Redesigning chromosomes to optimize conformation capture (Hi-C) assays

Muller Heloise[1,2*], Scolari F. Vittore[1,2*], Mercy Guillaume[1,2], Agier Nicolas[3,4], Descorps-Declere Stephane[5], Fischer Gilles[3,4], Mozziconacci Julien[6,7], and Koszul Romain[1,2]

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

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

[3] Sorbonne Universités, UPMC Univ. Paris 06, Institut de Biologie Paris-Seine UMR 7238, Biologie Computationnelle et Quantitative, F-75005, Paris, France

[4] CNRS, Institut de Biologie Paris-Seine UMR7238, Biologie Computationnelle et Quantitative, F-75005, Paris, France

[5] Institut Pasteur, Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), Paris, F-75015, France

[6] Sorbonne Universités, Theoretical Physics for condensed matter lab, UPMC University Paris 06, 75005 Paris, France

[7] CNRS, UMR 7600, 75005 Paris, France

Abstract

In all chromosome conformation capture based experiments the accuracy with which contacts are detected varies considerably because of the uneven distribution of restriction sites along genomes. Here, we redesigned and reassembled in yeast a 145kb region with regularly spaced restriction sites for various enzymes. Thanks to this design, we enhanced the signal to noise ratio and improved the visibility of the entire region as well as our understanding of Hi-C data, while opening new perspectives to future studies.

Results

Genomic derivatives of the capture of chromosome conformation assay (3C, Hi-C, Capture-C)(Lieberman-Aiden *et al*, 2009; Dekker *et al*, 2002; Hughes *et al*, 2014) are widely applied to decipher the average intra- and inter-chromosomal organization of eukaryotes and prokaryotes (Sexton *et al*, 2012; Le *et al*, 2013; Dekker *et al*, 2013; Marbouty *et al*, 2014a). Formaldehyde cross-linking followed by segmentation of the genome by a restriction enzyme (RE) are the first steps of the experimental protocol. The basic unit of "C" experiments therefore consists in restriction fragments (RFs) that are subsequently religated and captured to identify long range contacts. The best resolution that can be obtained is directly imposed by the positions of the RE sites along the genome. Both 6-cutter and 4-cutter REs have been used (Marie-Nelly *et al*, 2014; Sexton *et al*, 2012; Rao *et al*, 2014; Le *et al*, 2013), the latter with the expectation that the resolution increases with the number of sites. However, this approach suffers from a major caveat: restriction sites (RSs) are not regularly spaced along the genomes. The distribution of RFs lengths follows a geometric distribution, with important variations along the genome that depend on the local GC content and the specific sequence recognized by the RE. Given that the likelihood for a RF to be crosslinked by formaldehyde during the first step in the procedure depends on its length (Cournac *et al*, 2012), the probability to detect a given fragment in any 3C experiment will in turn be strongly affected by this parameter (**Fig 1A**). Normalization procedures have been developed in order to correct the signal (Cournac *et al*, 2012; Imakaev *et al*, 2012) but these methods involve filtering out fragments with unusually low or high signal and aggregating the contact data over several consecutive fragments in longer bins

of fixed genomic length, at the expense of actual resolution (Lajoie *et al*, 2015). Overall, the definition of Hi-C resolution has remained empiric, because of the lack of a control sequence where RF biases would be alleviated.

In order to investigate and increase the resolution of 3C-based experiments, we designed and assembled a dedicated "synthetic" genomic region. As a proof of concept of this strategy, we describe here a redesigned ~150kb region (called here synIV-3C) of budding yeast chromosome 4. This designer chromosome closely resembles the native chromosome with respect to genetic elements (see **Supplementary Note 1** and **Fig S1**), but was "designed" to yield high resolution and high visibility in 3C experiments by providing nearly equally spaced restriction sites. The RSs of four different enzymes were removed from the native sequence with point mutations and subsequently reintroduced within the sequence at regularly spaced positions (400bp, 1,500bp, 2,000bp and 6,000bp for DpnII, XbaI, HindIII and NdeI, respectively; **Fig 1B** and **Fig S2**). As shown on **Fig 1C**, the DpnII and HindIII RFs sizes in the redesigned synIV-3C region are normally distributed when compared to the highly skewed, native genome-wide distributions. Besides providing a way to increase the resolution of the 3C experiment, the design can also be used to focus on specific functional contacts, for instance between promoters and terminator regions of genes (**Fig S2**). When possible, coding sequences were targeted preferentially and modified using synonymous mutations (**Fig S1**). We identified a 150kb window on chromosome 4 for which the uniformity of RFs lengths was maximized while the number of potentially deleterious base changes was minimized (the final choice for the region can also take into account sequence annotation and guided primarily by specific interests of the end-user). From this design, DNA building block were purchased and assembled as described (Annaluru *et al*, 2014; Muller *et al*, 2012) (**Supplementary Note 2**). Sequencing confirmed that 144kb within the targeted region were replaced by the redesigned sequence and that 100% of the mutations were introduced at the correct positions corresponding to a total of 2% divergence with the reference genome. No significant growth defects were detected in the synthetic strain (**Fig S3**).

We then performed Hi-C experiments on the strain carrying the synIV-3C redesigned chromosome as well as in a wild type strain using DpnII and HindIII (**Supplementary Note 3**). The raw DpnII contact map of chromosome 4 exhibited a remarkably "smooth" pattern within the redesigned region compared to the native flanking regions (**Fig D**). The read coverage over the region also exhibits a dramatic and compelling change, with a more homogeneous and regular distribution in the synthetic regions for both enzymes compared to a highly heterogeneous distribution in the native sequence (**Figs 2A**, **B**). Interestingly, careful examination of this distribution indicates that besides its own length, the capture frequency of a given fragment is also influenced by the length of its neighbors. To quantify the improvement in the SynIV-3C region we compared the signal with the signal over the same region obtained in the WT strain using the same number of aligned read pairs and identical bins of various sizes (**Figs 2C**, **D**). At the smallest resolution tested (600bp for DpnII and 2,400bp for HindIII) the WT contact map exhibited numerous blind regions with no detectable contacts (empty bins), in sharp contrast with its synthetic counterpart (**Fig 2C**, **D**). When fragments were aggregated in bins of increasing sizes (hence, resulting in a loss of resolution) these blind regions gradually disappear, although the heterogeneity of the data remains consistently higher in the WT compare to synIV-3C strain, as showed by the increased span of the color-scales of the WT maps.

In order to further quantify this heterogeneity, we computed the cumulative distributions of the number of contacts between bins separated by a given genomic distance ***s*** (bp) in the synIV-3C region and in its native counterpart for DpnII and HindIII (**Figs 2C** and **2D**, respectively). The redesigned region systematically exhibited more homogeneous contacts counts and narrower distributions than the WT region, both at short (s = 2 x bins sizes; **Figs 2C** and **D** middle panels) and longer distances (**Supplementary Note 5** and **Figs S4, S5**). Some of the bins in the native region remain almost invisible to the assay as a result of the heterogeneity in RF distribution (blue squares on **Figs 2C** and **D** middle panels). We computed the coefficient of variation CV (i.e. standard deviation /mean) of these distributions for multiple values of ***s***. We use this value as an indication of the signal to noise ratio (**Figs 2C** and **D** right panels). Interestingly, we found that even for large bins, the CV is significantly and consistently smaller in the synthetic region, again indicating improved resolution. These results also clearly illustrate

the advantage of using a frequent cutter (DpnII vs. HindIII) restriction enzyme with respect to resolution since the distribution of contact counts between bins remains much more spread with HindIII than with DpnII, even for native sequences (**Fig 2B**).

Chromosome conformation capture is a dynamic field: two approaches using modified restriction patterns have been recently used to increase/improve the resolution, DNAse Hi-C and Micro-C (Hsieh *et al*, 2015) (note also that enrichment steps of regions of interest do not alleviate the limitations associated to the natural restriction pattern described above, thus have no effect on the resolution *per se*). DNAse-HiC captures contacts between open chromatin sites. DNAse Hi-C was not been performed in yeast and therefore we did not compare Syn-3C with this approach. However, given the fact that DNAse sensitive sites are found approximately every 3 kb along the yeast genome (Ma *et al*, 2015), it is expected that DNAse Hi-C would give results comparable to Hind-III Hi-C. Micro-C, on the other hand exploits micro-Coccal nuclease (Mnase) to digest DNA rather than a restriction enzyme. This approach generates nonspecific cuts in-between nucleosomes (every ~160bp), resulting in a relatively regular restriction pattern. Micro-C reads were reprocessed and the outcome compared to Syn-3C redesigned region along chromosome 4. Although the Micro-C reads density is overall more regular than for a WT Hi-C experiment, nucleosome free regions generate some inhomogeneity in the distribution. At short distances (600bp) Micro-C and Syn-3C compared well, but the signal to noise ratio quickly drops for Micro-C at larger distances. In this frame, the two approaches aim at different objectives: whereas Micro-C captures well small domains, Syn-3C appears as an approach of choice to concomitantly i) improve the visibility of any given region from ~500bp and above, and most importantly ii) track *trans* interactions as well as iii) homologs.

The yeast genome presents a relatively homogeneous GC content and few repeated sequences. The gain in resolution achieved by redesigning RS along the genome should therefore be even higher in organisms with more heterogeneous genomic content and will enable unbiased tracking of entire regions that are otherwise inaccessible to the experiment. One could envision, for instance, assembling the redesigned chromosome in yeast (Benders *et al*, 2010), before targeting the sequenced to replace its native counterpart in the organisms of

interest (such as a bacteria, or eventually on mammalian cells). Other advantages of the approach include the modularity of the assembly step (**Supplementary Note 2**), that allows the introduction of building blocks carrying genetic elements of interest within the redesigned region. For instance, one could introduce highly expressed promoters in the middle of "gene desert" areas, to investigate the effect of gene expression on the local chromatin structure. One can also "shuffle" some of these building blocks, to look at the influence of specific DNA binding proteins on the contact networks. In addition, an interesting follow up to this study is to cross our synIV-3C strain with a WT strain (or a strain with a different design) in order to resolve at the same time both homologs in a single experiment. Finally, the combination of Capture-C (Hughes *et al*, 2014) like approaches, which enrich the regions of interests (tough without alleviating the inherent biases) to investigate the synthetic region will also boost the analysis depth to unprecedented levels. This specific 3C-friendly design is the first time, to our knowledge, where a large (>100kb) region of chromosome is specifically redesigned and assembled for the purpose of improving an assay so that we can now address more precisely and accurately specific questions related to the biology of the cell. It paves the way to more studies exploiting the power of synthetic biology to boost, refine, and maybe reshape traditional molecular biology approaches through orthogonal ones.

Acknowledgements

We thank Jef Boeke for the Sc2.0 PCRTags sequences for chromosome 4 and for fruitful discussions and comments on the manuscript. We thank Elodie Pirayre and Ivan Moszer for contributing to the initial steps of the design of the algorithm, and Axel Cournac, Martial Marbouty, Luciana Lazar-Stefanita, Olivier Espeli and Bertrand Llorente for discussions. 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), from Agence Nationale pour la Recherche (MeioRec ANR-13-BSV6-0012-02), and from ERASynBio and Agence Nationale pour la Recherche (IESY ANR-14-SYNB-0001-03).

Author contributions

HM assembled the chromosome and performed the Hi-C experiments with help from GM. HM, SDD, NA, GF and RK designed the sequence. VS and JM analyzed the data. JM and RK wrote the article. RK conceived the study.

Figure Legends

Figure 1 - synIV-3C design and assembly.

A Number of contacts made by RFs as a function of their size (HindIII (red) or DpnII (blue) in the native sequence. Top panel: log-lin scale. Bottom panel: log-log scale.).

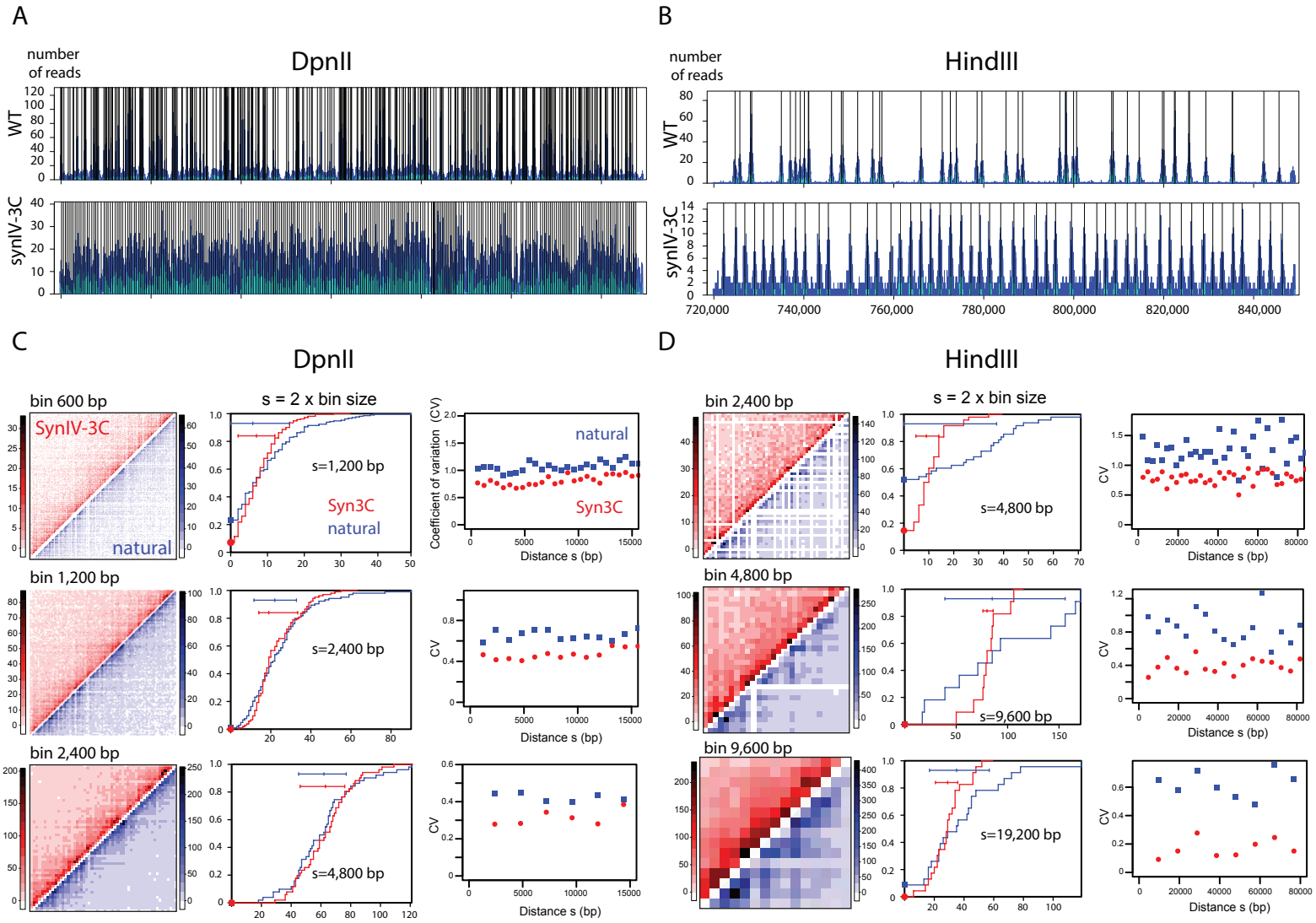
B Illustration of the design principles of the synIV-3C sequence for the DpnII and HindIII RSs. Black arrow: chromosome. Grey rectangles: genetic elements. Blue and red vertical lines represent the RSs positions for the enzymes DpnII and HindIII, respectively. Top panel: restriction pattern of a (hypothetical) native sequence. Bottom panel: restriction pattern after synIV-3C design, with the RSs defining regularly spaced intervals.

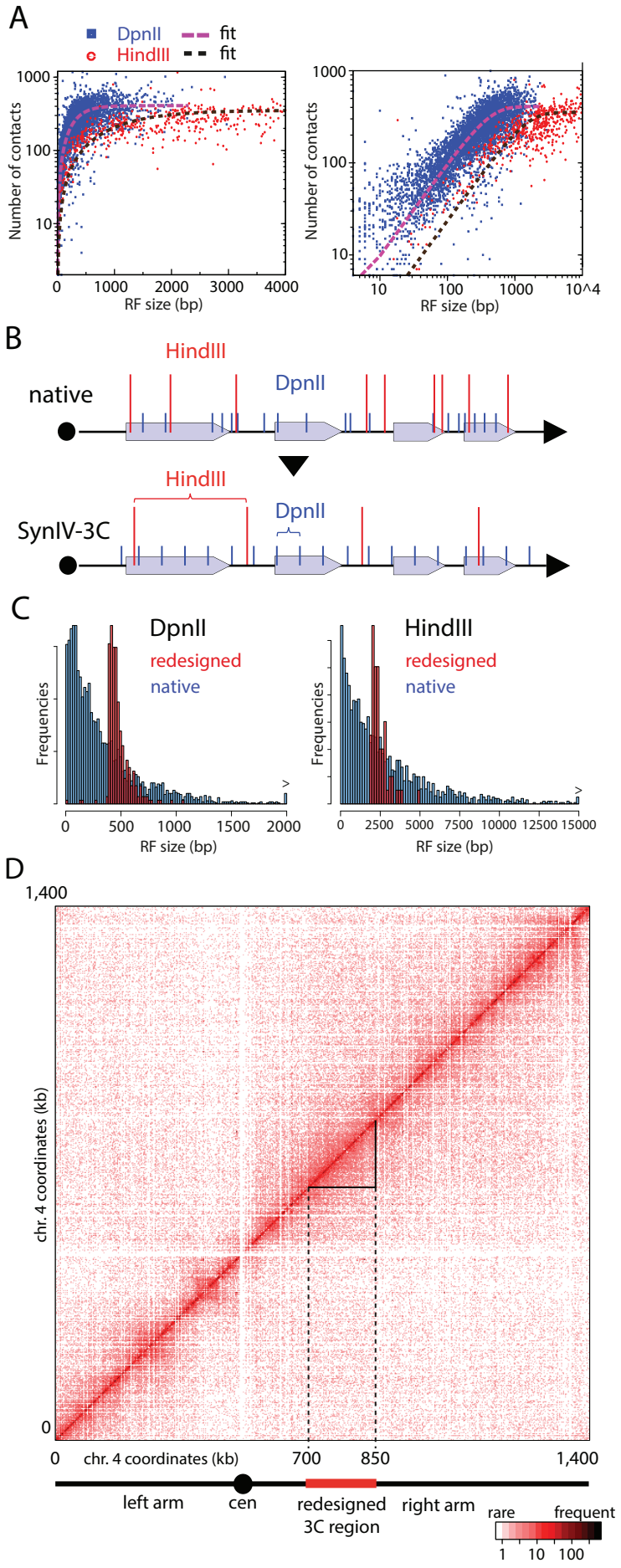
C Distribution of the DpnII (left) and HindIII (right) RFs sizes in both the native and synIV-3C 150kb redesigned sequence (red and blue, respectively).

D Raw DpnII contact map of the Hi-C experiment performed on G1 daughter cells synchronized through elutriation(Marbouty *et al*, 2014b). Dashed lines: borders of the redesigned region. Plain black lines: borders of the contact map analyzed in **Fig. 2**.

Figure 2 – Reads coverage from Hi-C experiments performed with DpnII (A) and HindIII (B) restriction enzymes in synIV-3C and native strains, and mapped against the synthetic region and its natural counterpart, respectively. Note that the scale of the y-axis illustrates the heterogeneity of the coverage, with some positions in the DpnII map being overrepresented with respect to others.

C, D Analysis of the contact counts along the synIV-3C region for DpnII (C) and HindIII (D). First column: synIV-3C (in red) and chromosome 4 native counterpart (blue) Hi-C contact maps. For each experiment, three different fixed bin sizes were analyzed (600 bp, 1200 bp and 2400 bp for DpnII, 2400 bp, 4800 bp and 9600 bp for HindIII). Middle panels: cumulative distribution of the number of contacts between bins located at a genomic distance ***s*** from each other's (***s*** = 2 x bin size). Right panels: distribution of the coefficient of variation (CV) as a function of ***s***.


A
DpnII
number of reads
WT
synIV-3C
B
HindIII
number of reads
WT
synIV-3C
720,000
740,000
760,000
780,000
800,000
820,000
840,000
C
DpnII
bin 600 bp
SynIV-3C
natural
s = 2 x bin size
s=1,200 bp
Syn3C
natural
Coefficient of variation (CV)
natural
Syn3C
Distance s (bp)
bin 1,200 bp
s=2,400 bp
CV
Distance s (bp)
bin 2,400 bp
s=4,800 bp
CV
Distance s (bp)
D
HindIII
bin 2,400 bp
s = 2 x bin size
s=4,800 bp
CV
Distance s (bp)
bin 4,800 bp
s=9,600 bp
CV
Distance s (bp)
bin 9,600 bp
s=19,200 bp
CV
Distance s (bp)


A
DpnII
HindIII
fit
fit
Number of contacts
RF size (bp)
B
HindIII
DpnII
native
HindIII
DpnII
SynIV-3C
C
DpnII
redesigned
native
HindIII
redesigned
native
Frequencies
RF size (bp)
D
1,400
chr. 4 coordinates (kb)
0
chr. 4 coordinates (kb)
700
850
1,400
left arm
cen
redesigned
3C region
right arm
rare
frequent
1
10
100

References

Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, Richardson SM, Dymond JS, Kuang Z, Scheifele LZ, Cooper EM, Cai Y, Zeller K, Agmon N, Han JS, Hadjithomas M, Tullman J, Caravelli K, Cirelli K, Guo Z, London V, et al (2014) Total Synthesis of a Functional Designer Eukaryotic Chromosome. *Science* **344:** 55–58

Benders GA, Noskov VN, Denisova EA, Lartigue C, Gibson DG, Assad-Garcia N, Chuang R-Y, Carrera W, Moodie M, Algire MA, Phan Q, Alperovich N, Vashee S, Merryman C, Venter JC, Smith HO, Glass JI & Hutchison CA (2010) Cloning whole bacterial genomes in yeast. *Nucleic Acids Res.* **38:** 2558–2569

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

Dekker J, Marti-Renom MA & Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14:** 390–403

Dekker J, Rippe K, Dekker M & Kleckner N (2002) Capturing chromosome conformation. *Science* **295:** 1306–1311

Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N & Rando OJ (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162:** 108–119

Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, De Gobbi M, Taylor S, Gibbons R & Higgs DR (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46:** 205–212

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J & Mirny LA (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9:** 999–1003

Lajoie BR, Dekker J & Kaplan N (2015) The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods San Diego Calif* **72:** 65–75

Le TBK, Imakaev MV, Mirny LA & Laub MT (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342:** 731–734

Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES & Dekker J (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326:** 289–293

Ma W, Ay F, Lee C, Gulsoy G, Deng X, Cook S, Hesson J, Cavanaugh C, Ware CB, Krumm A, Shendure J, Blau CA, Disteche CM, Noble WS & Duan Z (2015) Fine-scale chromatin

interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods* **12:** 71–78

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

Marbouty M, Ermont C, Dujon B, Richard G-F & Koszul R (2014b) Purification of G1 daughter cells from different Saccharomycetes species through an optimized centrifugal elutriation procedure. *Yeast* **31:** 159–166

Marie-Nelly H, Marbouty M, Cournac A, Liti G, Fischer G, Zimmer C & Koszul R (2014) Filling annotation gaps in yeast genomes using genome-wide contact maps. *Bioinforma. Oxf. Engl.* **30:** 2105–2113

Muller H, Annaluru N, Schwerzmann JW, Richardson SM, Dymond JS, Cooper EM, Bader JS, Boeke JD & Chandrasegaran S (2012) Assembling large DNA segments in yeast. *Methods Mol. Biol. Clifton NJ* **852:** 133–150

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES & Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A & Cavalli G (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148:** 458–472

METHODS

Redesigning chromosomes to optimize conformation capture (Hi-C) assays

Muller Heloise[1,2*], Scolari F. Vittore[1,2*], Mercy Guillaume[1,2], Agier Nicolas[3,4], Descorps-Declere Stephane[5], Fischer Gilles[3,4], Mozziconacci Julien[6,7], and Koszul Romain[1,2]

p.2 **Supplementary Note 1.** Design principles of Syn-3C chromosomes.

p.7 **Supplementary Note 2.** Assembly of the redesigned chromosome.

p.9 **Supplementary Note 3.** Hi-C experiments and contact maps generation

p.10 **Supplementary Note 5.** Statistical analysis

p.11 **Supplementary Note 4.** Processing of the reads and contact maps generations

Design principles of Syn-3C chromosomes.

We aimed at modifying the native sequence of a budding yeast chromosome according to our design principles while introducing as little modifications as possible. Because we were planning on re-assembling only a 150kb window within the genome, we scanned through the overall sequence using a scoring quality function to look for the candidate regions qualifying as the ideal target, i.e. where our principles would introduce a minimal number of mutations.

The starting material was the *S. cerevisiae* SK1 strain genome sequence and annotations (Liti *et al*, 2009) and a list of 9 restriction enzymes (EcoRI, HindIII, NdeI, PstI, SacI, SacII, SalI, XbaI, XhoI and DpnII). RE were selected based on their low cost and restriction efficiency. A genome index file was then computed, that contained the following information for each base pair:

- Whether it consists of a "forbidden mutation" site, defined by us as follow: i) start and stop codons of known ORFs, ii) regulatory transcription pre-initiation complexes binding regions identified through ChIP-Seq exo, encompassing TATA-box binding sites (Rhee & Pugh, 2012), iii) the consensus sequence of Autonomous Replicating Sequences (ARS), i.e. the core sequence within *S. cerevisiae* replication origins (list of ARS obtained from oridb (Siow *et al*, 2012), iv) intron borders, v) centromeres, vi) tRNA.
- Whether the position belongs to a restriction site.
- If it belongs to an intergenic or coding region, and in the latter case, the codon it belongs to and its position.

Sliding windows of 150 kb moving with 10kb steps were then generated over the entire genome.

In parallel, we defined the restriction pattern we wanted to generate:

- Regularly spaced intervals for 400, 1,500, 2,000 and 6,000 bp
- Gene promoter/terminator (substitutions within a coding sequence strongly preferred)

For each window, we computed all possible changes to apply to the genome so that all combinations of five out of the eight chosen 6-cutter enzymes were repositioned to generate all expected new restriction patterns. For each combination of 5 enzymes, all sites were first removed from the genome before being reintroduced at ideal positions. A margin of error in the positioning of the "ideal" position was tolerated (10% of the window size) to maximize the probability of introducing only synonymous mutations within the coding sequence. Once a RS was positioned, the position of the adjacent RS was adjusted based on the newly positioned site so that overall, the distribution of RFs remains as close as possible to the theoretical distribution. Overall, for each enzyme, a quality score was computed for each window based on the difference between the expected distribution of the site, and the real distribution. For each combination of enzyme, a global score corresponding to the sum of the individual scores of each enzyme was computed (see **Fig S1** for schema and more details).

Overall, we selected the 10 "best" windows located at least at 150 kb from either a centromere or a telomere. The quality score was weighted by the presence of "forbidden positions" within the window, for instance when a start codon overlaps a restriction site to be deleted. Finally, a manual curation, aiming at fixing potential conflicts (such as 2 RSs overlapping the same bases, or accidental re-creation of a RS of one enzyme when processing a second one), followed, and was performed on the genome windows presenting the best quality scores.

We chose the final window based also on our research interests, i.e. containing at least two early replicating replication origins (Siow *et al*, 2012; Raghuraman *et al*, 2001), and several hotspots of meiotic DNA double-strand breaks (Pan *et al*, 2011). We also attempted to avoid too many retrotransposable elements or other DNA repeats. The final window was positioned on chromosome IV::700,000-850,000, with restriction patterns as follow: DnpII ↔ 400 bp window; XbaI ↔ 1,500 bp window; HindIII ↔ 2,000 bp window; NdeI ↔ 6,000 bp window; HhaI ↔ promoter/terminator (see summary on **Fig S2**). 1037 mutations were present in the sequence, the vast majority corresponding to the modifications necessary to reorganize DpnII RS (**Table S1**). Overall, 1037 mutations were introduced, corresponding to 0.7% divergence.

Table S1. Mutations necessary to remove and generate new sites along chromosome 4 700,000::850,000 window.

	deletion	**new sites**
HindIII	58	61
NdeI	34	23
XbaI	25	76
dpnII	442	310
Total	559	470

In addition to these mutations, a number of other modifications were introduced into the sequence. First, PCRTags similar to those used in the Sc2.0 design(Annaluru *et al*, 2014; Dymond *et al*, 2011) specific to either the native or synthetic sequence were also introduced within the window. Performing PCR using these primers allow testing for the presence and absence of the synthetic and native sequence, respectively. PCRTags were manually curated to adapt them to the restriction design, and overall 59 PCR tags out of 154 needed to be modified accordingly.

Overall, a total of 3229 bp were modified (2% of the 150,000 bp window). 743 codons were modified, but no change in the sequences of the corresponding proteins were introduced.

Although we took great care in the design of the sequence and algorithm, our ongoing experiments nevertheless suggest slight modifications in the design principles that have the potential to facilitate both the design and the analysis of the experiment. First, windows of 400 and 1,000 bp are probably sufficient to assay the structure at a high resolution. Second, during curation one could manually remove the extremely small RFs generated throughout the process and that nevertheless pass all the quality filters (such as the tiny RFs still visible in the DpnII restriction pattern of the Syn-3C region in **Fig 1C**). Third, SNPs have to be introduced in repeats or low complexity DNA to facilitate mapping of the reads. Fourth, that the WT and Syn-3C sequence are sufficiently divergent to be analyzed using cheaper short-read technologies (such as paired-end 50bp Illumina for instance) can also be imposed by introducing extra silent mutations in the sequence. Finally, a possibility to find the best

frame for each interval/pattern is to start with the positions of RS overlapping forbidden sites as seed.

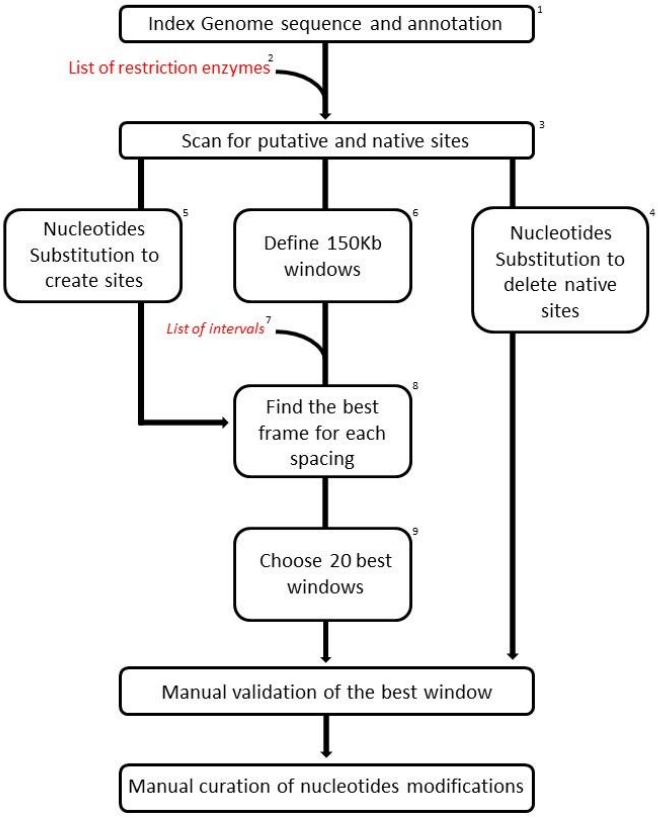


Figure S1. Diagram of the workflow.

[1] Annotation corresponds to CDS, ARS, telomere regions, retrotransposable elements, mating type loci, tRNA, Sn/Sno RNA, rDNA, ncRNA, intron motives, TATA box. All those features but CDS were labelled as « forbidden », preventing any nucleotide substitution in these regions.

[2] DpnII, HindIII, SacI, EcoRI, NdeI, SacII, SalI, XbaI and XhoI

[3] Were considered as putative restriction sites DNA sequences differing with one base pair from the RS recognized by an RE.

[4] The sequence modifications were allowed only in non-forbidden positions. In CDS silent mutations were introduced. When two sites overlapped the minimum changes needed were selected. When possible, we favored A <=> G and C <=> T substitutions. A validation step to

test whether or not the deletion of one site creates a new site was performed after each modification, and if so, a new modification was sought for.

[5] Modifications to generate new sites were also only introduced at non-forbidden positions. Only silent mutations were introduced within coding regions.

[6] 583 x 150 kb windows with 10 kb overlaps were generated over the entire genome, excluding telomeres and 75 kb from each side of centromeres.

[7] Here, 400 bp, 1,500 bp, 2,000 bp and 6,000 bp

[8] For each 150Kb window and each interval the following steps were performed:

- for each enzyme, for each starting point: putative sites within the first bin of the window (0 - 0+spacing)
- find the putative sites at position n+1 at a distance interval +/-10% from position n until the end of window

[9] For each window, a score is calculated as follows:

- for each interval, a score is calculated for each enzyme based on the Median Absolute Deviation (MAD)
- the best enzyme exhibiting the lowest score was chosen for each interval. Each spacing must have a different enzyme, so multiple combination of enzymes were computed for each window.
- The window score is calculated as the sum of the 4 chosen interval score

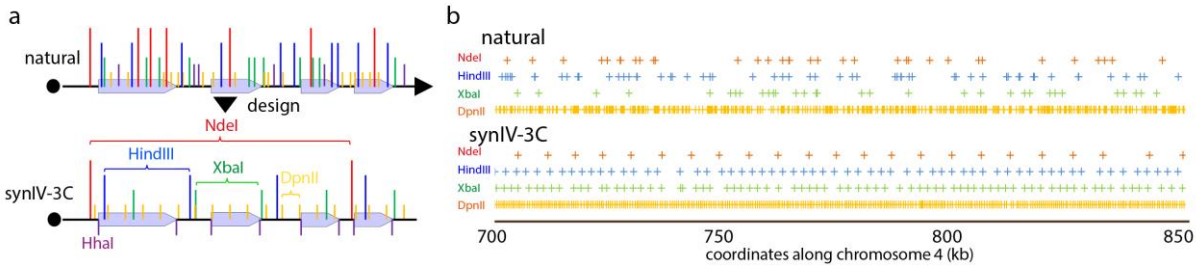


Figure S2. (**a**) Illustration of the design principles of the synIV-3C sequence. Black arrow: chromosome. Grey rectangles: genetic elements. Yellow, green, blue, red and purple vertical lines represent the restriction sites positions for the enzymes DpnII, XbaI, HindIII, NdeI and HhaI, respectively. Top panel: restriction pattern of a (hypothetical) native sequence. Bottom panel: restriction pattern after synIV-3C design, with four enzymes defining regularly spaced intervals, and the fifth one decorating genes promoter and terminator. (**b**) Distribution of the

DpnII (left) and HindIII (right) RFs sizes in both the native and synIV-3C 150kb redesigned sequence (red and blue, respectively).

Assembly of the redesigned chromosome.

The redesigned sequence was split into 52 fragments of ~3,000 bp (i.e., block), with 200 bp overlaps between them. In addition sequences corresponding either to the auxotropHi-C marker genes *URA3* or *LEU2* were added to blocks 20, 37, 52 (*URA3*) and blocks 11, 28, 47 (*LEU2*), followed by 200 bp sequences of the WT neighboring chromosomal region. The replacement of the native sequence of strain BY4742 with the redesigned blocks was performed through a succession of six transformations, up to 11 blocks at a time(Muller *et al*, 2012).

After each transformation, independent colonies were sampled and PCRs performed at the PCR tags positions to identify the transformants that have replaced all of the native sequence with the redesigned one (**Fig S3**). Upon the last transformation, the selected transformant genome was sequenced and the region 707,556-852,114 (144,558 bp) was found to be replaced by the synthetic blocks.

In parallel, growth assays were performed to see if the transformant exhibited small losses in fitness. Little to no growth defect could be identified when blocks 1 to 47 replaced the native sequence. Interestingly, the last transformation using blocks 48 till 52 led repeatedly to the recovery of transformants exhibiting a slow-growth, petite phenotype (Slonimski, 1949), reflecting a block in the aerobic respiratory chain pathway and a decrease in ATP. Since the region concerned by the 6th transformation only involved a few kb, we decided to move further regarding the analysis exploiting the 145 kb already successfully reassembled. We also observed that crossing with a WT strain gave diploids without growth defects. Sporulation of these diploids gave offsprings with growth rates also similar to WT, suggesting stable complementation of mitochondrial genomes.

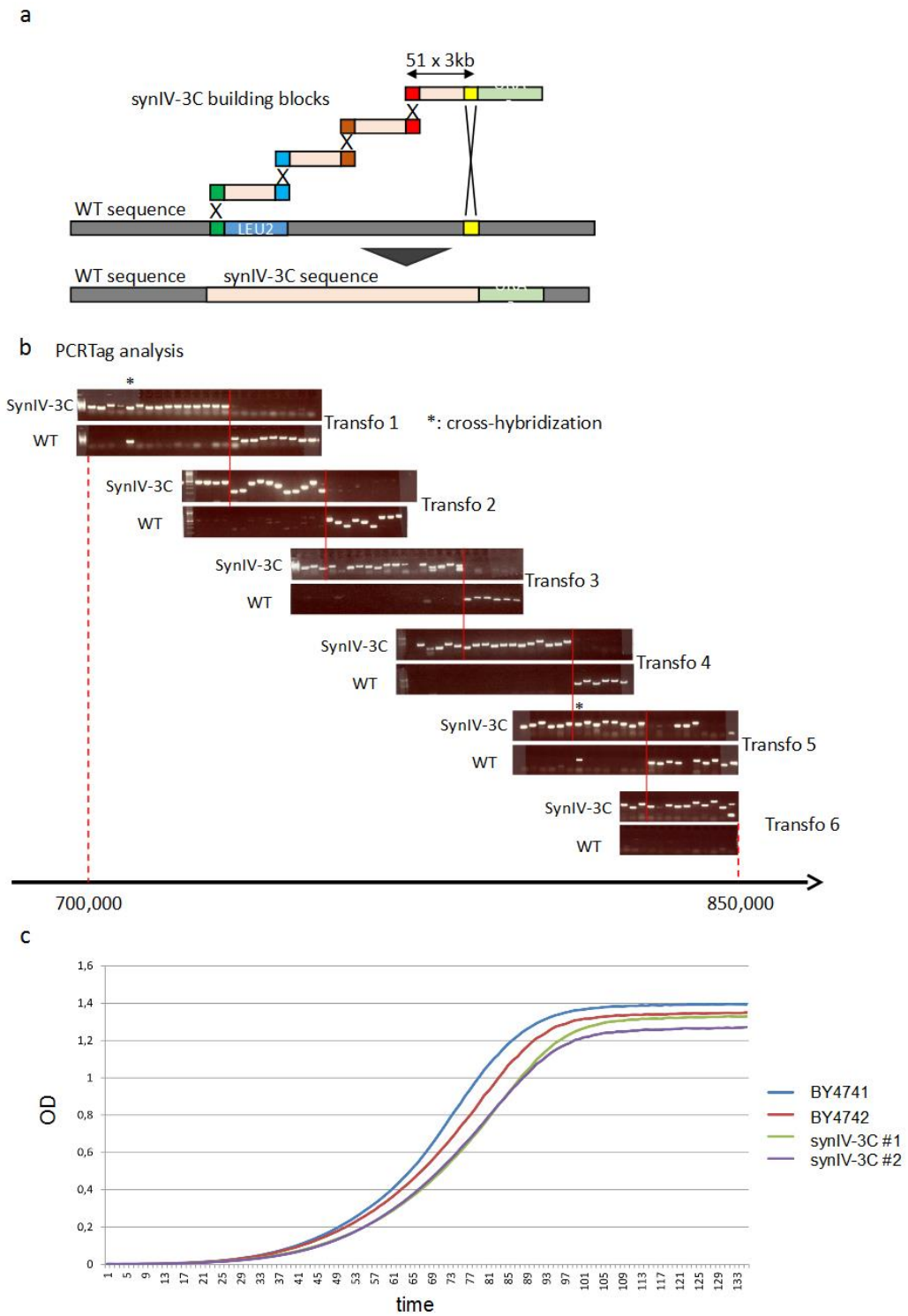


Figure S3. Strain assembly strategy. (**a**) Building blocks are iteratively integrated in the genome of S. cerevisiae through homologous recombination following transformation. An alternating auxotropHi-C selection and counterselection of uracil and leucine is performed to select for transformants likely to have replaced their native sequence by the redesigned one between the two extremities of the targeted region (see Muller *et al*, 2012; Annaluru *et al*,

2014 for details). (**b**) That the replacement of the native sequence by the synthetic one occurred over the entire targeted region is then controlled by PCR in transformants cells. For each transformation, PCR tags amplifying either the native or the redesigned genome confirm that the redesigned DNA was integrated over the entire region. The genome of the selected transformant is then sequenced as a control. (**c**) Growth curve of two independent subclones of the selected transformant vs two independent subclones of the parent lineage. Each curve is computed out of 8 independent cultures.

RNA Isolation from Yeast for RNA Sequencing

Total RNA isolation and analysis of BY4742 and Syn3C strains, were realized on three biological replicates. Single yeast colonies were grown in a 2 mL culture in YPD overnight at 30°C. The next morning, 10 mL cultures in YPD were started from 10^6 cells/mL until they reached 2.10^7 cell/mL. The cells were pelleted by spinning at 5000 rpm at 4°C for 5 min. The pellet was resuspended in 0.5 mL of Tris-HCl (10 mM, pH 7,5) and transferred to a microfuge tube. The cells were pelleted again by spinning briefly and discarding the supernatant. The cells were resuspended in 400 µL RNA TES buffer (10 mM Tris-HCl, pH 7.5 10 mM EDTA 0.5% SDS). 400 µL of acid phenol/chloroform was then added to the cells and vortexed for 1 min, and heated at 65°C for 30 minutes, briefly vortexing some time to time. The cells were placed on ice for 5 min band centrifuge at 13000 rpm at 4°C for 5 min. The aqueous layer (~400 µL) was then transferred into another microfuge tube; an equal amount of phenol/chloroform acid was added a second time, mixed well and centrifuged at 13000 rpm at 4°C for 5 min. The RNA (~400 µL) was precipitated by adding 40 µL of sodium acetate (3 M) and 1,1 mL of absolute ethanol and incubating the tube at −80˚C for a least 30 min. The

RNA was pelleted by centrifuging at 13000 rpm at 4˚C for 20 min. The RNA pellet was then washed with 500 µL of 70% ethanol, air-dried and then resuspended in 50 µL of water. 15 µg were treated with 2U of DNase TURBO (Invitrogen) and cleaned up by phenol extraction and ethanol precipitation before being prepared for sequencing.

RNA-Seq Analysis of synIII

Single-end non-strand-specific RNA-seq of the Syn3C and BY4742 were performed using Illumina Nextseq and standard TruSeq preparations kits, after depletion of ribosomal RNA. Reads were mapped using Bowtie2 to the reference S. cerevisiae BY4742 and Syn3C genome. For each gene, reads were counted if mapping quality was lower than 30 and analyzed for differential expression using DESeq2, with standard parameters.

Hi-C experiments and contact maps generation

S. cerevisiae G1 daughter cells of the redesigned strain were recovered from an exponentially growing population through an elutriation procedure (Marbouty *et al*, 2014). Hi-C libraries were generated as described (Cournac *et al*, 2015; Dekker *et al*, 2002) with introduction of a biotin-ligation step in the protocol (Lieberman-Aiden *et al*, 2009). G1 daughter cells were cross-linked for 20 minutes with fresh formaldehyde (3% final concentration). To generate the libraries with different restriction enzymes, aliquots of 3 x 10^9 cells were resuspended in 10 ml sorbitol 1M and incubated 30 minutes with DTT 5mM and Zymolyase 100T (C_{Final}=1 mg/ml) to digest the cell wall. Spheroplasts were then washed first with 5 ml of sorbitol 1M, then with 5 ml of 1X restriction buffer (depending on the restriction enzyme used). The spheroplasts were then resuspended either in 3.5 ml of the corresponding restriction buffer (NEB). For each aliquot/experiment, the cells were then split into three tubes (V=500µL) and incubated in SDS (3%) for 20 minutes at 65°C.

Crosslinked DNA was digested at 37°C overnight with 15 units of the appropriate restriction enzyme (NEB, DpnII, HindIII or NdeI). The digestion mix was then centrifuged for 20 minutes at 18000 g qnd the supernatant discarded. The pellets were then resuspended and pooled into 400 µL of cold water. Depending on the sequence of the restriction site overhangs, the extremities of the fragments were repaired in the presence of either 14-dCTP biotin or 14-dATP biotin (Invitrogen). Biotinylated DNA molecules were then incubated 4 hours at 16°C in presence of 250 U of T4 DNA ligase (Thermo Scientific, 12.5 ml final volume). DNA purification was achieved through an overnight incubation at 65°C in presence of 250µg/ml proteinase K in 6.2mM EDTA followed by precipitation step in presence of RNAse.

The resulting 3C libraries were sheared and processed into Illumina libraries using custom-made versions of the Illumina PE adapters (Paired-End DNA sample Prep Kit – Illumina – PE-930-1001). Fragments of sizes between 400 and 800 bp were purified using a PippinPrep apparatus (SAGE Science), PCR amplified, and paired-end (PE) sequenced on an Illumina platform (HiSeq2000; 2 x 75 bp). The accession number for the data reported in this paper is [Database]: [xxxx] (under completion).

Processing of the reads and contact maps generations

The raw data from each 3C experiment was processed as follow: first, PCR duplicates were collapsed using the 6 Ns present on each of the custom-made adapter and trimmed. Reads were then aligned using Bowtie 2 in its most sensitive mode against *S. cerevisiae* reference genome (native genome) or against the *S. cerevisiae* reference adapted for the Syn-3C region on chromosome 4 (SynIV-3C genome). An iterative alignment procedure was used: for each read the length of the sequence mapped increases gradually from 20 bp until the mapping became unambiguous (mapping quality > 30). Paired reads were aligned independently and each mapped read was assigned to a restriction fragment. Religation events has been filtered out through the information about the orientation of the sequences as described in(Cournac *et al*, 2012). The distribution of the reads along the synthetic region and its native counterpart is represented in **Fig 2**.

Contact matrices were built for the wild type and the mutant by binning the aligned reads into units of single restriction fragments. DpnII and HindIII contact maps for the SynIV-3C region and its native counterpart were randomly resampled in order to present the same number of contacts. The raw contact maps were then subsequently binned into units (i.e. bins) of 600, 1,200, 2,400, 4,800 and 9,600 base pairs. Contacts maps were generated using the *levelplot* function of the R *lattice* package.

Matrices for the synthetic region and 7 other control regions (see **Supplementary Note 5** below) were subsequently obtained by extracting the diagonal blocks for bins falling in the 719,756bp to 849,206bp interval, for the synthetic region, and 460,856bp to 590,306bp, 590,306bp to 719,756bp, 849,206bp to 978,656bp, 978,656bp to 1,108,106bp, 1,108,106bp to 1,237,556bp, 1,2375,56bp to 1,367,006bp, 1,367,006bp to 1,496,456bp for the controls. Outliers has been removed from the matrices if the number of the contacts surpassed by 20 times the top 5‰ threshold of the number of contacts between restriction fragment pairs.

The accession number for the sequences reported in this paper is BioProject: XXX. http://www.ncbi.nlm.nih.gov/bioproject/XXX

Statistical analysis

Construction of contact histogram. Cumulative histograms were generated from contact maps for the different bin sizes (**Fig S4**).

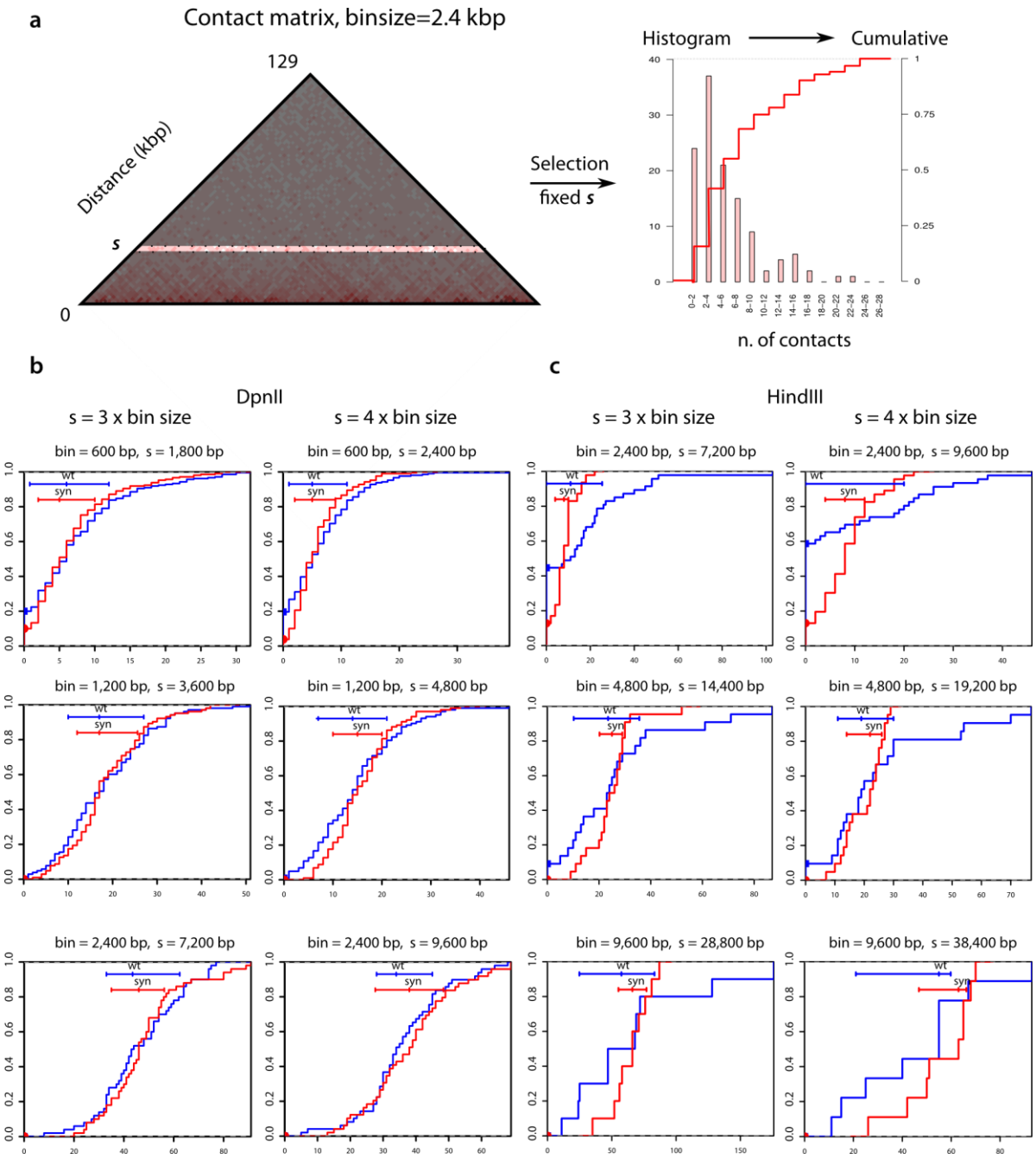


Figure S4. Generation of contact histograms. (**a**) Contacts made at fixed distance *s* are positioned along a diagonal (dark mask) that runs parallel to the main diagonal (triangle base). The histograms and the cumulative histogram as a function of *s* were then computed. The cumulative histograms were then computed for several values of *s*, here we report cumulative distributions of contacts as a function of s and bin size for DpnII (**b**) and HindIII (**c**) Hi-C contact maps. Histograms for selected values of bin sizes and distances *s* has been

reported in **Fig 2** in the main text. Blue line: native region. Red line: synthetic region. For small bin sizes (*s* = 600bp), the distribution of contacts of the redesigned region appeared systematically narrower than for the native region, with most bins being "visible", i.e. containing at least one read. The gain in resolution somehow fades away for the frequent cutter when the bin size increases, but, interestingly, the visibility of the bins remains nevertheless systematically better. For HindIII, the gain in resolution is always considerably better in the redesigned vs. the native sequence.

Q*uality improvement:* the CV is defined as the ratio between the standard deviation and the mean of the contact histograms at fixed distance *s*; to take into account the finite-size effect, we discarded bins at the edge of the contact matrix in order to keep the statistics (number of bins) for different values of *s* constant, up to *s* < 15,000 bp in DpnII and *s* < 70,000 bp in HindIII datasets. **Fig 2** show a consistent improvement in the CV within the redesigned SynIV-3C compared to the native counterpart. In order to show that the improvement is specific to the new restriction pattern and is unlikely to be find spontaneously within the genome, we compared the SynIV-3C results with seven regions of similar size (see **Supplementary note 4**) along chromosome 4. The quality improvement was assessed by computing the logarithm of the ratio of the CVs of the SynIV-3C and native region (**Fig S5**).

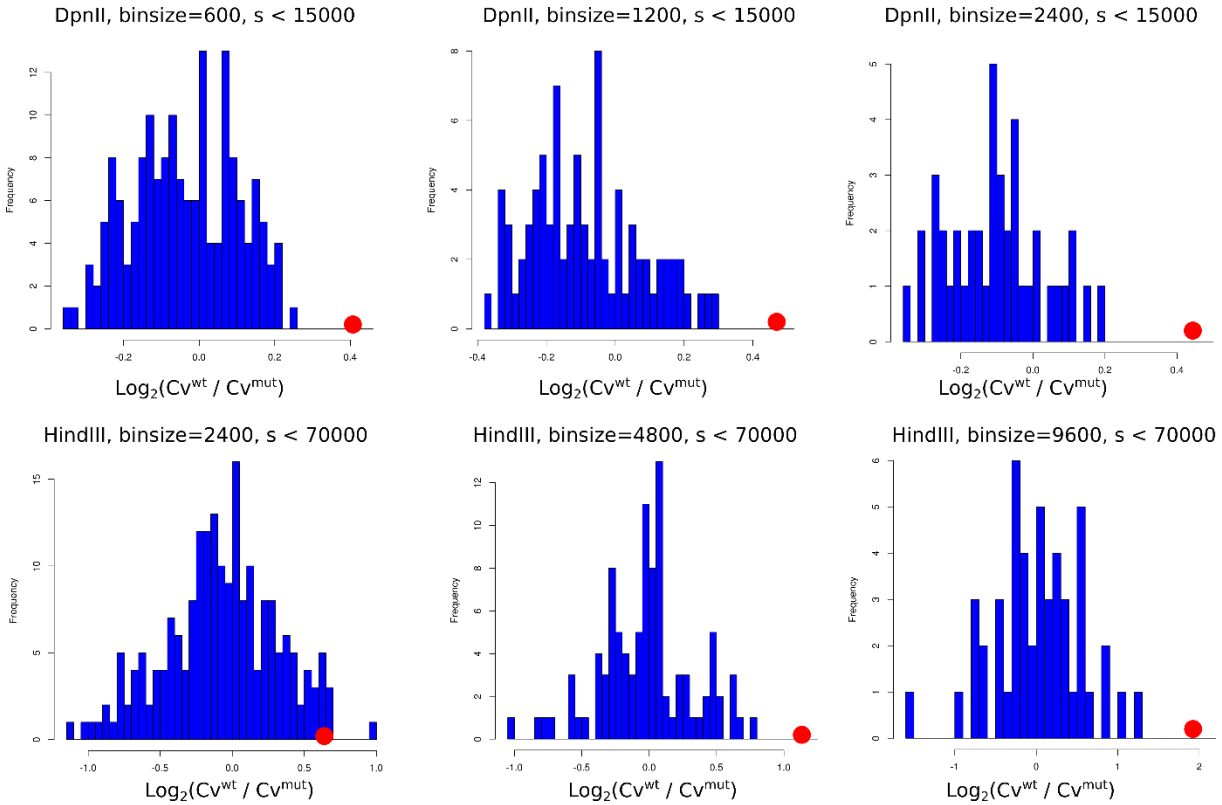


Figure S5. The histogram of ratio of the CVs between SynIV-3C and WT strain, for all values of distance *s* (for *s* < 15,000 in DpnII and *s* < 70,000 in HindIII datasets) in the control regions (blue bars) compared to the mean over *s* of this value in the synthetic region (red dot). Only the synthetic region show a quantitative improvement of statistical significance.

References

Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, Richardson SM, Dymond JS, Kuang Z, Scheifele LZ, Cooper EM, Cai Y, Zeller K, Agmon N, Han JS, Hadjithomas M, Tullman J, Caravelli K, Cirelli K, Guo Z, London V, et al (2014) Total Synthesis of a Functional Designer Eukaryotic Chromosome. *Science* **344:** 55–58

Cournac A, Marbouty M, Mozziconacci J & Koszul R (2015) Generation and Analysis of Chromosomal Contact Maps of Yeast Species. *Methods Mol. Biol.*

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

Dekker J, Rippe K, Dekker M & Kleckner N (2002) Capturing chromosome conformation. *Science* **295:** 1306–1311

Dymond JS, Richardson SM, Coombes CE, Babatz T, Müller H, Annaluru N, Blake WJ, Schwerzmann JW, Dai J, Lindstrom DL, Boeke AC, Gottschling DE, Chandrasegaran S, Bader JS & Boeke JD (2011) Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* **477:** 471–476

Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N & Rando OJ (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162:** 108–119

Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES & Dekker J (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326:** 289–293

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJT, van Oudenaarden A, Barton DBH, Bailes E, Nguyen AN, Jones M, Quail MA, et al (2009) Population genomics of domestic and wild yeasts. *Nature* **458:** 337–341

Marbouty M, Ermont C, Dujon B, Richard G-F & Koszul R (2014) Purification of G1 daughter cells from different Saccharomycetes species through an optimized centrifugal elutriation procedure. *Yeast* **31:** 159–166

Muller H, Annaluru N, Schwerzmann JW, Richardson SM, Dymond JS, Cooper EM, Bader JS, Boeke JD & Chandrasegaran S (2012) Assembling large DNA segments in yeast. *Methods Mol. Biol. Clifton NJ* **852:** 133–150

Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, Tischfield SE, Zhu X, Neale MJ, Jasin M, Socci ND, Hochwagen A & Keeney S (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* **144:** 719–731

Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, Conway A, Lockhart DJ, Davis RW, Brewer BJ & Fangman WL (2001) Replication dynamics of the yeast genome. *Science* **294:** 115–121

Rhee HS & Pugh BF (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483:** 295–301

Siow CC, Nieduszynska SR, Müller CA & Nieduszynski CA (2012) OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Res.* **40:** D682–D686