdes [24] are among the most popular algorithms using DBG representation.

Despite their empirical efficiency, these algorithms encounter a number of important limitations that in some cases strongly affect their results. Confronted with perfectly repeated sequences longer than the reads, all current approaches will lead in the best case to contig disruption and in the worst case to misassembled regions. In addition, the level of ploidy of the genome adds several layers of complexity to the problem. Finally, the outputs of existing assemblers are often fragmented and may contain many errors (ranging in size from single nucleotide substitutions to artifactual large-scale rearrangements or copy number variations) but statistical tools to robustly assess the validity of the assemblies are still missing, the development of which represents an active field of research [11,25–28].

Since no present-day assembler is able to directly produce superstrings that correspond to complete chromosomal sequences of eukaryotes, a second step called "scaffolding" is generally attempted once contigs have been generated. Scaffolding aims at ordering and orienting the contigs as accurately as possible into "supercontigs", or "scaffolds", as well as estimating the distances between them in order to generate a more global sequence backbone representative of the genome sequenced. In recent years new techniques have been developed that strongly improved this step: mate-pair sequencing [29], optical mapping [33] and single-molecule, real-time sequencing (SMRT) [30]. Mate-pair sequencing consists in cutting the genome into long DNA fragments (typically 1 to 5 kb, but sometimes up to 10–20 kb) and sequencing their extremities. The resulting pairs of sequences are therefore known to be separated by a genomic distance roughly equal to the size that was selected for, and this information can be used to detect structural variants [29] or to connect contigs over repeated regions (thereby improving *de novo* genome assembly; e.g. [31]). Optical mapping is another approach for scaffolding contigs [32–34] that can also be used to validate and/or correct contigs as well as to phase them into haplotypes [35]. In this approach, the DNA molecules of interest are first labeled with fluorescent probes directed toward specific sequences. These molecules are then stretched and elongated using microfluidic devices so that optical imaging of the probes allows determining the relative positions of their target sequences. Optical mapping has been successfully used in several genome sequencing projects, such as the domestic goat [36] and rice [37]. Finally, SMRT sequencing [30] generates long reads (up to 20 kb long) that can be used either to generate *de novo* assemblies of small genomes [38] and to scaffold contigs and fill up gaps in scaffolds obtained from large genomes [8], thereby efficiently solving many of the problems posed by repeats in assembling genomes.

Although mate-pair sequencing, optical mapping and SMRT sequencing alleviate some of the problems posed by repeats and structural complexity, they are usually unable to solve them all. First, mate pairs can only bridge regions up to a few tens of kilobases long and cannot solve complex structural variations easily; besides, mate-libraries are usually contaminated with erroneous paired-end reads, leading to even more misassemblies. Second, optical mapping requires a complex and costly experimental set-up not readily accessible to many labs involved in genomic projects, and this approach is unable to order and orient small contigs. Last, SMRT sequencing is plagued with a high error rare (about 15%, mostly indels), because of which even 20-kb reads may not map unambiguously to a single genomic location: notably, this approach cannot solve gaps caused by large repeats (>20 kb) of nearly identical sequences [8]. Overall, improving the quality of an assembly remains fastidious, time-consuming, and costly: as a result, *de novo* draft genomes usually contain numerous errors and gaps, including some that users may not be aware of. Therefore, new methods are actively sought that would allow the *de novo* assembly of finished genomes, using objective, quantitative and hypotheses-free approaches.

Over the last year, approaches have been proposed that exploit the three-dimensional (3D) physical signature of chromosomes to bring a new level of resolution to scaffolding and haplotyping as well as to metagenomic assembly. As these genomic approaches exploit the quantification of 3D contacts along the chromosomes, we dub this burgeoning field "contact genomics". These new techniques rely primarily on chromosome conformation capture (3C), which was originally developed to characterize the average 3D organization of chromosomes (see accompanying reviews; [39]). In the present review, we first introduce the supporting theory behind these methods before detailing their practical applications to genome scaffolding. We then present briefly contact genomic approaches to haplotyping and to metagenomic assembly, and conclude by outlining the future developments we envision in this field.

2. The theoretical foundations of contact genomics

As mentioned above, a typical assembly program generates a set of contigs that are subsequently scaffolded in an attempt to approximate the complete sequences of the chromosomes. Unlike mate-pair sequencing, optical mapping and SMRT sequencing, contact frequency data potentially provide a full spectrum of distances ranging from local to chromosome scale. Besides, because of the flexible nature of the chromatin fiber, loci that are in close proximity along its sequence are expected to interact much more than others that are farther apart [40,41]: hence, quantifying these contacts using genomic derivatives of 3C [42–44] makes it possible to estimate interaction frequencies between all loci within a genome and from there to infer genomic distances. Although many 3C derivatives exist (most notably Hi-C [42], but also 3C-seq [45] and Chicago [46]), they will all be referred to as "3C" in this review unless specified otherwise.

Typically, 3C starts with a crosslinking step that aims at "freezing" the organization of all the cellular components within a population of cells, including the chromosomes [39]. Crosslinked cells are incubated in the presence of a restriction enzyme, and the resulting complexes of proteins and DNA restriction fragments (RFs) are then ligated intramolecularly. The more frequently RFs are trapped together (because of their spatial proximity during crosslinking), the more likely they are to become ligated to one another and generate a molecule that is chimeric with respect to the genome sequence. The quantification of these religation events is currently best achieved through high-throughput paired-end sequencing of the 3C library, allowing the computation of detailed contact matrices (or contact maps) that reflect the contact frequencies between all the RFs in a genome [42,43].

In all organisms studied, genome-wide contact maps display a strong diagonal signal reflecting the frequent 3D contacts between

RFs located near each other along the chromosome(s) (Fig. 1a). As a result, one may assume a direct relationship between genomic distance and interaction frequency: loci that are in close proximity to each other along the chromosomes interact frequently, yielding a strong 3C signal, and reciprocally, strong 3C signals imply close genomic proximity. This relationship makes it possible to use contact genomics to establish the synteny (i.e., collinearity) of DNA loci across large distances, hence overcoming the current limitations of scaffolding, haplotyping, and even metagenomics analyses. In other words, the synteny information contained in a 3C library is similar to what mate-pair libraries can bring but spans distances that are up to 2 or 3 orders of magnitude larger, therefore potentially connecting loci across the entire length of each chromosome.

3. Application of contact genomics to scaffolding

The most direct application of contact genomic is to scaffold genome data, for example in order to identify large-scale chromosomal structural variations. When paired-end reads from a genomic 3C library are mapped on a reference genome sequence, strong incongruities (i.e., 3D signals outside of the expected diagonal) in the contact map are indicative of structural differences (Fig. 1b*i*), as was for instance noted in studies of oncogenic cell lines [47]. Like in a jigsaw puzzle, reordering the pieces (Fig. 1b*ii*) and reorienting them (Fig. 1b*iii*) to minimize the amount of incongruities results in a reconstruction of the true genome structure of the isolate that was sequenced. In this application, contact genomics provides strong hints about the connections between the contigs, revealing both their order and their orientation with respect to one another. In simple cases such as the example shown on Fig. 1b, the resulting jigsaw puzzle can be easily solved in a visual way thanks to the obvious incongruities in the pattern. However, the complexity of this procedure increases non-linearly with the number of contigs and assembly errors.

A simple, intuitive approach to improve a complex assembly using contact data is a "greedy", recursive algorithm that finds the best neighbors of a DNA region based on their contact frequencies. This approach represents each contig as an ordered string of oriented RFs, with each fragment having at most two adjacent (left and right) neighbors. The two RFs that interact most frequently with a given fragment are determined then recursively connected to each other until an incompatibility arises. Such a local method discards most of the long-range contact information contained in the 3C data: hence, although greedy approaches may perform well on ideal, simulated data without repeated elements, their performance drops quickly when data get sparser and genomes get more complex [48].

Things become even more complicated if simulations include statistical variations reflecting the fact that 3C is first and foremost a counting procedure. Indeed, 3C experiments quantify a signal that results from two complex, overlapping stochastic processes: first, the multistep experimental protocols used generate biases and artifacts (possibly linked directly to the DNA sequence itself) that must be taken into account when interpreting the result

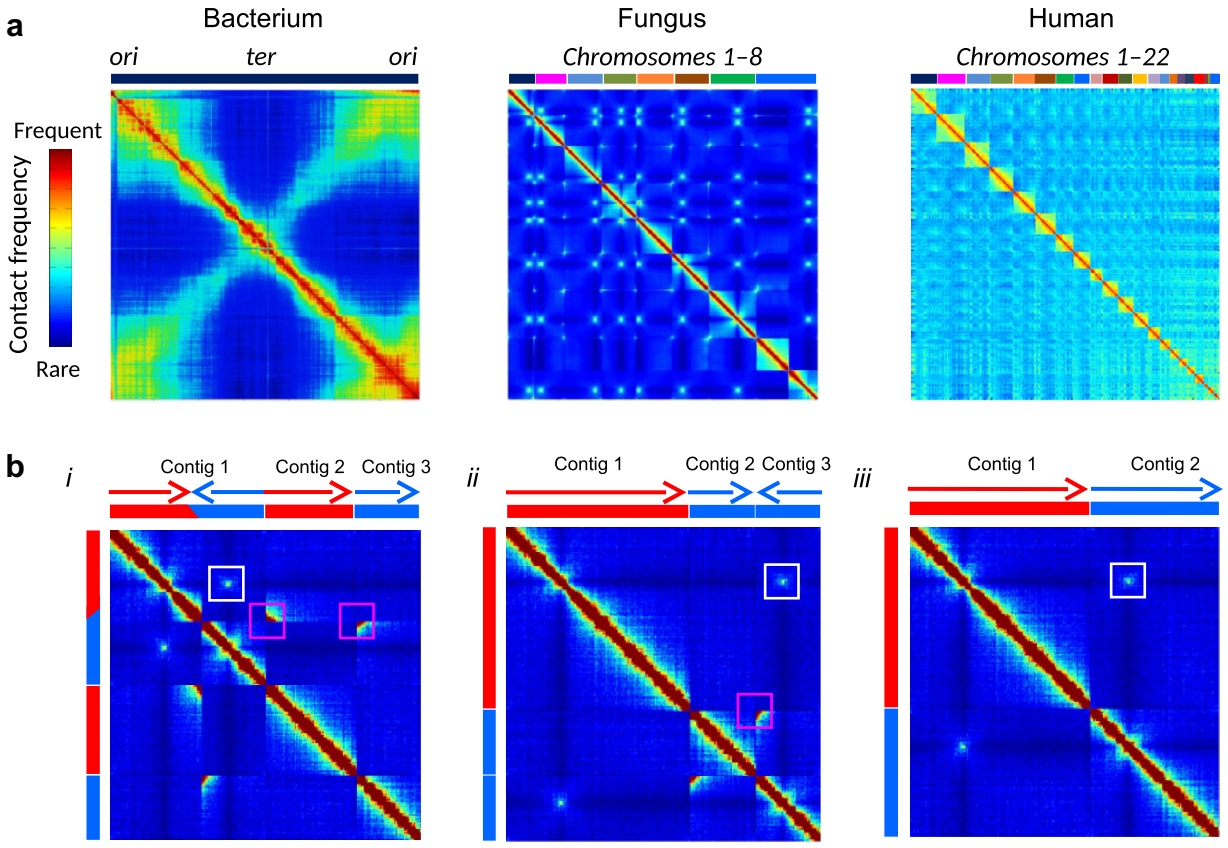

Fig. 1. Principle of genome assembly using chromosome contact data. (a) The flexible polymer properties of the DNA molecule explain the strong diagonal observed in the genomewide contact maps of all species studied using genomic 3C derivatives so far, as illustrated here for the genomes of *Bacillus subtilis* (a bacterium), *Naumovozyma castellii* (a fungus) and *Homo sapiens*. (b) 3C contact data mapped on the *de novo* assembly of chromosomes 4 (red bars) and 15 (blue bars) of the yeast *Saccharomyces cerevisiae* (panel *i*). The *de novo* assembly comprised errors and was fragmented, but one could easily re-order the fragments to produce an intermediate contact map (panel *ii*) in which the assembly errors were corrected, then from there scaffold the contigs in order to retrieve the correctly assembled genome (panel *iii*). In this experiment, the diagonal signal was two orders of magnitude stronger than the signal originating from inter-centromeric repeats (white squares), whereas the other extra-diagonal signals initially detected (pink squares) disappeared as the initial assembly was corrected and scaffolded.

[49–51]. Second, the experiments are performed on dynamic objects: chromosomes are dynamic polymers whose physical properties are likely to vary in time, in space, and even locally over their monomers [39,43,52,53]. Importantly, in the near-perfect situation of a 3C dataset simulated by considering the experiment as the output of a Poisson process, recursive algorithms fail to reconstruct the original contigs [48]. This failure illustrates the fact that even when there are no experimental artifacts and a hypothetical "shortest common supersequence" does exist, the raw contact counts cannot be used directly as a robust indication that two restriction fragments are located on the same chromosome. Therefore, the main challenge when using contact genomics to scaffold genomes appears to distinguish true interactions from background and statistical noise in order to reorder and reorient DNA regions properly. To reach this aim, the algorithms described in rest of this section adopt different strategies to handle contact data and generate outputs; all of them succeed in improving scaffold sizes by several orders of magnitude, leading for instance to scaffolds spanning the entire length of human chromosomes.

3.1. Clustering methods

Clustering solutions offer a quick and practical approach to group DNA contigs or fragments that are likely to be in the vicinity of each other because they are part of the same chromosome. To produce scaffolds, this first step has to be followed by a second one that aims at ordering and, ideally, orienting these DNA segments with respect to each other within a cluster (Fig. 2a). The programs dnaTri (the name of which stands for "DNA triangulation"; [54]) and Lachesis [55] use this strategy to explore the ability of genomic 3C to scaffold human chromosomes. Notable differences exist between the two approaches. For instance, Lachesis necessitates prior knowledge of the expected number of clusters to proceed, and this approach often clusters small chromosomes together. dnaTri, on the other hand, applies an average-linkage hierarchical clustering algorithm directly to a distance matrix approximated from the contact matrix, without making any prior assumption regarding the expected number of clusters. Despite these differences, these two programs were reported to successfully scaffold both simulated and *de novo* contigs into full-length chromosomes, paving the way for further development and applications.

Although clustering approaches are clear improvements over the greedy approach mentioned previously, several limitations remain. Notably, their two-step process potentially results in cumulative errors: unless specific care is taken to tackle such problems, a contig misplaced during the clustering step will not be reassigned to its correct chromosome during the second step. Also, these programs do not account for duplications and do not attempt to correct the assembly errors that may be present in the contigs that are fed to them.

3.2. Probabilistic methods

Alternatively, genome assembly based on contact data can be approached from a probabilistic perspective. Probabilistic approaches using Bayesian inference provide a robust framework to assess the validity of a genome in an objective and quantitative fashion [56]. Such an approach was implemented in the GRAAL program [57], which uses the highly redundant information encapsulated in genomic 3C data together with an analytic model inspired from polymer physics to compute the likelihood of a genomic structure (namely, the probability of observing the contact matrix at hand given the genome structure being evaluated). The program is initialized with a set of DNA sequences, a pool of "bins" from which the program repeatedly draws. For each bin, the program uses 3D contacts to find candidate neighbors among the other bins (Fig. 2b), then determines their most likely relationships within the genome by testing a large number of biologically inspired structural variations (including duplications, inversions, etc.; Fig. 2c). For each structure the program computes a likelihood score, and one of the structures with the highest scores is retained for the next iteration. As a result, upon thousands of iterations the procedure converges toward what is expected to be the most likely structure given the data. Once initialized with a set of contigs and the contact data, the program iterates automatically without further user intervention, with each bin being processed as many times as defined by the user (Fig. 2a). GRAAL was validated on both human and fungal genomes, and on both simulated and *de novo* datasets [45,57]. One disadvantage is that, at least in its current implementation, it does not attempt to guess the size of the remaining gaps in an assembly. Another assembler using a probabilistic approach based on likelihood comparisons was recently published (HiRISE; [46]) and appears similar to GRAAL in term of its capabilities (Fig. 2a), but a detailed comparison of the performance of these two programs is still wanting.

4. Application of contact genomics to chromosome-scale haplotyping

Most animal and plant species have diploid or polyploid genomes, the characterization of which poses challenges far beyond those of haploid organisms such as bacteria. Characterizing the genetic variations along all homologous chromosomes/sequences present in a diploid (or polyploid) cell is important not only for biomedical applications such as linkage analyses [58,59] but also for population genomics and evolutionary studies [60]. However, haplotype reconstruction remains limited by current sequencing technologies, by cost, and, in many instances, by the lack of robust genome scaffolds [61]. Here again, genome 3D physical signatures open new perspectives to phase single-nucleotide polymorphisms (SNPs) and indels among homologous chromosomes as well as to discriminate among different paralogous copies arising from copy-number variations. The underlying principle of these approaches remains the same as for scaffolding: physical linkage, i.e. the fact that two nucleotide variants are carried by the same chromosome and therefore belong to the same haplotype, can be assessed based on the frequency of the contacts between these positions. The basic tenet is that two variants present on the same haplotype (in *cis* positions) are much more likely, up to a certain distance, to be captured together in a 3C experiment than variants present on the two different haplotypes (in *trans* positions). Instead of ordering restriction fragments as in scaffolding applications, one can therefore use 3D contacts as long-distance anchors to cluster SNPs or indels, thereby unveiling the haplotypes (Fig. 3a).

Several teams recently investigated the potential of contact genomics for resolving haplotypes. Bing Ren and colleagues performed a Hi-C experiment on a diploid mouse cell line with two homologous chromosomal sets originating from homozygous strains whose genome sequences were already known [62]. As expected, most Hi-C contacts resulted from *cis* interactions (as a result of the spatial segregation of chromosomes within nuclei). By adapting the HapCUT program (originally developed to phase haplotypes from shotgun or mate-pair data [63]) to exploit the broader range of long-distance contacts generated by Hi-C libraries, they successfully phased *de novo* more than 99% of the known heterozygous sites along each chromosome in their mouse system. When applied to a human cell line carrying about ten times fewer heterozygous sites than their mouse strain, their

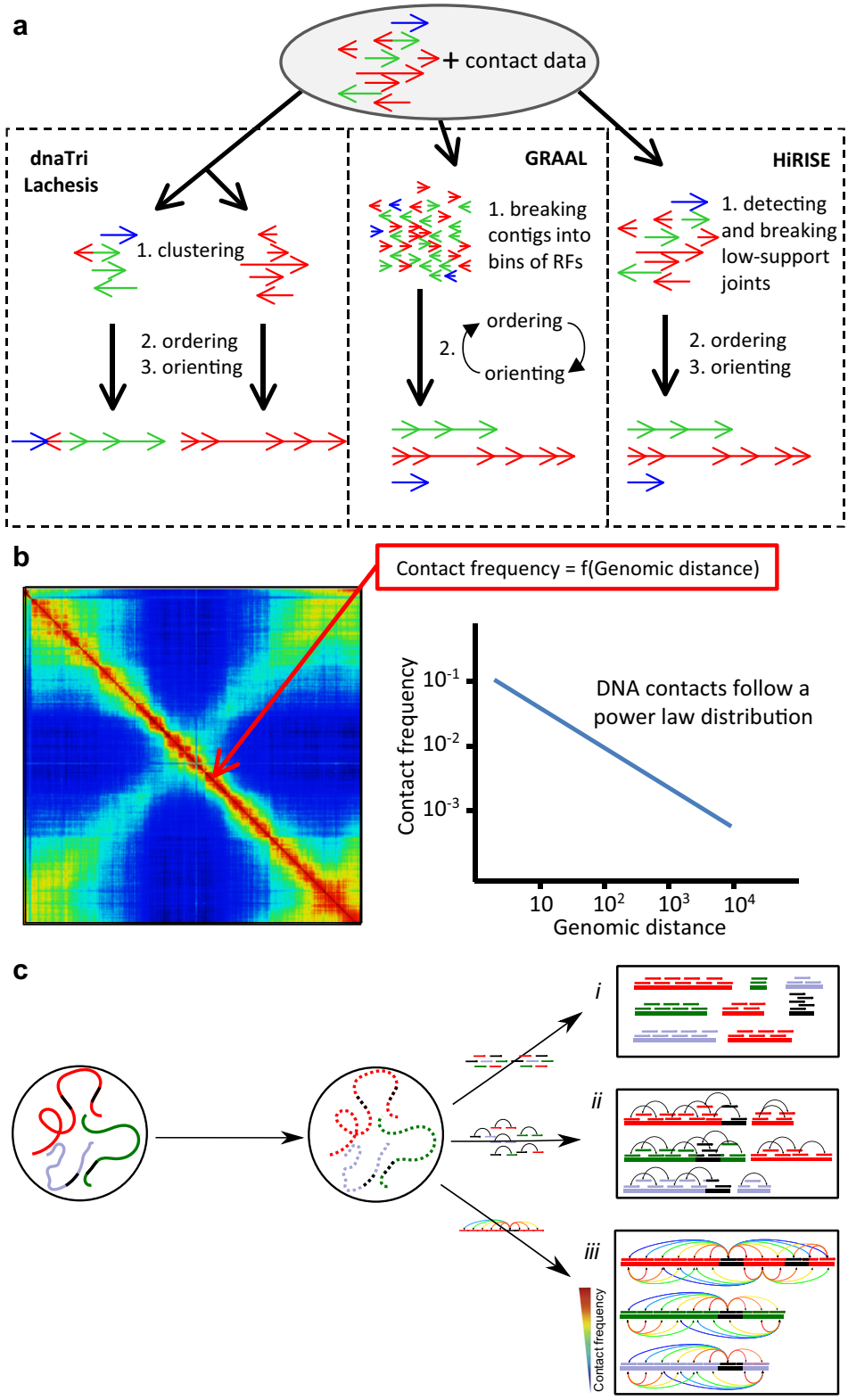

Fig. 2. Scaffolding application of contact genomics. (a) Simplified representation of the pipelines used by published algorithms to perform contact genomic scaffolding (dnaTri, Lachesis, GRAAL and HiRISE [46,54,55,57]). Blue, red, and green arrows represent contigs/scaffolds from an assembly presenting discrepancies with the genome of the species or cell line from which 3C data were obtained. (b) Left: contact map of a bacterial chromosome. Right: when plotted against genomic distances, contact frequencies between DNA regions exhibit a power law. This distribution can vary quantitatively depending on the experimental conditions or species, but its overall shape remains highly conserved. In the absence of contigs long enough to compute a distribution over large distances, one can initialize an assembly algorithm using a published distribution and then gradually replace it with one inferred from the actual dataset at hand. (c) Schematic representation of scaffolding using contact frequency distributions. Duplications within genomes (black lines) can be identified based on their contacts with their neighboring regions and then repositioned correctly using adequate algorithms such as GRAAL.

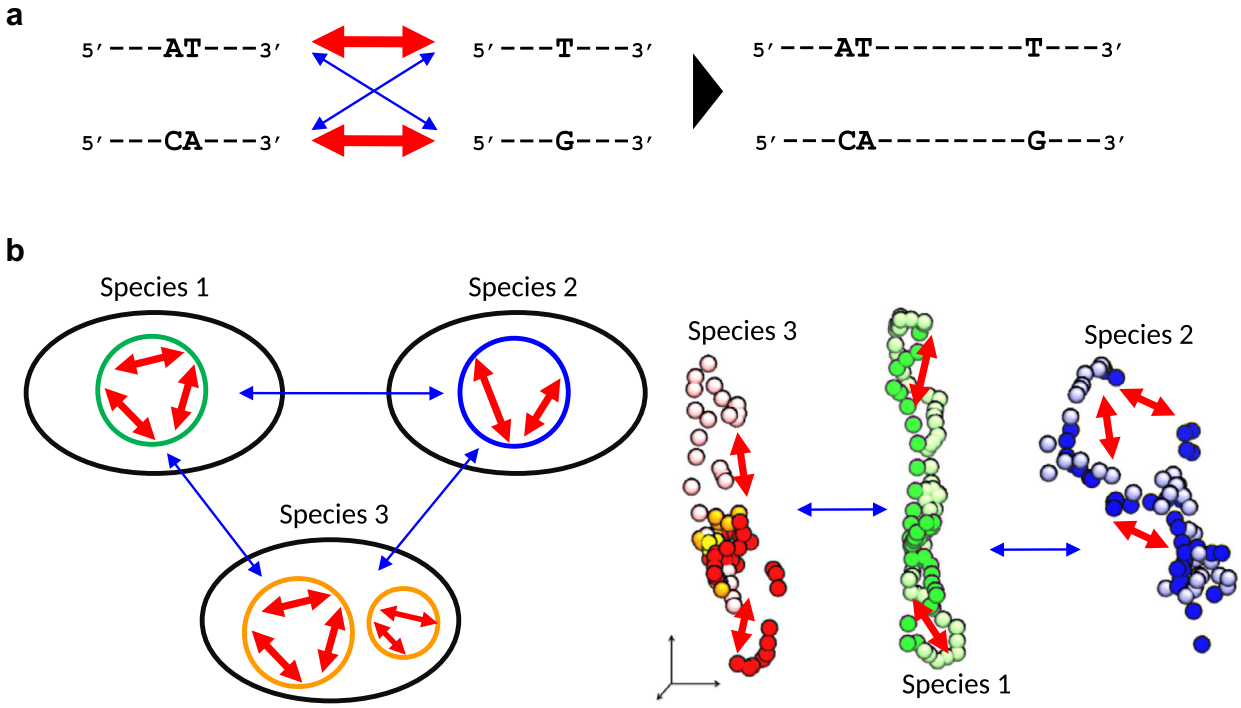

Fig. 3. Haplotyping and metagenomic applications of contact genomics. (a) Example of haplotype deconvolution based on 3D contacts. Two genomic variants occurring in *cis* will exhibit more contacts (red arrows) than variant positions located in *trans* on two different chromosomes (blue arrows). The syntenic variations can then be identified and positioned appropriately. (b) Illustration of the application of contact genomics to metagenomics. A metagenomic 3C (meta3C) experiment performed directly onto a mix of species reveals that 3D contacts are more frequent between DNA regions belonging to the same cellular compartment (red arrows) than between chromosomal sets in different compartments (blue arrows). This discrepancy can be exploited to distinguish the different chromosomal sets in each compartment, thereby separating the genomes of the different species present into the mix. Right panel: 3D reconstruction of the contact matrix recovered from an experiment performed on a controlled mix of species [45]. Each bead represents a ~30 kb DNA region and is positioned according to its contacts with the other beads. Each cluster of beads that appears on the figure corresponds to one species, illustrating the low amount of noise in these experiments.

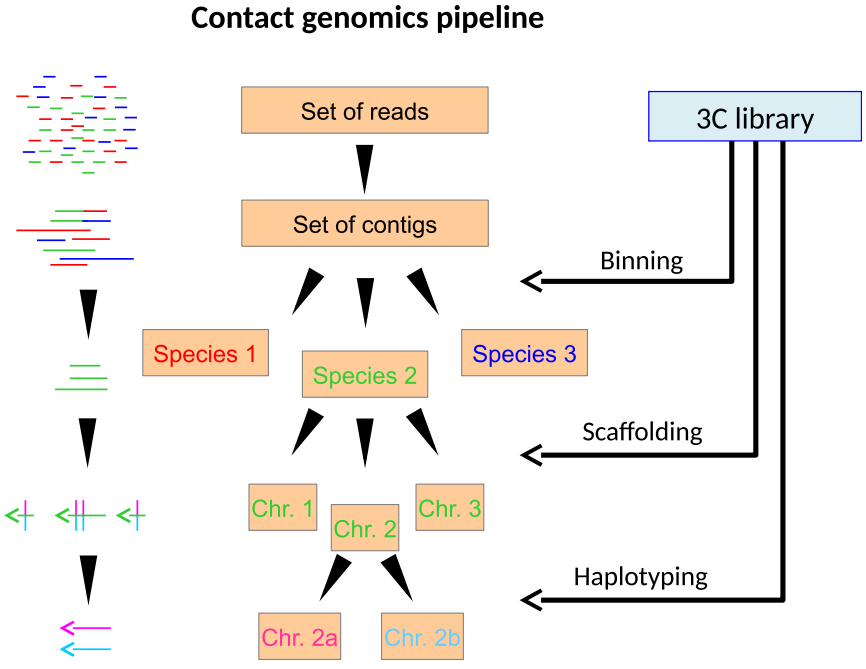

Fig. 4. Integrated contact genomics pipeline, from metagenomic assembly to scaffolding to haplotyping. The schematic representation illustrates how, from a set of DNA sequences recovered from a mixed population, one could theoretically exploit DNA physical contacts to scaffold the chromosomes of the species present in the mixture and phase their haplotypes. The color code illustrates how this process is gradually achieved.

approach (called HaploSeq) was able to resolve haplotypes with an average accuracy of 98% [62]. Another 3C derivative, targeted-locus amplification (TLA), focuses on specific regions of the genome and allows identification of structural variations, SNPs and other variations affecting the surroundings of these positions of interest [64].

5. Application of contact genomics to metagenomics

From the applications described above, it is easy to see how the principles of contact genomics can be extended to the analysis of genomes of different species cohabiting together, i.e. "metagenomes". When performing a 3C experiment directly onto a mix of species, one generally observes a very low frequency of intergenomic ligation events, which makes it possible to use 3D contact signatures to distinguish DNA segments of chromosomes belonging to different organisms: instead of scaffolding contigs into chromosomes from a single genome, one should simply consider individual genomes as DNA entities to be characterized based on their 3D contacts within a metagenome (Fig. 3b, left). Three studies have recently provided proofs-of-concept experiments showing that genome-wide 3C of a controlled mixed population can indeed generate sufficient contact information to infer the genomes of the species present in the mix [45,65,66] (Fig. 3b, right). However, these published approaches used general contact-genomic algorithms and not algorithms dedicated to metagenomics, leaving ample room for future improvements. Two of these proof-of-concept studies were performed using simulated contigs or prior assemblies of separate shotgun libraries as templates on which to align 3C reads [65,66], whereas the third article exploited the fact that 3C-seq libraries contain about 80% regular paired-end reads to generate contigs that were subsequently clustered and reassembled using GRAAL (during which some assembly errors present in the initial contigs were corrected) [45]. Importantly, contact genomics also allows detection of extra-genomic elements sharing the same compartment as a given genome, as was shown both for controlled mixes of species [45,65] and for a complex microbial community isolated from the environment [45]. For instance, a correlation analysis of the 3D contacts originating from a F plasmid (the fertility factor allowing bacterial conjugation [67]) detected in 3C data from a mix of three bacterial species revealed not only that this plasmid belonged to *E. coli*, but also that it carried a 140-kb copy of a portion of the genome of this bacterium [45]. A similar analysis performed on bacteriophage sequences in the same dataset also revealed which ones among these elements were extra-chromosomal and which ones had become integrated as prophages in the genome (RK, unpublished data). Finally, the genome haplotyping strategies described above may also be applied to phase closely related genomic variants occurring in a metagenome [65].

6. Conclusion

Taking advantage of the spatial signature of chromosomes to improve genomic analysis holds important promises, but these may shift in light of continuous technological developments. For instance, novel sequencing technologies such as nanopore membranes may alleviate the remaining challenges encountered to "fill the gaps" in repeated or otherwise complex regions of genomes [68]. However, we envision that the emerging contact genomics approaches described in this review will remain important for several applications. First, physical contacts make it possible to assess the quality of an assembly using an objective, independent source of information and to correct errors in the assembly [46,57]. Second, the application of contact genomics to haplotype resolution is likely to develop in the future, not only for single genomes but also for metagenomic analysis and for characterizing the multiple strains within a population of a given species.

Originally, contact genomics analyses were performed using genomic 3C datasets generated *in vivo*. Emancipation from the sometimes complex manipulation of living cells by performing 3C directly onto purified DNA *in vitro* appears a natural extension of this approach, which may follow either one of two possible paths. The first improvement would be to develop chemicals that crosslink the DNA molecule itself. To our knowledge, few chemicals have been synthesized and specifically used to perform interhelical DNA–DNA crosslinking. One such chemical was used in the early 1980s to study the packaging of the lambda phage genome [69,70]. The synthesis of this product remains fastidious, but it may be possible to develop crosslinking chemicals that are easier to synthesize, such as two intercalation molecules linked together by a long carbonate chain. Another alternative consists in simply reconstituting chromatin *in vitro* by mixing the molecules of interest with histones, which can be achieved using commercial kits. The latter approach was recently applied to DNA isolated from human and alligator with apparently good results [46] (although it remains difficult to assess its efficiency relative to other published approaches since the available preprint does not include such comparison). One potential advantage of using an *in vitro* procedure is to remove the 3D signal induced by biologically meaningful contacts (such as the clustering of centromeres in yeast (see Fig. 1b) and gene loops in mammals), as the latter may interfere with the signal originating from the linear structure of chromosomes. However, these contacts do not seem to present a challenge to scaffolding algorithms such as GRAAL or dnaTri [54,57], given that they have been reported to successfully scaffold hundreds of kilobases of individual chromosomes. Another potentially interesting feature of *in vitro* approaches is that they do not require living tissues but can be applied on mere DNA extracts, which is certainly advantageous when performing *post mortem* analyses. The advantage of *in vivo* experiments, on the other hand, is that beyond scaffolding, haplotyping and metagenomic analysis they also provide insights into the 3D structure of the genomes under scrutiny [45,57]. This structure, in turn, can eventually reveal the positions of functional elements such as point centromeres [71,72]. In addition, it is possible that the *in vivo* packaging of chromatin in the Hi-C and 3C-seq approaches improves the capture and assembly of long, repeated elements by increasing long distance contacts.

Beyond refinements in experimental procedure, we expect contact genomics to benefit greatly from the development of dedicated software. At present, the steps of assembling the reads into contigs, scaffolding them and phasing them are performed by different programs that may not take into account all the information available for each step (e.g., 3D contacts are not taken into account during contig formation) and that may have different input/output formats. Hence, developing a single, user-friendly program able to take as input both regular paired-end and mate-pair reads as well as 3C reads and possibly other type of information (such as PacBio and nanopore reads) then to assemble, scaffold and phase them using an explicitly probabilistic framework such as the one of GRAAL should be a priority direction for future research (Fig. 4). From a more technical viewpoint, the exploitation of graphic processors units (GPU [57]), combined with the development of new mathematical treatments of contact matrixes, will likely prove essential to the democratization of these methods by allowing them to run on cheaper computers.

Acknowledgments

The authors thank Ronnie de Jonge and Ken Kraaijeveld for fruitful comments. This research was supported by funding to R.K. from the European Research Council under the 7th Framework Program (FP7/2007-2013)/ERC grant agreement 260822. J.F.F. is supported by the European Research Council (ERC-2012-AdG 322790).

References

[1] Pagani, I., Liolios, K., Jansson, J., Chen, I.-M.A., Smirnova, T., Nosrat, B., Markowitz, V.M. and Kyrpides, N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40, D571–D579.

[2] Myers, E.W. (1995) Toward simplifying and accurately formulating fragment assembly. J. Comput. Biol. 2, 275–290.

[3] Idury, R.M. and Waterman, M.S. (1995) A new algorithm for DNA sequence assembly. J. Comput. Biol. 2, 291–306.

[4] Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1462.

[5] Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. and Al, E. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

[6] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Life with 6000 genes. Science 274 (546), 563–567.

[7] C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012–2018.

[8] Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J.M., Stamatoyannopoulos, J.A., Hunkapiller, M.W., Korlach, J. and Eichler, E.E. (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517, 608–611.

[9] Consortium, I.H.G.S (2004) Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

[10] Richards, S. and Murali, S.C. (2015) Best practices in insect genome sequencing: what works and what doesn't. Curr. Opin. Insect Sci. 7, 1–7.

[11] Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified. Nat. Rev. Genet. 14, 157–167.

[12] Miller, J.R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. Genomics 95, 315–327.

[13] Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H.J., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.-H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D. and Venter, J.C. (2000) A whole-genome assembly of *Drosophila*. Science 287, 2196–2204.

[14] Chevreux, B., Wetter, T. and Suhai, S. (1999) Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, 45–56.

[15] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463–5467.

[16] Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

[17] Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J. and Pallen, M.J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30, 434–439.

[18] Compeau, P.E.C., Pevzner, P.A. and Tesler, G. (2011) How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29, 987–991.

[19] Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.

[20] Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. and Birol, İ. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.

[21] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W. and Wang, J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience 1, 18.

[22] Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S. and Jaffe, D.B. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. 108, 1513–1518.

[23] Peng, Y., Leung, H.C.M., Yiu, S.M. and Chin, F.Y.L. (2012) IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428.

[24] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. and Pevzner, P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

[25] Howison, M., Zapata, F. and Dunn, C.W. (2013) Toward a statistically explicit understanding of de novo sequence assembly. Bioinformatics 29, 2959–2963.

[26] Howison, M., Zapata, F., Edwards, E.J. and Dunn, C.W. (2014) Bayesian genome assembly and assessment by Markov chain Monte Carlo sampling. PLoS One 9, e99497.

[27] Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Fabbro, C.D., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B. and Ho, I.Y. (2013) Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. GigaScience 2, 10.

[28] Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., Marçais, G., Pop, M. and Yorke, J.A. (2011) GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22, 557–567.

[29] Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C.E., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M. and Snyder, M. (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420–426.

[30] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138.

[31] Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W. (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579.

[32] Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M. and Kwok, P.-Y. (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. 30, 771–776.

[33] Zhou, S., Deng, W., Anantharaman, T.S., Lim, A., Dimalanta, E.T., Wang, J., Wu, T., Chunhong, T., Creighton, R., Kile, A., Kvikstad, E., Bechner, M., Yen, G., Garic-Stankovic, A., Severin, J., Forrest, D., Runnheim, R., Churas, C., Lamers, C., Perna, N.T., Burland, V., Blattner, F.R., Mishra, B. and Schwartz, D.C. (2002) A whole-genome shotgun optical map of *Yersinia pestis* strain KIM. Appl. Environ. Microbiol. 68, 6321–6331.

[34] Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J. and Wang, Y.K. (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. Science 262, 110–114.

[35] Anantharaman, T.S., Anantharaman, T.S., Mysore, V., Mysore, V., Mishra, B., and Mishra, B. (2005) Fast and cheap genome wide haplotype construction via optical mapping, in: Pacific Symposium on Biocomputing, pp. 385–396.

[36] Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., Tosser-Klopp, G., Wang, J., Yang, S., Liang, J., Chen, W., Chen, J., Zeng, P., Hou, Y., Bian, C., Pan, S., Li, Y., Liu, X., Wang, W., Servin, B., Sayre, B., Zhu, B., Sweeney, D., Moore, R., Nie, W., Shen, Y., Zhao, R., Zhang, G., Li, J., Faraut, T., Womack, J., Zhang, Y., Kijas, J., Cockett, N., Xu, X., Zhao, S., Wang, J. and Wang, W. (2013) Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). Nat. Biotechnol. 31, 135–141.

[37] Zhou, S., Bechner, M.C., Place, M., Churas, C.P., Pape, L., Leong, S.A., Runnheim, R., Forrest, D.K., Goldstein, S., Livny, M. and Schwartz, D.C. (2007) Validation of rice genome sequence by optical mapping. BMC Genomics 8, 278.

[38] Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W. and Korlach, J. (2013) Non-hybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10, 563–569.

[39] Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. Science 295, 1306–1311.

[40] Rippe, K. (2001) Making contacts on a nucleic acid polymer. Trends Biochem. Sci. 26, 733–740.

[41] Gennes, P.-G. (1979) Scaling Concepts in Polymer Physics, Cornell University Press, Ithaca, NY.

[42] Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S. and Dekker, J. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293.

[43] Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A. and Noble, W.S. (2010) A three-dimensional model of the yeast genome. Nature 465, 363–367.

[44] De Laat, W. and Dekker, J. (2012) 3C-based technologies to study the shape of the genome. Methods 58, 189–191.

[45] Marbouty, M., Cournac, A., Flot, J.-F., Marie-Nelly, H., Mozziconacci, J. and Koszul, R. (2014) Metagenomic chromosome conformation capture (meta3C)

unveils the diversity of chromosome organization in microorganisms. eLife 3, e03318.
[46] Putnam, N.H., O'Connell, B., Stites, J.C., Rice, B.J., Fields, A., Hartley, P.D., Sugnet, C.W., Haussler, D., Rokhsar, D.S., and Green, R.E. (2015) Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage, ArXiv 150205331 Q-Bio.
[47] Rickman, D.S., Soong, T.D., Moss, B., Mosquera, J.M., Dlabal, J., Terry, S., MacDonald, T.Y., Tripodi, J., Bunting, K., Najfeld, V., Demichelis, F., Melnick, A.M., Elemento, O. and Rubin, M.A. (2012) Oncogene-mediated alterations in chromatin conformation. Proc. Natl. Acad. Sci. 109, 9083–9088.
[48] Marie-Nelly, H. (2013). A probabilistic approach for genome assembly from high-throughput chromosome conformation capture data (Ph.D. dissertation). Université Pierre et Marie Curie – Paris 6.
[49] Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. and Mozziconacci, J. (2012) Normalization of a chromosomal contact map. BMC Genomics 13, 436.
[50] Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods 9, 999–1003.
[51] Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. 43, 1059–1065.
[52] Dekker, J. (2007) GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol. 8, R116.
[53] Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A. and Dekker, J. (2013) Organization of the mitotic chromosome. Science 342, 948–953.
[54] Kaplan, N. and Dekker, J. (2013) High-throughput genome scaffolding from *in vivo* DNA interaction frequency. Nat. Biotechnol. 31, 1143–1147.
[55] Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. (2013) Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat. Biotechnol. 31, 1119–1125.
[56] Rieping, W., Habeck, M. and Nilges, M. (2005) Inferential structure determination. Science 309, 303–306.
[57] Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D.P., Syan, S., Guillén, N., Margeot, A., Zimmer, C. and Koszul, R. (2014) High-quality genome (re)assembly using chromosomal contact data. Nat. Commun. 5, 5695.
[58] Gillanders, E.M., Pearson, J.V., Sorant, A.J.M., Trent, J.M., O'Connell, J.R. and Bailey-Wilson, J.E. (2006) The value of molecular haplotypes in a family-based linkage study. Am. J. Hum. Genet. 79, 458–468.
[59] Akey, J., Jin, L. and Xiong, M. (2001) Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? Eur. J. Hum. Genet. 9, 291–300.
[60] Flot, J.-F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E.G.J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.-M., Barbe, V., Barthélémy, R.-M., Bast, J., Bazykin, G.A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, I., Rodriguez, F., Ryan, J.F., Vakhrusheva, O.A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A.S., Mark Welch, D.B., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O. and Van Doninck, K. (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. Nature 500, 453–457.
[61] Bonizzoni, P., Vedova, G.D., Dondi, R. and Li, J. (2003) The haplotyping problem: an overview of computational models and solutions. J. Comput. Sci. Technol. 18, 675–688.
[62] Selvaraj, S., Dixon, J.R., Bansal, V. and Ren, B. (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol. 31, 1111–1118.
[63] Bansal, V. and Bafna, V. (2008) HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 24, i153–i159.
[64] de Vree, P.J.P., de Wit, E., Yilmaz, M., van de Heijning, M., Klous, P., Verstegen, M.J.A.M., Wan, Y., Teunissen, H., Krijger, P.H.L., Geeven, G., Eijk, P.P., Sie, D., Ylstra, B., Hulsman, L.O.M., van Dooren, M.F., van Zutven, L.J.C.M., van den Ouweland, A., Verbeek, S., van Dijk, K.W., Cornelissen, M., Das, A.T., Berkhout, B., Sikkema-Raddatz, B., van den Berg, E., van der Vlies, P., Weening, D., den Dunnen, J.T., Matusiak, M., Lamkanfi, M., Ligtenberg, M.J.L., ter Brugge, P., Jonkers, J., Foekens, J.A., Martens, J.W., van der Luijt, R., van Amstel, H.K.P., van Min, M., Splinter, E. and de Laat, W. (2014) Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nat. Biotechnol. 32, 1019–1025.
[65] Beitel, C.W., Froenicke, L., Lang, J.M., Korf, I.F., Michelmore, R.W., Eisen, J.A. and Darling, A.E. (2014) Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ 2, e415.
[66] Burton, J.N., Liachko, I., Dunham, M.J. and Shendure, J. (2014) Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3: Genes Genom. Genet. 4, 1339–1346.
[67] Lederberg, J., Cavalli, L.L. and Lederberg, E.M. (1952) Sex compatibility in *Escherichia coli*. Genetics 37, 720–730.
[68] Steinbock, L.J. and Radenovic, A. (2015) The emergence of nanopores in next-generation sequencing. Nanotechnology 26, 074003.
[69] Mitchell, M.A. and Dervan, P.B. (1982) Interhelical DNA-DNA crosslinking. Bis(monoazidomethidium)octaoxahexacosanediamine: a probe of packaged nucleic acid. J. Am. Chem. Soc. 104, 4265–4266.
[70] Mitchell, M.A. (1983). Interhelical DNA–DNA crosslinking of bacteriophage lambda: bis(monoazidomethidium)octaoxahexacosanediamine and bis(psoralen)nonaethyleneoxy ether, probes of packaged nucleic acid (Ph.D. dissertation). California Institut of Technology.
[71] Marie-Nelly, H., Marbouty, M., Cournac, A., Liti, G., Fischer, G., Zimmer, C. and Koszul, R. (2014) Filling annotation gaps in yeast genomes using genome-wide contact maps. Bioinformatics 30, 2105–2113.
[72] Hanson, S.J., Byrne, K.P. and Wolfe, K.H. (2014) Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus *Saccharomyces cerevisiae* system. Proc. Natl. Acad. Sci. USA 111, E4851–E4858.