Chapter 13

Generation and Analysis of Chromosomal Contact Maps of Yeast Species

Axel Cournac, Martial Marbouty, Julien Mozziconacci, and Romain Koszul

Abstract

Genome-wide derivatives of the chromosome conformation capture (3C) technique are now well-established approaches to study the multiscale average organization of chromosomes from bacteria to mammals. However, the experimental parameters of the protocol have to be optimized for different species, and the downstream experimental products (i.e., pair-end sequences) are influenced by these parameters. Here, we describe a complete pipeline to generate 3C-seq libraries and compute chromosomal contact maps of yeast species.

Key words Yeast, Chromosome conformation capture, 3C, Genome organization, Genome assembly, 3C analysis

1 Introduction

We present a method to characterize the tridimensional (3D) organization of budding yeast genomes (*see* [1], for the first genome-wide analysis performed in *S. cerevisiae*). Using 3C-seq, a derivative of chromosome conformation capture (3C; [2–4], this protocol generates genome-wide contact maps of various yeast species. An interest of the 3C-seq approach, compare to other 3C derivatives such as Hi-C [5, 6], is that it does not use any enrichment for ligation products and can be directly applied to the sequencing and assembly of unknown species, as described in [7]. Briefly, 3C gives access to the contact frequencies between restriction fragments (RFs) along a chromosome, reflecting the average chromosome organization within the nuclei within a population [2] and, eventually, unveiling functional reorganization upon changes in DNA-related metabolic processes such as DNA repair [8], homolog pairing during meiosis [2], or transcription [9]. Several methods have been published regarding the generation and analysis of 3C

Frédéric Devaux (ed.), *Yeast Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1361,
DOI 10.1007/978-1-4939-3079-1_13, © Springer Science+Business Media New York 2016

libraries, including a recent comprehensive discussion that recapitulates the overall experimental approach and analysis [10]. In its classical version, a cellular culture of a species of interest is treated with a cross-linking agent (typically formaldehyde) that generates covalent bounds between proteins and between DNA and proteins [2]. In each cell, cellular components, including the chromosomal set, will "freeze" in a disposition that is assumed to reflect the physiological configuration. To quantify the contacts between different DNA regions of the genome, two steps are necessary. First, the cells (and nuclei) are gently lysed and the cross-linked chromatin is digested with a carefully chosen restriction enzyme. The insoluble part of the raw chromatin extract is then isolated through centrifugation, diluted, and incubated in presence of DNA ligase. Using the insoluble fraction diminishes the background by removing small DNA molecules that were not cross-linked in large complexes [11, 12]. Ligating under diluted conditions aims in turn at alleviating the relegation of molecules which are trapped in the different cross-linked complexes. Following the ligation step, the cross-link is then reversed and the DNA purified. The resulting 3C library consists of a mix of different ligation products whose relative abundance reflects their average spatial proximity within the cell population at the time of the fixation step. The different religation events within a 3C library can be quantified using pair-end (PE) sequencing and genomic contacts maps generated through a variety of protocols [1–3, 5, 6, 13].

This section describes the experimental protocol for generating and sequencing a 3C library of a yeast species. The experimental part is then followed by a brief overview of the computational analysis necessary to extract meaningful contact information from the raw data sequencing. Generationand analysis of 3C libraries do not require special equipment (except obviously access to a sequencing apparatus able to process a large number of PE sequences). However, the preparation of the assay requires careful planning. The choice of the restriction enzyme and of the cross-linking conditions is critical for the success of the experiment, and must be thoughtfully envisioned before starting (*see* **Notes 1** and **2**).

2 Materials

2.1 3C Library Components

1. 50 mL disposable conical tubes.
2. Filtration unit 0.22 μm.
3. 1.5 and 2 mL lo-binding microcentrifuge tubes (Eppendorf, Hamburg, Germany).
4. VK05 Precellys tube (Bertin Corp, Rockville, Maryland, USA).
5. Yeast species of interest (genome sizes, ~10–15 Mb).

6. Restriction enzyme and corresponding restriction enzyme buffer (*see* **Note 2**).
7. 5 U/μL T4 DNA ligase (Weiss Units).
8. 20 mg/mL proteinase K in water.
9. 10 mg/mL DNAse-free RNAse A in water.
10. 37 % formaldehyde solution (v/v) (Sigma-Aldrich, Saint Louis, Missouri, USA).
11. 2.5 M glycine: weigh 75.07 g of glycine and transfer to a 1 L cylinder. Add water to a volume of 400 mL and dissolve glycine using a magnetic stirrer and a stir bar (*see* **Note 3**). Filtrate on a 0.22 μm filtering unit and store at room temperature (RT).
12. 10 % sodium dodecyl sulfate (w/v) (SDS) in water. Add 20 mL of 20 % SDS (*see* **Note 4**) in a 50 mL disposable conical tube. Add 20 mL of water. Mix gently by returning tube several times. Store at RT.
13. 20 % Triton X-100 (v/v) in water. Add 10 mL of Triton X-100 in a 50 mL falcon. Add 40 mL of water and incubate in a 37 °C water bath until complete dissolution (it can take several hours). Store at RT.
14. 10× ligation buffer (without ATP): 500 mM Tris HCl pH 7.4, 100 mM $MgCl_2$, 100 mM DTT. Add 100 mL of Tris–HCl pH 7.5, 20 mL of $MgCl_2$ 1 M, and 10 mL of DTT 2 M to a 500 mL cylinder. Add water to reach 200 mL, mix and filtrate on 0.22 μm filtering unit. Split as 10 mL aliquot and store at −20 °C.
15. 10 mg/mL bovine serum albumin (BSA) in water. Store as 1 mL aliquots at −20 °C.
16. 100 mM adenosine triphosphate (ATP) pH 7.0 in water. Weigh 1 g of ATP and transfer to a 50 mL falcon. Add 14 mL of water. Add 1.6 mL of NaOH 1 M. Complete to 16.7 mL with water. Check that the pH is around 7.0. Filtrate on 0.22 μm filtering unit. Store as 1 mL aliquots at −20 °C (*see* **Note 5**).
17. 500 mM EDTA in water, pH 8.0.
18. 3 M sodium acetate in water, pH 5.2. Weigh 204.12 g of sodium acetate and transfer to a 1 L cylinder. Complete with water to 400 mL, and adjust pH to 5.2 with acid acetic 100 %. Complete to 500 mL with water. Filtrate on a 0.22 μm filtering unit and store at RT.
19. Isopropanol.
20. 10:9:1 phenol–chloroform–isoamylalcohol pH 8.2.
21. 100 % Ethanol.
22. TE buffer, pH 8.0. Add 5 mL of TE 10× to a 50 mL falcon. Add 45 mL of water and filtrate on a 0.22 μm filtering unit. Store at RT.

23. Precellys (Precellys®24) (Bertin Corp, USA) (*see* **Note 6**).
24. 16 °C water bath.
25. 65 °C oven.
26. Magnetic stirrer and stir bar.
27. Variable temperature incubator (25, 30 and 37 °C).
28. Dry bath at 65 °C.
29. Refrigerated tabletop centrifuge (for 50 mL falcon tubes).

2.2 NGS Library Processing Components

1. Covaris S220 instrument (Covaris Ltd., Woburn, Massachusetts, USA).
2. Snap Cap microTUBE for Covaris (Covaris Ltd.).
3. Column PCR purification Kit (QIAgen, Venlo, Netherlands) (*see* **Note 7**).
4. Column MinElute PCR purification Kit (QIAgen) (*see* **Note 7**).
5. 1.5 mL lo-binding microcentrifuge tubes (Eppendorf).
6. Illumina paired-end adapters and amplification primers (*see* **Note 8**; Illumina, San Diego, California, USA).
7. Tabletop centrifuge.
8. NanoDrop (Thermo Fisher Scientific).
9. 10× ligation Buffer (New England Biolabs, Ipswich, Massachusetts, USA—NEB): 500 mM Tris–HCl (pH 8.0), 100 mM $MgCl_2$, 100 mM DTT, 10 mM ATP.
10. 10× NEBuffer 2 (NEB): 500 mM NaCl, 100 mM Tris–HCl (pH 7.9), 100 mM $MgCl_2$, 10 mM DTT.
11. 10 mM deoxynucleotide triphosphates in water. Add 100 μL of each dNTP (dNTP set 100 mM) in a microcentrifuge tube 1.5 mL. Complete to 1 mL with water. Make 50 μL aliquots and store them at −20 °C (*see* **Note 9**).
12. 1 mM deoxyadenosine triphosphate in water. Add 10 μL of dATP 100 mM (from the dNTP set) in a microcentrifuge tube 1.5 mL. Complete to 1 mL with water. Make 50 μL aliquots and store them −20 °C (*see* **Note 9**).
13. 10 U/μL T4 polynucleotide kinase.
14. 1 U/μL T4 DNA polymerase.
15. 5 U/μL Klenow DNA polymerase.
16. 5 U/μL Klenow (exo-) DNA polymerase.
17. 400 U/μL T4 DNA ligase (Cohesive End Unit).
18. Phusion polymerase (Thermo Fisher Scientific).

2.3 Data Processing

1. Computer with a UNIX system (Linux, MacOSX, Ubuntu). Large memory space and multiprocessor core are needed for efficient reads alignments.

2. Alignment program (for instance, Bowtie2).
   http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
3. A script language like Bash or python to manipulate files.
4. A tool to visualize big matrices like Matlab (license needed) or Octave (free).

3 Methods

3.1 Generation of a 3C Library of Mixed Species

The generation of the 3C library takes 3 days, and the generation of the sequencing library an additional 2–3. The 3C library can be stored at −80 °C and therefore the two processes can be separated. Whereas it remains difficult to prepare more than four libraries at a time, processing the samples for sequencing can be performed at a larger scale (up to eight libraries), the limiting step being then, to some extent, the purification of molecules of a size appropriate for sequencing (*see* **Note 10**). Therefore, timing is important criteria when planning to do the experiment! The overall schedule will require for an experienced experimentalist an afternoon (partly), a morning–afternoon (partly), a morning (full), followed by 2 full days (with several incubations steps).

3.1.1 Culture Fixation

1. Start culture of yeast species in your favorite medium. For instance, strains can be grown at 30 °C in 100 mL BMW medium [14] up to 1×10^7 cells/mL (this quantity will allow to realize two libraries) (*see* **Note 11**).
2. Add 8.5 mL of the fresh formaldehyde solution (i.e. 37 %) to the culture (final concentration of 3 %) (*see* **Note 2**).
3. The cells are incubated for 30 min at room temperature (RT) under gentle agitation with a magnetic stirrer.
4. Move the cell culture at 4 °C for another 30 min under gentle agitation.
5. Transfer the culture at RT and add 25 mL of Glycine 2.5 M (final concentration: 470 mM) to quench the remaining formaldehyde; incubate under agitation for 5 min at RT.
6. Relocate the culture at 4 °C and keep them under gentle agitation for an extra 15 min.
7. Pellet the fixed cells at 4 °C (3500 × *g*—10 min).
8. Wash the cells with 10 mL of the initial medium.
9. Pellet the fixed cells at 4 °C (3500 × *g*—10 min).
10. Suspend the cells into 2 mL of medium and transfer them into 2 × 1.5 mL microcentrifuge tubes.
11. Pellet the cells at 4 °C (3500 × *g*—10 min).

12. Remove the supernatant and flash freeze the pellet (i.e. in liquid nitrogen or dry-ice + ethanol).
13. Store pellets at −80 °C until use.

NB: Do not store pellet for more than 6 months (*see* **Note 12**).

3.1.2 3C Library Generation

Day one

1. Thaw the pellet on ice for 1 h.
2. Resuspend the cells in 4.5 mL of 1× restriction buffer (*see* **Notes 1** and **2**).
3. Transfer the cell suspension into 3× VK05 tubes (Precellys) (*see* **Note 6**).
4. Lyse the cell using the following program: 9 cycles × (6500 × *g*—30 s ON/60 s OFF).
5. Transfer lysate into 8 × 1.5 mL microcentrifuge tube (500 μL per tube).
6. Add 15 μL of 10 % SDS per tube (final concentration: 0.3 %).
7. Incubate tubes in a dry bath at 65 °C for 20 min.
8. Promptly transfer tubes on ice and incubate for 1 min.
9. Incubate tubes for 30 min at 37 °C and under agitation.
10. Add 50 μL of Triton X-100 20 % and 6 μL of 10× restriction buffer per tube.
11. Incubate tubes for 30 min at 37 °C and under agitation.
12. Put one tube aside as a non-digested control.
13. Add 150 units of restriction enzyme in each of the 7 remaining tubes.
14. Incubate overnight at the appropriate temperature for the chosen restriction enzyme.

Day two

15. The next morning, take the non-digested control and one of the digested samples (non-digested and digested controls, respectively). Add 100 μL of SDS 10 % and 30 μL of proteinase K to each tube and incubate them at 65 °C overnight (these controls will then be furthered processed at **step 30**).
16. Centrifuge the 6 remaining tubes at 16,000 × *g* for 20 min at temperature in order to isolate the insoluble fraction of the cross-linked chromatin [11].
17. Remove the supernatant and suspend each pellet in 500 μL of H_2O.
18. Pool the pellets three by three and dilute the two samples in 22.5 mL of a precooled (4 °C—on ice) ligation reaction mix (10× ligation buffer 2.4 mL, BSA 10 mg/mL 240 μL, ATP 100 mM 240 μL, water) in 50 mL conical tubes.

19. Add 125 units of T4 DNA ligase.
20. Homogenize the reaction by inverting the tubes 2–3 times.
21. Incubate for 4 h in a 16 °C water bath.
22. Transfer to a 25 °C water bath for an extra 45 min.
23. Add 200 μL of EDTA 500 mM per tube to stop the reaction.
24. Add 200 μL of proteinase K (20 mg/mL) and incubate the tube overnight at 65 °C.

Day three

25. The next morning, cool down the tubes at room temperature and transfer the solution to new 50 mL conical tubes.
26. Add 2.4 mL of 3 M Na Acetate pH 5.0 and 24 mL isopropanol and incubate at −80 °C for 1 h in order to precipitate DNA (*see* **Note 13**).
27. Centrifuge the tube in an appropriate centrifuge at 10,000 × *g* for 20 min.
28. Remove the supernatant and dry the pellet on the bench (*see* **Note 14**).
29. Suspend each pellet in 900 μL of TE buffer 1× and transfer them in 2 × 2.0 mL microtube.
30. Perform a DNA extraction for each tube using 900 μL of phenol–chloroform. Also extract the DNA from control samples from **step 15** using 500 μL of phenol–chloroform–isoamylalcohol.
31. Recover 2 × 400 μL of the aqueous phase (upper phase) for each tube (800 μL per tube in total) (and 1 × 400 μL for control tubes and transfer them into 1.5 mL microcentrifuge tube.
32. Add 40 μL of 3 M Na Acetate pH 5.0 and 1 mL of cold ethanol to each tube.
33. Vortex the tubes and incubate at −80 °C for 30 min.
34. Centrifuge the tubes at 16,000 × *g* for 20 min; discard the supernatants.
35. Wash each DNA pellet with 500 μL of cold 70 % ethanol.
36. Centrifuge tubes at 16,000 × *g* for 20 min and remove supernatant.
37. Dry pellet by incubating them on a 37 °C dry bath.
38. Suspend each pellet in 30 μL TE buffer 1× supplemented with RNAse A (0.1 μg/mL final concentration).
39. Incubate at 37 °C for 45 min.
40. Pool the tubes containing the 3C libraries.
41. Estimation of the quality and quantity on a 1 % agarose gel (Fig. 1; *see* **Note 15**).
42. *Optional: store at −80 °C as ~6 μg DNA aliquots .*

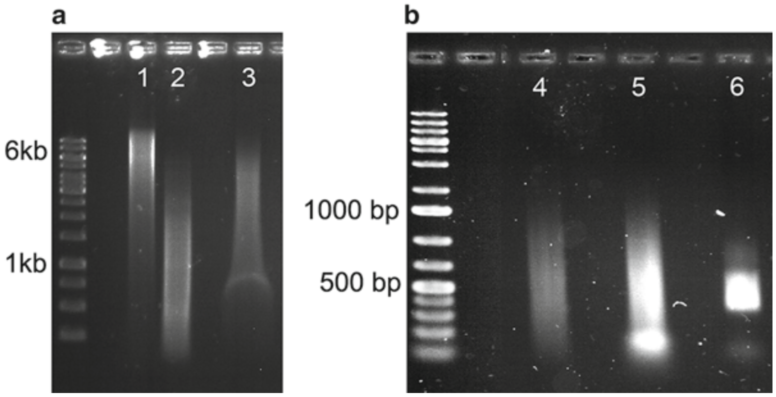

Fig. 1 Photography of gel electrophoresis migration of DNA at various steps of a 3C library construction. (**a**) Non digested control (*1*); Digested control (*2*); 3C library (*3*). (**b**) Processing of a 3C library for Illumina sequencing. Profile after shearing (*4*); profile following reparation, addition of 3′ A-tail and ligation of PE adapters (*5*); profile after size selection (400–800 bp) and PCR amplification (*6*)

3.1.3 Processing the 3C Library for Deep-Sequencing

The current protocol applies if the sequencing apparatus is anIllumi-nasequencer. For other brands/technologies, refer to the manual to design an appropriate protocol.

1. *Optional:* thaw a dry 3C sample on ice for 30 min.
2. Adjust the volume of a melted ~6 μg aliquot of a 3C library to 130 μL with water. With less DNA the protocol can nevertheless be pursued (down to 500 ng in our experience) but more amplification cycles will be necessary at the end.
3. Shear the library using your favorite instrument. For instance, we use a Covaris with the following settings: Peak Power: 105, Duty Factor 5 %, Cycles per Burst 200, Treatment time (s) 80 s, to obtain DNA fragments between 300 and 1500 bp (*see* **Note 16**).
4. Purify the DNA on a QIAquick column and elute with 5 μL of elution buffer (EB).
5. Quantify the DNA on a NanoDrop apparatus and prepare a tube with 5 μg of DNA and adjust the volume to 80 μL with water.
6. Add 12 μL of 10× ligase buffer (NEB), 4 μL of dNTP 10 mM, 15 μL of T4 DNA polymerase, 5 μL of T4 polynucleotide kinase, 1 μL of Klenow DNA polymerase) and complete to 120 μL with H_2O. Incubate at RT for 30 min.
7. Purify on QIAgen MinElute column and recover 30 μL of DNA in EB (to do so, add 31 μL of EB on the column in order to recover 30 μL).

8. Add 5 μL of 10× buffer NEB2, 10 μL of dATP 1 mM, 3 μL of Klenow DNA polymerase (exo minus) and complete to 50 μL with water. Incubate at 37 °C for 30 min followed by 20 min at 65 °C.
9. Purify on QIAgen MinElute column and recover 20 μL of DNA in EB.
10. Add 3 μL of 10× ligation buffer (NEB), 4 μL of adapters 10 μM, and 3 μL of T4 DNA ligase (NEB). Incubate at room temperature for 2 h (it is also possible to incubate overnight at 4 °C).
11. Purify fragment between 400 and 800 bp with your favorite method (Gel, Pippin Prep, caliper—*see* **Note 10**). Recover DNA in a volume of 40 μL in EB or TE.
12. Determine the optimal number of cycle and quantity of matrix to generate enough library for sequencing. Prepare several PCR reaction (phusion DNA polymerase—Volume of 50 μL per reaction) with different amount of library (Typically 1 and 2 μL). Temperature profile of the reaction is as follow: 30 s at 98 °C followed by 9, 12, or 15 cycles of 10 s at 98 °C, 30 s at 65 °C, 30 s at 72 °C, and a final 7 min extension at 72 °C.
13. Run the PCR reaction on a 1 % agarose gel and determine the optimal conditions.
14. Prepare 8 PCR reaction using the determine conditions and run them.
15. Purify on two QIAquick MinElute columns and recover around 40 μL of DNA.
16. Quantify on NanoDrop and check the profile on gel.
17. Run your library on an Illumina sequencing platform.

3.2 Analysis of Pair-End Sequencing Reads

Each 3C based protocol present peculiarities likely to generate noise or specific biases in the data. These caveats can be attenuated by an appropriate, specific preprocessing of the data and by proper normalization. Several approaches have been described that aim at correcting biases, or alleviating it to improve the quality of subsequent analysis [15–17]. This part presents the main steps to process the 3C-seq data described above. We provide commands and software's that we currently use. This description is only an illustration of what can be done, since many other bioinformatics tools exist and are available to the community.

3.2.1 Mapping Along the Genomes of the Mixed Species Yeast

1. If there is a reference genome for the species you are studying, recover or generate the fasta file (*see* **Note 17**). If the genome is unknown, *see* Marbouty et al. [7]. For the sake of illustration, we will use in the following the file "*genomes_yeasts.fa*" as the name of the reference genome.

2. Indexing the reference genome. Aligner software such as Bowtie 2 or BWA are needed for this task and routinely used [18]. The first step is to index your reference genome.

```
> bowtie2 -build genomes_yeasts.fa genomes_yeasts_index
```

3. Align each mate (from the file sequences_mate1.fastq) independently using the most sensitive mode of the alignment software. We recommend being very stringent when mapping the reads against the genome (whether from mixed samples of from unique samples) to minimize alignment mistakes (*see* **Note 18**).

```
> bowtie2 -x genomes_yeasts_index -p6 --sam-no-hd --sam-no-sq --quiet --local --very-sensitive-local -S p1.sam sequences_mate1.fastq
```

```
> bowtie2 -x genomes_yeasts_index -p6 --sam-no-hd --sam-no-sq --quiet --local --very-sensitive-local -S p2.sam sequences_mate2.fastq
```

4. The alignment software generates a SAM file (Sequence Alignment/Map format). You can select and keep the fields relevant for subsequent analysis to save memory space using awk. For instance,

```
> awk '{print $1,$3,$4,$2,$5;}' p1.sam > p1.sam.select
```

To recover the pair-end information, a convenient and fast way is to use the bash commands "sort" and "paste":

```
> sort -T /path_to_temporary_repository p1.sam.select > p1.sam.select.sorted
> sort -T /path_to_temporary_repository p2.sam.select > p2.sam.select.sorted
> paste p1.sam.select.sorted p2.sam.select.sorted > p1_p2.select.merged
```

where /path_to_temporary_repository points to a temporary repository where storage space is available for the sort command to be executed.

5. Finally, complete the mapping procedure by filtering ambiguous hits on the genome. Only the pairs of mapped reads with a quality above a certain threshold will be retained (*see* **Note 19**).

```
> awk '{$5 >= 40 && $10 >= 40) print $0;}' p1_p2_merged > p1_p2_merged.MQ40
```

3.2.2 Building the Contact Network

Assign every pair of mapped read on their restriction fragment by crossing the coordinates of the mapped reads with the restriction map of the genome. This can be done for instance with the "restrict" function from the bioinformatics suite EMBOSS (*see* **Note 20**).

3.2.3 Filtering Out of Non-informative Contacts and Construction of the Contact Map

1. A possibility that always arise at the ligation step is the formation of a loop from a long DNA molecule that contains a successive of restriction fragments that are not cut by the restriction enzyme. These events can lead to the detection of false long

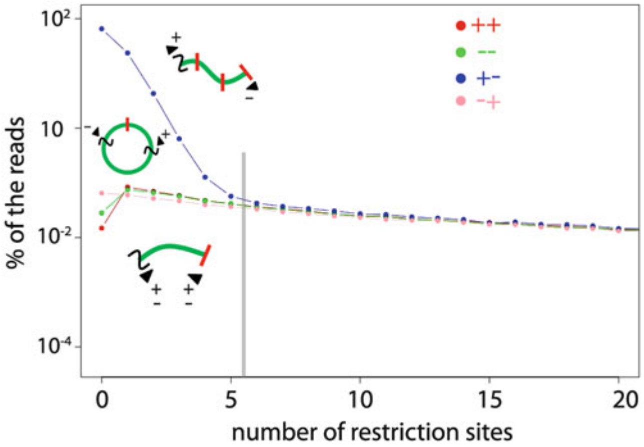

Fig. 2 Distribution of the different types of molecules (*green lines*) sequenced from a 3C library, plotted as a function of the number of restriction sites (*red bars*) between the two pair-reads (*black arrowheads*). Sonicated sites are indicated with *black twisted lines*. The directionality of the reads according to the coordinates of the reference genome are indicated by + and – symbols and are used to characterize different sub-population in the library (*see* Cournac et al. [16] for details). If the religation events were truly random between two RFs, each one of the four extremities of two restriction fragments would have the same probability to be ligated with each of the three others. However, because restriction efficiency is far from being 100 % and because of the occurrence of other types of religation events (circularization etc.), neighboring RFs present variations in the distribution of ligation events, with "uncuts" or "loops" events being overrepresented which do not reflect "contact" information but biochemical or physical biases. The *grey bar* indicates the number of restriction sites needed before all the different categories of events are equally represented within the population, corresponding to molecules that are retained for further analysis

range contacts and to avoid them, it is necessary to filter out some contacts ([16]; Fig. 2). Once this threshold is estimated PE reads that do not present this significant number of RF between them are discarded from the analysis (*see* **Note 21**).

2. To reduce the dimension of the contact map, and alleviate some local variations resulting from the size of RF, neighboring RF can be pooled into "bins" regrouping the sum of contacts of successive RF. These bins can be made either of a fixed number of successive RF, or in a window constant in size (*see* **Note 22**).

3.2.4 Normalization and Representation of the Contact Map

1. The raw matrix is then normalized to attenuate biases inherent to the protocol (*see* **Note 23**). Several methods can be used to achieve this step, including the Sequential Component Normalization (SCN; [16]; *see* **Note 24**).

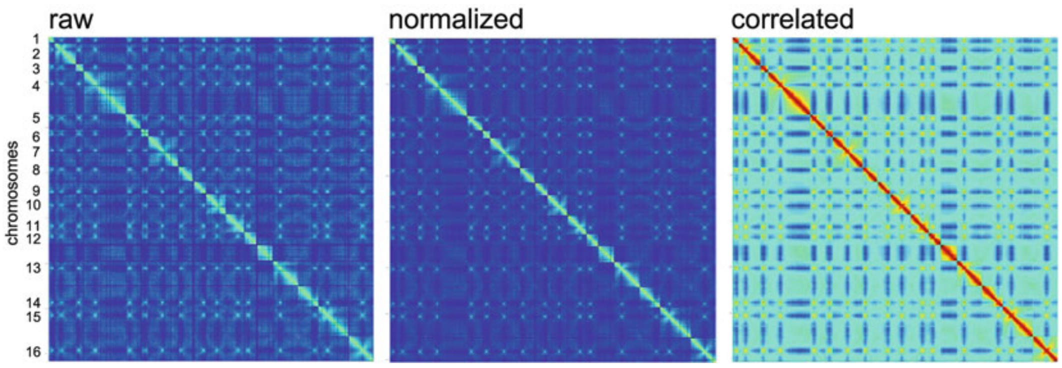

Fig. 3 (**a**) Raw genomic contact map of *Saccharomyces cerevisiae*, with each vector of the matrix representing ten restriction fragments. (**b**) The same matrix after SCN normalization. (**c**) Pearson correlation representation of the normalized matrix

For instance, using MatLab (Mathworks, Natick, MA, USA), the contact map Mat is normalized through several iterations of the following commands:

```
> for i=1:1:n1
> Mat2(:,i)=Mat(:,i)/norm(Mat(:,i));
> End
> for i=1:1:n1
> Mat_scn(i,:)=Mat2(i,:)/norm(Mat2(i,:));
> end
```

2. The correlation contact map of the normalized matrix can also be computed (*see* **Note 25**; Fig. 3). Using Matlab, the corrcoef function calculates the Pearson correlation coefficient between each line and column of the normalized map.

```
> Mat_corr=corrcoef(Mat_scn);
```

3. To visualize the contact map, the imagesc function from Matlab is a convenient tool. The contrast can be improved by raising all the elements of the matrix to the power n, with for instance n=0.4 (*see* **Note 26**).

```
> figure, imagesc(mat.^0.4);
```

3.2.5 Statistical Analysis

1. Although genomic contact maps unveil the global genome organization of a population of cells, such as centromere clustering that reveal the position of centromeric sequences in yeast species [1, 4], a large fraction of the information contained in the data is not directly visible on the contact map and need to be statistically exploited. The statistical analysis aims at determining whether the contacts observed between two or more DNA regions of interests is higher than expected by chance. One way to do that is to calculate the mean of normalized interactions between the different members of the group

of interest and compare them to a random set to evaluate significance (*see* **Note 27**). This implies carefully designing a null model taking into account the specificity of the global chromosome organization.

4 Notes

1. The choice of the restriction enzyme and buffer is a key component of this experiment. Several restriction enzymes become inactive under the experimental conditions described in the protocol (i.e. cellular brut extract of yeasts). The cheapest enzymes—usually the best characterized—provide the best candidates to generate a 3C library. Consequently, we highly recommend choosing «classical» restriction enzymes when designing the experiment. However, it is still possible that the enzyme selected is not active enough (for instance SacII works pretty bad in our hands; MM, personal communication). Restriction buffer used to generate 3C libraries have to contain DTT. Consequently, we strongly suggest avoiding new NEB buffer (NEBuffer 1.1, 2.1, 3.1 and CutSmart).
2. Building a 3C library is also entirely dependent on the match between the restriction enzyme chosen and the condition of the fixation step (Fig. 4). The likelihood for a RF to be cross-linked is dependent on the probability for one bp to be cross-linked and thus on the incubation parameters in the presence of a fixative agent level [16]. Notably, a 4-cutter (restriction enzyme recognizing a 4pb site) will require a higher concentration of cross-linking agent (or longer incubation, to

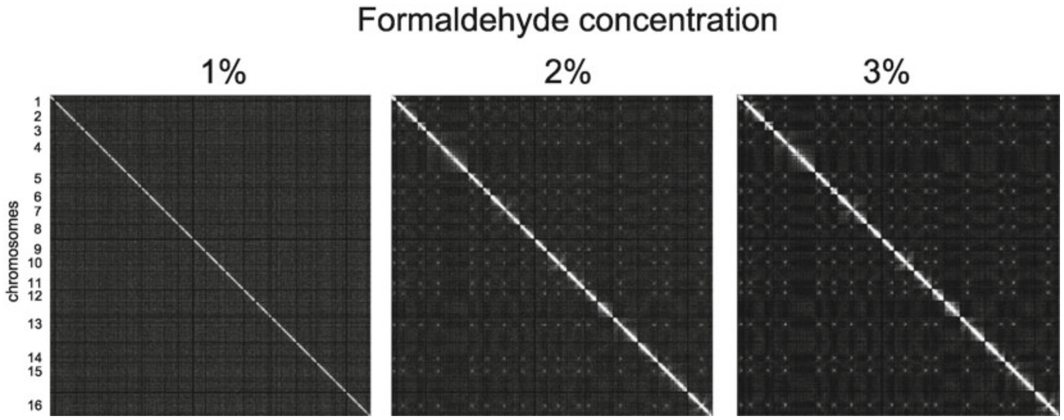

Fig. 4 Illustration of the influence of cross-link concentration on the genome-wide contact profile. *S. cerevisiae* contact maps obtained as described in this methods, but with varying formaldehyde concentrations (1, 2, and 3 %). Each contact map contains approximately the same amount of PE reads

some extent) than a 6-cutter. The protocol described in this article is designed for enzymes that generate RFs with an distribution average lower than 500 bp (± 200 bp). In general, we do not recommend enzymes that generate RFs with a distribution average lower than 300 bp.

3. Dissolving glycine at such concentration can take several hours. The process can be accelerated by gently warming the solution (40–50 °C).
4. SDS treatment is a critical step. We noticed an important drop in the quality of the library when the SDS begins to precipitate (warming prior use does not solve the problem). Therefore the SDS solution has to be changed immediately if signs of precipitation are visible.
5. ATP is a critical cofactor of the ligase reaction. In order to avoid any problem due to ATP degradation, discard the thawed aliquot after use.
6. Some yeast species—or some metabolic states—appear somehow resistant to zymolyaze treatment. Lysis is thus obtained through mechanical treatment using for instance a Precellys (Bertin Corp, Rockville, Maryland, USA).
7. We have noticed a decrease in the efficiency of the QIAgen PE buffer (i.e., wash buffer) over time. To avoid this problem we strongly recommend preparing the required amounts extemporarily to the experiment.
8. Sequences of Pair-End Adapters (with index) and Pair-End Amplification Primers can be found at http://support.illumina.com/downloads/illumina-customer-sequence-letter.html.
9. As for ATP, discard aliquot once thawed and used.
10. For size selection, we routinely use a Pippin Prep apparatus (Sage Science), though gel purification works well.
11. The mixed culture of 100 mL with a concentration of 1 × 10^7 cells of genome sizes ~10–15 Mb is sufficient to generate two libraries in the conditions described in the protocol. For other conditions, the cross-linking step will have to be adapted, as it will change the DNA–protein–formaldehyde ratio (*see* **Notes 1** and **2**).
12. We have noticed a quality decreased after storage of more than 6 months.
13. After 1 h at −80 °C, solution will froze. Prompt freezing is necessary for a good recovery of libraries.
14. The pellet does not have to be entirely dry, since DNA will be subsequently extracted by a phenol–chloroform step and re-precipitated.

15. Quantify the libraries on a gel using an image quantification software (such as Image J or Quantity One). Indeed, large DNA fragments and impurities prevent the use of NanoDrop or Qbit quantification.

16. Shearing can also be done using Bioruptor or nebulizer.

17. Importantly, the paired-end mode of the software Bowtie2 must not be used to align 3C or Hi-C data. Indeed, this mode sometimes favors wrong positions for ambiguous reads. Notably, it can favor a *cis* position for two ambiguous PE reads against a distant or *trans* alignment. This leads to "speckles" in *trans* positions between the regions involved, which often correspond to repeated sequences (for example ribosomal protein encoding genes) and can generate artefacts when analyzing co-localization of DNA regions.

18. Based on our experience, the Mapping quality threshold must be set as high as possible. Generally use a quality value of 40. Even then, incorrect mapping can still be detected. Another possibility to filter ambiguous reads is to keep the reads that do present a 'XS' field in the SAM file. The 'XS' option contains the score of the second best alignment and therefore is an indicator of a nonunique alignment. This approach is used in the python library hiclib [17]. However, in our hands, this approach is relatively less stringent than putting a threshold on the Mapping Quality.

19. A typical 3C-seq library contains a large amount of molecules that do not result from religation of two non-adjacent RF, and therefore that do not bring information about the 3D genome structure and have to be removed. To do so, plot the distribution of events according to the orientation of the pair-end reads compare to the reference genome coordinates. We distinguish at least three categories [16]. First, «uncut» events that correspond to mate pairs separated by none, or a few consecutive RFs, most likely not digested. "Loops", that correspond mostly to one or several consecutive RFs circularized during the ligation step and subsequently sheared. Finally, "weird" events which are pairs of reads belonging to the same restriction fragment but with same directions with respect to the reference genome, and may eventually be exploited to look at sister chromatids or homologs behavior. Indeed, thistype of events is largely accountable from the presence of several copies of the same DNA molecule within the same cellular compartment, as shown notably by the small number of religation events between DNA molecules belonging to different cellular compartments [7]..

20. We assign each read to its restriction fragment along the genome using a custom-made C routine. The C language is

fast and allows precise allocation of memory. At this step, the distance between each read and the associated restriction site and keeps as well the size of the associated restriction fragment can also be calculated. Several other filters can be applied at this stage to remove incorrect events. Notably, the size of the fragment sent to the sequencer can also be calculated and the distribution can be checked to be in accordance with the experimental one (*see* **step 11** selection with Pippin Prep). A filter can also be applied to reads too close of their associated restriction site.

21. The "visibility" of RFs in the PE reads will depend principally of their size, and whether they contain repeated sequences or not. This variability will impact on the visibility of the bins. Therefore, working with fixed size bins may reflect this variation in visibility. There is no perfect solution, since working with bins made of successive RFs (and therefore representing regions of different sizes) can also generate visual discrepancies between bins: for instance, a bin made of a successive long RFs will be represented in the contact map the same way as a bin made of successive small RFs. Working with fixed size bins can also simplify the analysis and comparison between contact maps. In this case, the genome is divided in equal size bins and every reads is attributed to the bin where its start position belongs.

22. The rationale behind the normalization procedure is that each bin has an equal probability to be detected overall. Plotting the distribution of contact per bin reveal a subset of elements that are undetectable or present only a handful of contacts. These bins can either result from repeated sequences that prevent confident mapping of reads in these regions, but potential structural variations between the genome of the strain being tested and the reference genome can also generate a similar outcome (for instance, deletion can easily be detected with such approaches, since consecutive bins will present no contacts; [19]. The tail of this distribution has to be removed since these bins correspond clearly to "invisible" regions. The remaining graph presents a clear Gaussian distribution, meaning that some bins are less detectable than others. Such differences result notably from variations in the distribution of RS if the bins in the contact map are made of constant sizes, or from differences in the sizes of RF binned together if this binning approach has been retained. To attenuate the differences of detection, divide each matrix element by the sum of elements of the line it belongs then do the same by dividing each matrix element by the sum of the column it belongs. Iterate this process until the matrix converges to a stable one. To our experience, a few iterations (5–10) is sufficient to have a stable

matrix. The normalization procedure ensures that the sum over the column and lines of the matrix equals 1, which reduces the noise and biases inherent to the protocol. You can do that in a matlab script, C code, or python code.

23. The SCN procedure is based on similar approaches and mathematical operations than the ICE procedure published concomitantly [17]. To our knowledge, it gives similar results.

24. Each element of the correlation matrix corresponds to the Pearson coefficient between the line and column vectors. Correlation map are to be handled with care, since they do not reflect necessary important contacts between two elements, but provide indications about their behavior similarity. This representation will increase the contrast between elements presenting similar neighbors, and others positions, but will not provide indications on the strength of the contacts between these elements.

25. "Beautification" of the contact map can be increased by applying a blurring effect on the matrices. Applying a convolution matrix with as kernel the 3×3 matrix [0.05 0.05 0.05; 0.05 0.05 0.05; 0.05 0.05 0.05] to the contact map will result in such effect. The convolution has to be repeated to emphasize the structures. You can as well display the matrices with R which contains several Gaussian filters built in functions or in python with the tool *imshow* which contains an interpolation function as default. It has to be noted that this type of image processing adds information to the initial data, and that statistical analysis cannot be done with such processed matrixes.

26. Choosing a relevant null model is not trivial. The minimalist null model must respect the distribution along the different chromosomes as proposed in [20]. A more stringent null model has to take into account the global organizational features of the genome being studied. For yeasts, whose chromosomes are organized under a Rabl organization with centromeres co-localizing and chromosome arms extending from there in the nuclear space, the positions along the genome of the elements being studied has to take into account their distance from the centromeres and, eventually, from the subtelomeric regions. Otherwise, if a subset of genes appear to be positioned at equal distances from their respective centromeres they will mechanically present enriched contacts due to the constraint imposed by centromere clustering. For instance, studying colocalization of coregulated, paralogous genes in *S. cerevisiae* raises this question accurately, since many of those genes originated from whole genome duplication events and have remained at relatively equal distances from centromeres. Failing to take into account this disposition in the null model

will lead to the conclusion that coregulated genes are colocalizing in space, which may be true, but impossible to assert since this colocalization can alsosimply reflect the distance separating them from their respective centromeres. Similar precautions apply when studying the colocalization of regions positioned along the same chromosome, which will mechanically present enriched contacts compared to the average contacts over the entire genome, if the distance separating them not included in the null model.

Acknowledgements

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 20100600373.

References

1. Duan Z, Andronescu M, Schutz K et al (2010) A three-dimensional model of the yeast genome. Nature 465:363–367
2. Dekker J, Rippe K, Dekker M et al (2002) Capturing chromosome conformation. Science 295:1306–1311
3. Sexton T, Yaffe E, Kenigsberg E et al (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148:458–472
4. Marie-Nelly H, Marbouty M, Cournac A et al (2014) Filling annotation gaps in yeast genomes using genome-wide contact maps. Bioinformatics 30(15):13
5. Lieberman-Aiden E, van Berkum NL, Williams L et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293
6. de Laat W, Dekker J (2012) 3C-Based technologies to study the shape of the genome. Methods 58:189–191
7. Marbouty M, Cournac A, Flot J-F et al (2014) Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. eLife 3, e03318
8. Oza P, Jaspersen SL, Miele A et al (2009) Mechanisms that regulate localization of a DNA double-strand break to the nuclear periphery. Genes Dev 23:912–927
9. O'Sullivan JM, Tan-Wong SM, Morillon A et al (2004) Gene loops juxtapose promoters and terminators in yeast. Nat Genet 36:1014–1018
10. Lajoie BR, Dekker J, Kaplan N (2015) The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods 72:65–75
11. Gavrilov AA, Gushchanskaya ES, Strelkova O et al (2013) Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic Acids Res 41(6):3563–3575
12. Nagano T, Lubling Y, Stevens TJ et al (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502: 59–64
13. van de Werken HJG, Landan G, Holwerda SJB et al (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods 9:969–972
14. Thompson DA, Roy S, Chan M et al (2013) Evolutionary principles of modular gene regulation in yeasts. eLife 2, e00603
15. Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 43:1059–1065
16. Cournac A, Marie-Nelly H, Marbouty M et al (2012) Normalization of a chromosomal contact map. BMC Genomics 13:436

17. Imakaev M, Fudenberg G, McCord RP et al (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 9:999–1003
18. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
19. Marie-Nelly H, Marbouty M, Cournac A et al (2014) High-quality genome (re)assembly using chromosomal contact data. Nat Commun 5:5695
20. Witten DM, Noble WS (2012) On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. Nucleic Acids Res 40:3849–3855