Metagenome Analysis Exploiting High-Throughput Chromosome Conformation Capture (3C) Data.

Martial Marbouty, Romain Koszul

Metagenome analysis exploiting high throughput chromosome conformation capture (3C) data

Martial Marbouty[1,2] and Romain Koszul[1,2,†]

[1]Institut Pasteur, Department of Genomes and Genetics, Groupe Régulation Spatiale des Génomes, 75015 Paris, France

[2]CNRS, UMR 3525, 75015 Paris, France

†Corresponding author. E-mail: romain.koszul@pasteur.fr (R.K.)

Keywords : Metagenome assembly;meta3C metaHi-C; Hi-C;3C;Metagenomics;contact genomics

Abstract (100-120 words)

Microbial communities are extremely complex and constitute critical actors of our environment. Genomic analyses of these populations are a dynamic research area but remain limited by the difficulty to assemble full genomes of individual species. Recently, a new method for metagenome assembly/analysis based on chromosome conformation capture has emerged (meta3C). This approach quantifies the collisions experienced by DNA molecules to identify those sharing the same cellular compartments, allowing the characterization of genomes present within complex mixes of species. The exploitation of these chromosome 3D signatures holds important premises regarding reaching at the genome sequences from discrete species in complex populations. It also has the potential to correctly assign extra-chromosomal elements, such as plasmids, mobile elements and phages, to their host cells.

Limitations of metagenomic analysis

Microbial communities and their biochemical activities are ubiquitous and essential components of environmental and natural ecosystems. They often play important roles in the maintenance of environmental ecosystems as well as in sustaining animal and plant life [1–3]. Microbes that associate with a macroscopic host organism can also interact with it and exert a major influence on its metabolism with consequences on health, growth and fitness [4,5]. However, and despite their established importance, these mixes of microorganisms have often proven difficult to characterize to the full extent of their complexity, which often encompasses hundreds or more species engaged in co-dependant relationships. Notably, the vast majority (90-99%; reviewed in [6]) of microbes cannot be cultivated and therefore isolated, impairing more refined characterization of these species. This bottleneck can be bypassed, to some extent, by studying directly DNA extracted from the microbial community. Metagenomics (see Glossary) aims at studying microbial ecosystems globally through the characterization of the genes that are present in the community as a whole. Pioneering metagenomics projects have focused on the ubiquitous 16S RNA gene to explore the diversity and nature of the species present within populations. They brought to light the tremendous diversity of microbes, unveiling a largely unexplored and unknown world, and paving the way to this flourishing field [7–9]. Thanks to important breakthrough in sequencing technologies and computational analysis, numerous studies based on shotgun sequencing of metagenomes have then allowed exploring the DNA molecules present in a microbial community [10–12]. However, those analysis are somehow limited as it remains difficult to assemble these short DNA sequences into larger molecules (i.e. contigs and scaffolds; see Glossary; [13]), if not full genomes, hindering in-depth analysis of the system as a whole. Improving the ability to discriminate, characterize and assign DNA sequences to a specific species, so that eventually its complete genome can be characterize, is challenging for complex communities. Furthermore, a complete picture of the genetic content of a microbial community goes much beyond the sum of the genomes/chromosome(s) of the microorganisms that compose it. Indeed the population also contains numerous DNA molecules, either independent from

the core genome or exhibiting dynamic behaviours (plasmids, transposon, viruses/prophages) that add extra and important layers of complexity to the system.

The analysis of the individual chromosomes contained within a metagenomic sample has improved thanks to continuous advances in sequencing technologies: single-molecule, real-time (SMRT) sequencing [14] as marketed by Pacific Biosciences [15] or Nanopore [16] are promising approaches likely to become prominent in the field in the future. Alternatively or in complement, single-cell technologies offer interesting perspectives as they preserve the cellular integrity [17], though they remain prone to important biases, contamination and overall relatively out of reach of many laboratories. Computational methods can also overcome to some extent some of those limitations. *De novo* assembly programs such as MetaVelvet [18] and IDBA-UD [19] have been adapted to metagenomic data [20]. Finally, assembly of short reads into discrete genomes have also been improved by exploiting biases in base composition [21], by leveraging variations in gene abundance [22,23] between the genomes of different species, or by combining multiple parameters and searching for correlations for instance between variations in reads coverage and cultivation into different growth media [24]. However, the assumption that contigs/reads with similar characteristics are likely to originate from the same genome is contradicted by horizontal gene transfer events that generate genetic heterogeneity. Also, if multiple strains of a species coexist within the mix it is difficult or impossible to identify their different haplotypes. Alternative approaches to bypass these limitations are therefore actively sought.

Exploiting contact genomic to analyse metagenomes

A recent and potential breakthrough came from the realization that chromosome 3D physical signatures can be exploited to infer 1) the linear sequence of a chromosome and 2) whether different chromosomes occupies different cellular compartments or not [25]. This new field, dubbed ‘contact

genomics' [25], aims at exploiting the physical contacts experienced by small DNA molecules and quantified through genomic derivatives of the chromosome conformation capture technique (3C; see Glossary and Box 1**;** [26,27]). 3C relies primarily on a fixation step, typically achieved using formaldehyde cross-linking, to freeze and capture contacts made by DNA regions within (*cis*) and between (*trans*) chromosomes. The cross-linked DNA is then digested by a restriction enzyme, diluted, religated, and pair-end sequenced (Figure 1A and Box 1**).** Chimeric molecules resulting from religation events between non-adjacent DNA regions are identified from the pair-end reads by mapping them along chromosomes to generate a genome-wide matrix of contact frequencies. Performed on a population of cells, these contact maps presumably reflect the average genome organization. An immediate observation made from genomic 3C data is that contact frequencies between DNA regions decrease as a function of the distance separating them following, to some extent, a power law distribution [28]. Given a frequency of contacts, one should therefore be able to get a robust assessment of the distance that separates two loci. This makes it possible to exploit the broad distribution of statistically significant *cis*-contact, ranging from a few kb to up to a Mb, to bridge the numerous assembly gaps present in most published genomes (reviewed in [25]). Several approaches have recently tackled this challenge and developed programs aiming at improving genome assembly through the use of large 3C contact datasets [29–31]. These promising approaches, though perfectible, have allowed the scaffolding of contigs (see Glossary) up to several hundreds of Mb from incomplete assembly of genomes of various sizes (for instance of human contigs obtained *de novo*, [30,31]).

Beyond the great promises held by contact genomics approaches to chromosome scaffolding/genome assembly, this concept has now started to be applied to the field of metagenomics. By analogy with the physical contacts allowing discriminating multiple chromosomes within a single nuclear/cell compartment of a homogeneous population, metagenomic 3C (meta3C or metaHi-C) considers the entire population as a large ensemble of cellular compartments with specific chromosomal sets, and exploits their contact frequencies to discriminate DNA molecules

occupying different compartments [32–34] (Figure 1). As, physical contacts provide quantitative and objective information regarding whether or not two pieces of DNA share the same cellular compartment, without requiring any prior knowledge of the content of the sample.

To test the validity of this approach, a first important unknown was to characterize the "background noise" frequency with which two DNA RFs originating from 2 different genomes end up fused together after a meta3C experiment. To do so, pair-end reads obtained from meta3C experiments performed on mixes of species with known genomes were aligned against those sequences, and inter-species ligation events were quantified. Religation events between RFs coming from different genomes proved extremely low for both prokaryotic and eukaryotic mixes (< 1%; [32–34] (Figure 2A-B). One must notice that experimental conditions, notably fixation, have an important impact on the end results, with low amounts of crosslinking agent leading to noisier signal, especially in bacteria [33,34]. Overall, these control experiments confirmed that DNA/DNA contacts are a convenient and sound measurement of the co-existence of two DNA regions within the same compartment, if not the same genome (below).

In the case of environmental/natural samples, prior knowledge regarding the species present in the mix remains generally sparse and incomplete. Obtaining the genomes of the species present within the sample remains important challenge that is rarely met except for dominant members of the community [35]. To apply meta3C to such natural sample, the first step consists in assembling the metagenome from the sequencing reads in order to reduce the complexity of the data, which typically generates a large number of small contigs of various depths coverage. The paired-end reads containing the 3D contact information (Hi-C or 3C-seq, see Glossary) are then mapped on this set of contigs, producing a large meta3C network in with the nodes representing the contigs and the edges the contacts. This network is typically complex, and exhibits a community structure with subgraphs groups of densely connected nodes and sparse connections to the rest of the network. The challenges associated with complex network analysis are familiar to many scientific fields such as

sociology (e.g. friendship/collaborations network) [36,37] and neurobiology [38], and important efforts have been made to develop efficient clustering algorithms in recent years [39–43].

To test this approach, several group performed blind analyses of meta3C networks obtained from various controlled mixes (from 3 to 12 species) [32–34]. Metagenome assembly was performed using either SOAP de novo [32] or IDBA-UD [33,34] (Table 1), either from an independent library [32,33] or directly from the meta3C reads [34]. The later solution is advantageous when only a limited amount of biological material is available, especially if the meta3C data was generated using 3C-seq rather than Hi-C. Indeed, in the 3C-seq protocol the 3C library is directly sequenced, without any enrichment step for chimeric ligation products as in Hi-C [28]. Consequently, the sequenced library is composed of 80-90% regular paired-end reads that are mainly used during the assembly step, whereas the 10-20% chimeric read pairs corresponding to religation events do not result in a significant increase of assembly errors [34]. The interaction network between the *de novo* generated contigs is then achieved taking advantage of the 10-20% chimeric read pairs that correspond to religation events.

Different clustering algorithms can be used to identify communities of contigs within the global network of interactions generated using the meta3C PE reads (see Table 1). In the studies reviewed here, 80% to 98% of the entire genome sequences present in the controlled mixes were appropriately pooled together after processing. Interestingly, the clustering also delineated non-chromosomal DNA sequences such as plasmids. Notably, the distribution of contacts made by a F' plasmid sequence confirmed its presence in the vicinity of the *E. coli* chromosome, whereas a correlation analysis of the these contacts signal further revealed that it carried a large segmental duplication of this genome [34]. The clustering also pooled together chromosomes that shared the same cellular compartment, and identified multiple chromosomes within a single species.

Current limitations of meta3C approaches

Although promising, existing meta3C approaches remain perfectible. First, the presence of closely related species and/or of strains of the same species in the mix may impair the proper clustering of the corresponding sequences [32,33]. Notably, errors occurring during assembly step (resulting for instance from repeated sequences) can lead to chimeric contigs and misleading contact signal. Cutting the contigs into pieces of DNA of similar sizes can alleviate this problem [34], but further computational developments will be needed to fully exploit the information contained within meta3C libraries. In addition, and as mentioned above, metagenomes are more than a mix of genomic sequences. Homologous sequences, repeats, plasmids, mobile elements or horizontal gene transfer events are present, resulting in DNA sequences interacting with different host genomes at the same time. Those elements will bend the network by being equally connected to multiple communities as evidenced by an experiment performed on an completely unknown metagenome stemming from a LB-enriched sediment sample [34] (Figure 2C-D)**.** Detection of overlapping communities is a complex problem in the field [44–46], that will require further developments to be fully addressed in the case of meta3C data. Finally, another limitation is inherent to the 3C technology and results from the enzymatic restriction step (Box 1). The frequency of restriction sites is highly correlated with the GC content of the sequence, which can introduce important biases [47,48]. Such bias will notable affect meta3C experiments when members of a metagenomic sample exhibit large differences in GC content. One possibility to alleviate this problem is to build meta3C libraries with different restriction enzymes, so that all genomes are visible in at least one condition. Alternatively, other restriction approaches, such as partial DNAse digestion [49], could be envisioned.

Future directions

3C/Hi-C based approaches appear as promising tools to study the genomes of microbial communities since they give access not only to the chromosomes of the species present within the mix but also provide information regarding mobile elements such as plasmids and phages that share

the same compartment [32–34]. Such data are promising to the study of bacterial genomes dynamics and plasticity within population and over time. Microbial communities are not only composed of numerous different species of bacteria but also of a tremendous amount of phages, sometimes considered to be the most abundant “replicating” organisms on earth [50] and important vectors of DNA fragments [51,52]. Although their influence on the balance of bacterial communities is gradually being unveiled, not much is known about it because of the difficulty to access their sequences [53,54]. Consequently, our understanding of their roles in the size, structures and function of microbial communities is only starting. Finding out their bacterial hosts and their precise role in the plasticity of bacterial genomes will be necessary to fully understand phages and their integration in microbial communities. By analysing contact between phages and bacterial genomes, Meta3C appears as a promising method to answer to these important questions.

Although it is tempting to envision reassembling *de novo* the individual genomes identify within these complex communities using programs such as GRAAL [31,34], this remains challenging with such complex samples. However, the stakes of this computational challenge are high, considering the importance of these data to understand and characterize the physiology of non-cultivable organisms and the population equilibrium. Developments in both assembly programs and sequencing technologies will therefore likely allow full metagenome reconstruction in the near future.

Interestingly, the meta3C approach also gives access to the genome organization of the microorganisms directly captured in their environment. An emerging picture in the field is that the high-order chromosome organization in these species reflects, to some extent, metabolic regulation and/or cell cycle progression. In *Bacillus subtilis* for instance, cycles of condensation/decompaction of the origin of replication correlate with replication regulation [55]. In *Saccharomyces cerevisiae*, long-term survival following diauxic shift and entry into quiescence necessitate the settling of a “hyper-cluster” of telomeres [56]. Therefore, knowing the genome organization of the

microorganisms within the mix will provide access to their metabolic state, paving the way for an integrated analysis of the dynamic physiology of the population.

Concluding remarks

Metagenomic analyses provide important insights on the dynamics of natural microbial communities and their response or interplay with the environment. Approaches based on contact genomics, such as meta3C, have the potential to alleviate some of the limitations of traditional metagenomic approaches. DNA is an ubiquitous and stable molecule, and using it as a marker of "compartmentalization" at cellular and population levels hold great promises. Indeed, the objective and quantitative information provided by the collisions of DNA molecule sheds light not only on the chromosomal content of various cell compartments but also on the extra-genomic sequences present within the population. This paves the way for an exhaustive description of complex communities and gene dynamics taking into account the entire pool of DNA molecules present within a community, which appears within reach using meta3C (see Outstanding Questions). Taken together, contact genomic investigations using meta3C represent a great opportunity to address the current challenges metagenomics, such as deciphering how environmental changes influence the genome organization, dynamics and plasticity of the organisms present in a given ecological niche, and how they adapt in response.

Acknowledgement

The authors thank Axel Cournac, Lyam Baudry and Jean-François Flot for comments on this manuscript and discussions about this work. 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. M.M. is the recipient of an Association pour la Recherche sur le Cancer fellowship (number 20100600373).

Bibliography

1 Cryan, J.F. and Dinan, T.G. (2012) Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* 13, 701–712
2 Hanski, I. *et al.* (2012) Environmental biodiversity, human microbiota, and allergy are interrelated. *Proc. Natl. Acad. Sci.* 109, 8334–8339
3 Kau, A.L. *et al.* (2011) Human nutrition, the gut microbiome and the immune system. *Nature* 474, 327–336
4 Joice, R. *et al.* (2014) Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab.* 20, 731–741
5 Manor, O. *et al.* (2014) Mapping the inner workings of the microbiome: genomic- and metagenomic-based study of metabolism and metabolic interactions in the human microbiome. *Cell Metab.* 20, 742–752
6 Sharpton, T.J. (2014) An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* 5,
7 Rappé, M.S. and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.* 57, 369–394
8 Franzosa, E.A. *et al.* (2015) Sequencing and beyond: integrating molecular “omics” for microbial community profiling. *Nat. Rev. Microbiol.* 13, 360–372
9 Rappé, M.S. *et al.* (1997) Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hatteras, North Carolina. *Limnol. Oceanogr.* 42, 811–826
10 Hasan, N.A. *et al.* (2014) Microbial Community Profiling of Human Saliva Using Shotgun Metagenomic Sequencing. *PLoS ONE* 9, e97699
11 Mangrola, A. *et al.* Deciphering the microbiota of Tuwa hot spring, India using shotgun metagenomic sequencing approach. *Genomics Data* DOI: 10.1016/j.gdata.2015.04.014
12 Venter, J.C. *et al.* (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* 304, 66–74
13 Simpson, J.T. and Pop, M. (2015) The Theory and Practice of Genome Sequence Assembly. *Annu. Rev. Genomics Hum. Genet.* 16, null
14 Eid, J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138
15 Fichot, E.B. and Norman, R.S. (2013) Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome* 1, 10
16 Steinbock, L.J. and Radenovic, A. (2015) The emergence of nanopores in next-generation sequencing. *Nanotechnology* 26, 074003
17 Rinke, C. *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437
18 Zerbino, D.R. and Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829
19 Peng, Y. *et al.* (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinforma. Oxf. Engl.* 28, 1420–1428
20 Afiahayati, null *et al.* (2015) MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* 22, 69–77
21 Saeed, I. *et al.* (2012) Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.* 40, e34
22 Carr, R. *et al.* (2013) Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution. *PLoS Comput Biol* 9, e1003292

23 Hug, L.A. *et al.* (2013) Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* 1, 22
24 Albertsen, M. *et al.* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538
25 Flot, J.-F. *et al.* (2015) Contact genomics: Scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. *FEBS Lett.* DOI: 10.1016/j.febslet.2015.04.034
26 Dekker, J. *et al.* (2002) Capturing Chromosome Conformation. *Science* 295, 1306–1311
27 Dekker, J. *et al.* (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14, 390–403
28 Lieberman-Aiden, E. *et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326, 289–293
29 Burton, J.N. *et al.* (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125
30 Kaplan, N. and Dekker, J. (2013) High-throughput genome scaffolding from in-vivo DNA interaction frequency. *Nat. Biotechnol.* 31,
31 Marie-Nelly, H. *et al.* (2014) High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* 5,
32 Beitel, C.W. *et al.* (2014) Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* 2, e415
33 Burton, J.N. *et al.* (2014) Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 Bethesda Md* 4, 1339–1346
34 Marbouty, M. *et al.* (2014) Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife* 3, e03318
35 Pelletier, E. *et al.* (2008) “Candidatus Cloacamonas acidaminovorans”: genome sequence reconstruction provides a first glimpse of a new bacterial division. *J. Bacteriol.* 190, 2572–2579
36 Weng, L. *et al.* (2013) Virality Prediction and Community Structure in Social Networks. *Sci. Rep.* 3,
37 Kim, Y.-H. *et al.* (2013) Two Applications of Clustering Techniques to Twitter: Community Detection and Issue Extraction. *Discrete Dyn. Nat. Soc.* 2013, e903765
38 Bullmore, E. and Sporns, O. (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198
39 Blondel, V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008, P10008
40 Campigotto, R. *et al.* (2014) A Generalized and Adaptive Method for Community Detection. *ArXiv14062518 Phys. Stat* at <http://arxiv.org/abs/1406.2518>
41 van Dongen, S. and Abreu-Goodger, C. (2012) Using MCL to extract clusters from networks. *Methods Mol. Biol. Clifton NJ* 804, 281–295
42 Lancichinetti, A. and Fortunato, S. (2009) Community detection algorithms: a comparative analysis. *Phys. Rev. E* 80,
43 Liu, W. *et al.* (2014) Detecting Communities Based on Network Topology. *Sci. Rep.* 4,
44 Chen, Y. *et al.* (2014) Overlapping community detection in networks with positive and negative links. *J. Stat. Mech. Theory Exp.* 2014, P03021
45 Palla, G. *et al.* (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818
46 Wang, W. *et al.* (2013) Fuzzy overlapping community detection based on local random walk and multidimensional scaling. *Phys. Stat. Mech. Its Appl.* 392, 6578–6586
47 Cournac, A. *et al.* (2012) Normalization of a chromosomal contact map. *BMC Genomics* 13, 436
48 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
49 Ma, W. *et al.* (2015) Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods* 12, 71–78
50 Edwards, R.A. and Rohwer, F. (2005) Viral metagenomics. *Nat. Rev. Microbiol.* 3, 504–510

51 Muniesa, M. *et al.* (2013) Potential impact of environmental bacteriophages in spreading antibiotic resistance genes. *Future Microbiol.* 8, 739–751
52 Stern, A. *et al.* (2012) CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* 22, 1985–1994
53 Ogilvie, L.A. *et al.* (2013) Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences. *Nat. Commun.* 4, 2420
54 Thurber, R.V. *et al.* (2009) Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* 4, 470–483
55 Marbouty, M. *et al.* (2015) Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol. Cell* 59, 588–602
56 Guidi, M. *et al.* (2015) Spatial reorganization of telomeres in long-lived quiescent cells. *Genome Biol.* 16, 1
57 Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46
58 Dixon, J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380
59 Le, T.B.K. *et al.* (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* 342, 731–734
60 Bennett, S. (2004) Solexa Ltd. *Pharmacogenomics* 5, 433–438
61 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380
62 Rothberg, J.M. *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475, 348–352
63 Bastian, M. *et al.* (2009) , Gephi: An Open Source Software for Exploring and Manipulating Networks. , in *Third International AAAI Conference on Weblogs and Social Media*
64 Jacomy, M. *et al.* (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS One* 9, e98679
65 Jarvis, R.A. and Patrick, E.A. (1973) Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans Comput* 22, 1025–1034
66 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 95, 14863–14868

Box 1. 3C approaches unveil the multiscale organization of genomes

Chromosomes from prokaryotes and eukaryotes are long filaments that are tightly folded within a relatively confined space. Although the volume occupied by the DNA fibre is only a fraction of the total nucleus or nucleoid volume, its folding is of primary importance since it determines its accessibility and the cell's ability to mobilise the information carried by the DNA in a timely manner. Hence, complex mechanisms have evolved that drive and regulate this organization by folding the 1D genome into a 3D structure. The principles, role and dynamics of this folding have been investigated since the birth of cytology, with remarkable advances obtained from microscopy studies. However, an important step was recently reached by combining high-throughput sequencing with the chromosome conformation capture (3C) technique [26], giving access to the multiscale 3D organization of the DNA sequence. 3C and its genomic derivatives [27] have in recent years become popular tools to study spatial chromosome organization. In a 3C assays, a population of cells or a single cell is fixed using a crosslinking agent such as formaldehyde. This results in covalent bounds between histone proteins, which trap together DNA regions occurring in the vicinity of each other's. The cross-linked DNA is digested with a restriction enzyme then religated under dilute conditions (to favour intramolecular ligation over intermolecular ones). After DNA purification, the resulting 3C library consists in a collection of restriction fragments (RFs) ligated together, resulting in molecules that are chimeric with respect to the reference genome. The relative abundance of the pairs of RFs involved in a chimeric religation event reflects the frequency with which these two chromatin segments were crosslinked, hence their spatial proximity. The relative proportions of these chimeric product were initially quantified using semi-quantitative PCR [26] but 3C assays have recently boomed, mainly thanks to the advent of high-throughput sequencing techniques [57] allowing sequencing the extremities of the DNA molecules present in the 3C library and estimating in a quantitative way contact frequencies of pairs of RFs within genomes. Genome-wide contact maps are built from this quantification step. 3C-based assays have provided important advances in our understanding of the genomic architecture of mammals [28,58] or bacteria [59]. In all species,

chromosomes appear as well individualized entities sometimes presenting large topological domains (TADs – Topological Associated Domains). Their precise roles and regulatory mechanisms are the subject of intensive investigations, with increasing evidence that the 3D architecture of chromosomes is an essential part of the cellular process and cell physiology. The next decade promises to be really exciting in this research field as improvements in sequencing technology and data treatment will certainly uncover the relationships between genome architectures and cellular processes.

Glossary

3C (Chromosome Conformation Capture): see **Box 1**.

3C-seq: next-generation sequencing of a 3C library without enriching it for chimeric read pairs (unlike Hi-C).

Contig: a consensus DNA sequence inferred from a set of overlapping DNA fragments. In a contig the order of the base pairs is known with a high confidence, without gaps.

Hi-C: next-generation sequencing of a 3C library that was enriched in chimeric read pairs using biotinylation.

Metagenomics: scientific field that aims at studying genetic material directly obtained from environmental samples. It could also be referred as ecogenomics or environmental genomics.

NGS (next-generation sequencing): generic term applying to a group of sequencing technologies developed at the turn of the century (Solexa [60], 454 [61], Ion Torrent [62]) that allow affordable, massive and fast sequencing of DNA. Typically, NGS generate tens of millions of short sequences (or read), and is commonly applied to genomics and sequencing projects. Also called high-throughput sequencing.

Scaffold: a DNA sequence made out of multiple contigs with intervening gaps (the length of which is generally not precisely known).

Figures & tables

Table1: Summary of published processes and data using 3C-based metagenome assembly.

Figure1: Principle of a meta3C experiment.

(A) Starting from a mix of species (metagenomicssample), a shotgun library is generated and used to generate a preliminary assembly (a 3C library can also be used here).

(B) Starting from the same sample, a 3C/HiC library is generated (see **Box1**).

(C) Pair-end reads from the 3C library, some of which reflect the collision frequencies between all the pairs of DNA restriction fragments present within the population, is then mapped on the contigs.

(D) Representation of the complex network resulting from the step C). Left panel: contact map representation of the contigs. Right panel: 3D representation of the data using a clustering visualization tool such as Gephi [63].

(E) The disordered contigs are then clustered and reordered based on their frequencies of interactions, unveiling the genomic sequences of the organisms present within the mix.

Figure 2 : meta3C experiment on different mix of species [34].

(A) Chromosomal contact map of a mixture of three bacteria (*Bacillus subtilis, Escherichia coli, Vibrio cholerae*), with the color code representing contacts between DNA regions from low (white) to high (red) frequencies (a.u.). Frequencies of inter-species (chimeric) pairs of reads are directly reported on the matrix.

(B) Contact frequencies plotted as a function of genomic distance (for all 3 bacterial genomes together). The score shows a clear decrease at the genome size of these bacteria (i.e. 4Mb).

(C) Meta3C contact map of the largest 11 communities of contigs found by analyzing a river sediment sample. Each square in the matrix corresponds to a community grouping contigs that exhibit significantly more contact with each other than with other communities. Red squares indicate signal outside the main diagonal due to contigs exhibiting important contacts between several communities (i.e. overlapping communities).

(D) Illustration of the interactions between the 11 largest communities of contigs using the force-directed graph-drawing algorithm Force Atlas 2 [64]. Each node corresponds to a contig and each link represents at least one meta3C interaction. The colors correspond to the communities identified by the Louvain algorithm and described in (A). The red square highlights the overlapping communities described in (A).

Table1

Ref.	Samples	Assembly				Network resolution		Results
		Library used		Strategy		Library used	Algorithm	
[32]	synthetic community of 5 bacteria (*P. pentosaceus, L. brevis* + 2 plasmids, *E. coli BL21, E. coli K12, B. thailandensis*	**Type**	reads simulated	**Programm**	SOAP de novo	**Type** HiC	Markov Clustering Algorithm (MCL) [41]	4 clusters of contigs detected (100% of contigs used after filtering out small contigs).
		PE	12 M	**Contigs**	7687	**PE** 20 M		
		length	2x165 pb	**N50**	87 Kb	**length** 2x160pb		
				Assembly size	15.57Mb	**RE** HindIII		
[33]	synthetic commmunity of 16 yeasts strains (four strains of *S. cerevisiae*, 12 species of ascomycetes with reference genome available)	**Type**	shotgun + MatePair	**program**	IDBA-UD	**Type** HiC	Jarvis-Patrick nearest-neighbor clustering algorithm [65] followed by hierarchical agglomerative clustering [66]	12 distinct clusters (covering 82.2% of the total assembly - contigs that do not contain HindIII sites)
		Shotgun	92 M	**contigs**	48511	**PE** 81 M		
		Mate Pair	9.2 M	**N50**	17 Kb	**length** 2x100pb		
		length	2x100pb	**Assembly size**	136 Mb	**RE** HindIII		
	synthetic community of 8 yeast, 9 bacteria (1 bacteria harbor 2 chromosomes and 1 plasmid), 1 archeon	**Type**	assembly simulated	**program**	simulation of contigs by cutting references genomes into contigs of 10Kb	**Type** HiC		18 distinct clusters (covering 99.6% of the total simulated assembly)
		PE	xxxx	**contigs**		**PE** 14 M		
		length	xxxx	**N50**		**length** 2x100pb		
				Assembly size		**RE** HindIII		

	synthetic community of 3 bacteria (*E. coli* + 1 plasmid, *B. subtilis, V. cholerae*)	**Type**	3C seq	**program**	IDBA-UD	**Type**	3C seq		3 communities of contigs detected (covering 95% of the total assembly)
		PE	4 M	**contigs**	2436	**PE**	4 M		
		length	2x91pb	**N50**	55 Kb	**length**	2x91pb		
				Assembly size	12.5 Mb	**RE**	HpaII		
[34]	synthetic community of 11 yeasts species	**Type**	3C seq	**program**	IDBA-UD	**Type**	3C seq	Louvain algorithm [39]	13 communities of contigs (11 major representing 98% of the assembly, one contaminant, one corresponding to misassembled contigs)
		PE	60 M	**contigs**	47 614	**PE**	60 M		
		length	2x91pb	**N50**	6.9 Kb	**length**	2x91pb		
				Assembly size	138 Mb	**RE**	DpnII		
	LB-enriched river sediment	**Type**	3C seq	**program**	IDBA-UD	**Type**	3C seq		184 significant communities (19 larger than 1Mb)
		PE	67 M	**contigs**	130 713	**PE**	67 M		
		length	2x91pb	**N50**	1.2 Kb	**length**	2x91pb		
				Assembly size	111 Mb	**RE**	HaeIII		

Figure 1

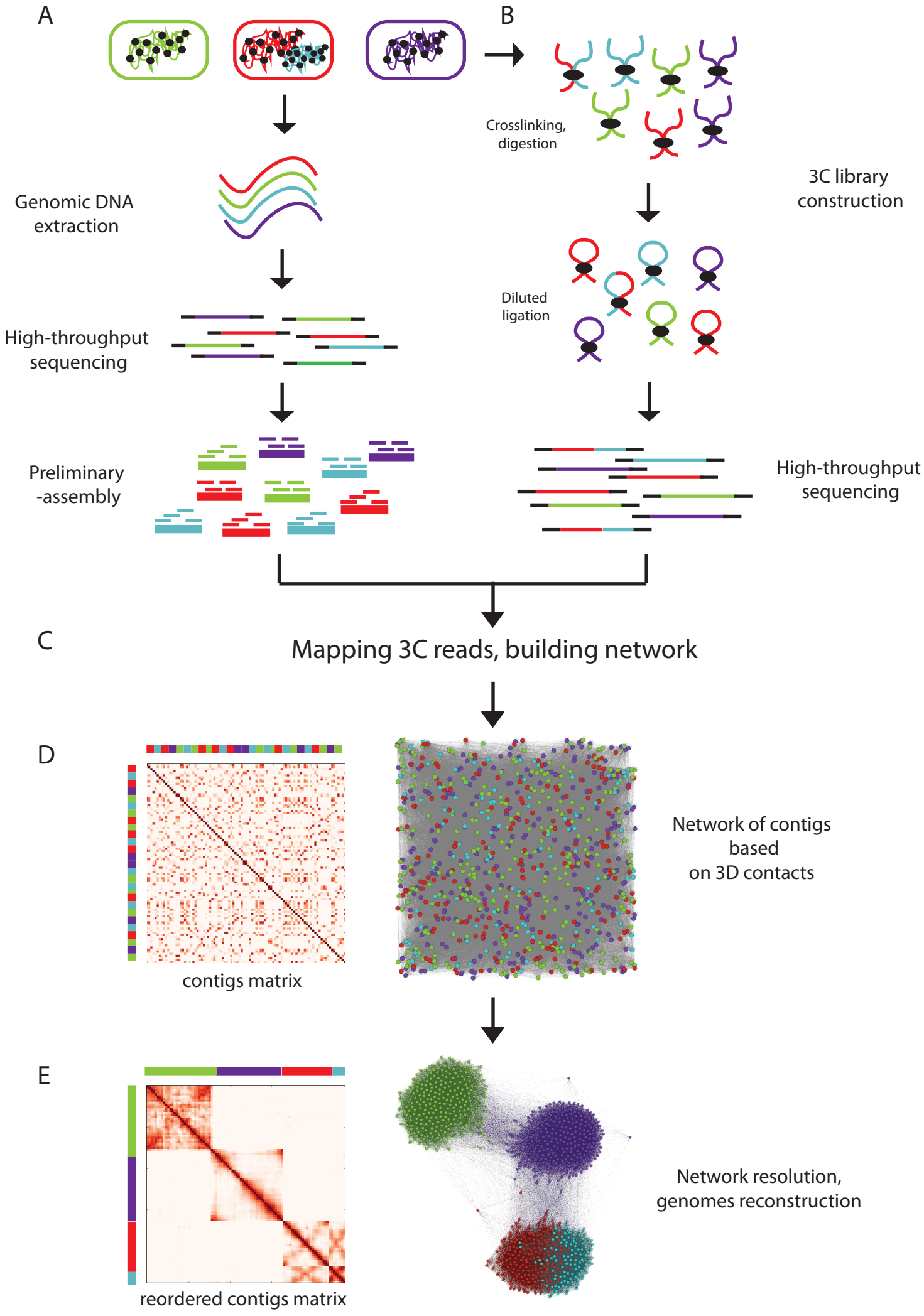
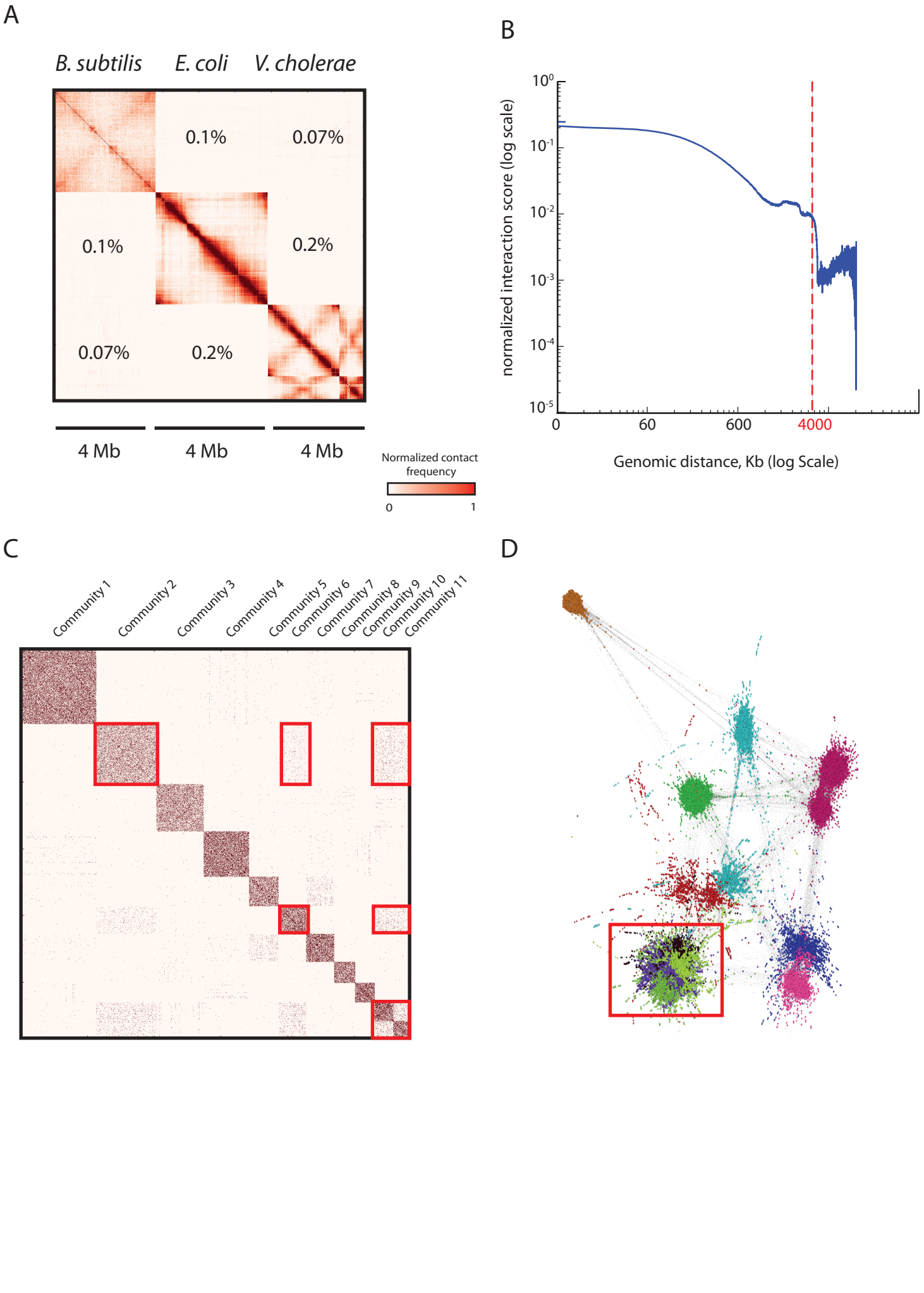

A
B
Crosslinking,
digestion
3C library
construction
Genomic DNA
extraction
Diluted
ligation
High-throughput
sequencing
Preliminary
-assembly
High-throughput
sequencing
C
Mapping 3C reads, building network
D
contigs matrix
Network of contigs
based
on 3D contacts
E
reordered contigs matrix
Network resolution,
genomes reconstruction

Figure 2


A
B. subtilis
E. coli
V. cholerae
0.1%
0.07%
0.1%
0.2%
0.07%
0.2%
4 Mb
4 Mb
4 Mb
Normalized contact frequency
0
1
B
normalized interaction score (log scale)
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
60
600
4000
Genomic distance, Kb (log Scale)
C
Community 1
Community 2
Community 3
Community 4
Community 5
Community 6
Community 7
Community 8
Community 9
Community 10
Community 11
D