Filling annotation gaps in yeast genomes using genome-wide contact maps.

Hervé Marie-Nelly, Martial Marbouty, Axel Cournac, Gianni Liti, Gilles Fischer, Christophe Zimmer, Romain Koszul

Original paper

Filling annotation gaps in yeast genomes using genome-wide contact maps

Herve Marie-Nelly[1,2,3,4,8], Martial Marbouty[1,2,8], Axel Cournac[1,2], Gianni Liti[5], Gilles Fischer[6,7], Christophe Zimmer[3,4] & Romain Koszul[1,2,*]

[1]Institut Pasteur, Groupe Régulation Spatiale des Génomes, 75015 Paris, France

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

[3]Institut Pasteur, Unité Imagerie et Modélisation, 75015 Paris, France

[4]CNRS, URA 2582, 75015 Paris, France

[5]Institute for Research on Cancer and Ageing of Nice (IRCAN), CNRS UMR 7284 - INSERM U108, Université de Nice Sophia Antipolis, 06107 Nice, France

[6]CNRS, UMR7238, Biologie Computationnelle et Quantitative, F-75005, Paris, France

[7]Sorbonne Universités, UPMC Univ Paris 06, UMR7238, Biologie Computationnelle et Quantitative, F-75005, Paris, France

[8]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

[*]ABSTRACT

Motivations: *De novo* sequencing of genomes is followed by annotation analyses aiming at identifying functional genomic features such as genes, non-coding RNAs or regulatory sequences, taking advantage of diverse datasets. These steps sometimes fail at detecting non-coding functional sequences: for example, origins of replication, centromeres and rDNA positions have proven difficult to annotate with high confidence. Here, we demonstrate an unconventional application of Chromosome Conformation Capture (3C) technique, which typically aims at deciphering the average three-dimensional organization of genomes, by showing how functional information about the sequence can be extracted solely from the chromosome contact map.

Results: Specifically, we describe a combined experimental and bioinformatic procedure that determines the genomic positions of centromeres and ribosomal DNA clusters in yeasts, including species where classical computational approaches fail. For instance, we determined the centromere positions in *Naumovozyma castellii*, where these coordinates could not be obtained previously. Although computed centromere positions were characterized by conserved synteny with neighboring species, no consensus sequences could be found, suggesting that centromeric binding proteins or mechanisms have significantly diverged. We also used our approach to refine centromere positions in *Kuraishia capsulata*, and to identify rDNA positions in *Debaryomyces hansenii*. Our study demonstrates how 3C data can be used to complete the functional annotation of eukaryotic genomes.

Availability and Implementation: The source code is provided in the supplementary material. This includes a zipped file with the Python code and a contact matrix of *S. cerevisiae*.

[*]To whom correspondence should be addressed.

1 INTRODUCTION

De novo sequencing of genomes is typically followed by analyses aiming to identifying functional genomic features such as genes, non-coding RNAs or regulatory sequences. This important so-called annotation step raises non-trivial questions, and led to the development of complex bioinformatics approaches taking advantage of multiple datasets. For instance, transcriptome analysis is conveniently used to annotate expressed coding sequences (Grabherr *et al.*, 2011; Saha *et al.*, 2002) and synteny conservation between related species can reveal or confirm the presence of regulatory elements (Gordon *et al.*, 2011; Kellis *et al.*, 2004). Complementary to automated annotation through comparative approaches, experimental approaches such as ChIP-seq or MNase-seq have been conveniently used to map epigenetic marks, replication origins, or other functional elements of the genome (Roy *et al.*, 2010; Wang *et al.*, 2012).

However, such tools are sometimes unable to detect non-coding functional sequences: for example, origins of replication, centromeres and rDNA positions have sometimes proven difficult to annotate with a high degree of confidence in genomes. A compelling example is the failure of comparative genomics to identify the centromeres of the hemiascomycetes species *Naumovozyma castellii* through comparative genomics (Gordon *et al.*, 2011). As another example, the number of rDNA clusters in the genome of *Debaryomyces hansenii* is not known precisely, but only estimated

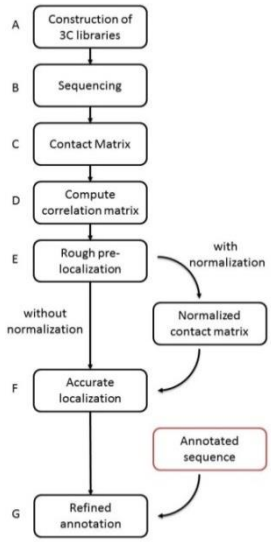
to lie between one and three and not currently annotated in the genomic sequence (Dujon *et al.*, 2004; Jacques *et al.*, 2010).

Genomic chromosome conformation capture (3C) assays measure the physical contact frequencies between DNA sequences (Dekker *et al.*, 2002; Lieberman-Aiden, Berkum, *et al.*, 2009; Duan *et al.*, 2010), providing important insights into both genomic organization and topological changes of chromatin domains that accompany cell differentiation or development. 3C data are typically analyzed in light of epigenetic marks and other genomic annotations. In an alternative application, 3D contact can be interpreted as an indicator of co-linearity of two DNA segments, and were recently used to improve the scaffolding of the human genome (Kaplan and Dekker, 2013; Burton *et al.*, 2013). Here, we use genome-wide 3C data to unveil functional elements of eukaryotic genomes that escape comparative genomic analysis. Specifically, we take advantage of nuclear architecture features to precisely determine the positions of centromeres in the yeast species *N. castellii* (Cliften *et al.*, 2006). We show that this bioinformatic approach discriminates ambiguous results from bioinformatics analysis, such as in *K. capsulata*. Finally, it also allowed us to complete the annotation of rDNA in the *D. hansenii* genome and confirm its centromere annotation (Lynch *et al.*, 2010).

2 METHODS

Centromeres and rDNA clusters lead to characteristic architectural features in the yeast nucleus. Taking advantage of these features, we developed a robust approach to characterize computationally centromere and rDNA positions from 3C data.

Yeast centromeres are tethered near a pole of the nucleus via microtubules attached to the microtubule organizing center (MTOC, or Spindle Pole Body in yeast), leading to centromere clustering. In the budding yeast *Saccharomyces cerevisiae*, this clustering causes distinct peaks of inter-



chromosomal contact frequencies in the raw genome-wide contact matrix, reflecting the convergence of 32 chromosomal arms towards a discrete region of the nucleus (Duan *et al.*, 2010).

Fig.1. Experimental and computational workflow.

We developed an algorithm that automatically recognizes these specific contact enrichments and estimates the genomic coordinates of centromeres. Centromeric positions are therefore determined based on their core biological function, rather than by sequence motif recognition, as is usually done. This approach can discriminate between multiple candidate positions obtained by sequence analysis.

Ribosomal DNA is organized as a cluster of repeats in the genome of all eukaryotes sequenced so far. These rDNA repeats give rise to the nucleolar compartment(s), which in *S. cerevisiae*, and other species, consists in a single subnuclear volume located opposite the SPB. This organization, combined with the large size of the rDNA cluster, creates an apparent intra-chromosomal barrier within the contact matrix of the chromosome carrying the rDNA locus. The position of a rDNA cluster in a genome is therefore easily identifiable, even in the absence of any annotation or sequence in the reference sequence. We developed an algorithm to identify the presence of rDNA clusters in these genomes.

The flowchart in Figure 1 provides an overview of the experimental and computational workflow, each of which will be described in a distinct subsection below.

2.1 Generation of genome-wide chromosome contact frequency matrices

3C libraries of the yeast species *S. cerevisiae* (BY4741), *N. castellii* (CBS4309), *D. hansenii* (CBS767), and *K. capsulata* (CBS1993) were generated from log-phase cells growing in YPD medium and as previously described (Dekker *et al.*, 2002; Oza *et al.*, 2009), but using a frequently cutter enzyme (DpnII) as in (Sexton *et al.*, 2012). Briefly, the cells were cross-linked with 1% formaldehyde, and resuspended in DpnII restriction buffer, which were subsequently processed into Illumina libraries. 3C libraries were sheared and resulting fragments with sizes between 400 and 800 bp were sequenced using 100 bp pair-end sequencing on an Illumina HiSeq2000. The raw data from all 3C-seq experiments was then processed as follow: first, short reads were mapped on the genomes of *S. cerevisiae* (GCF_000146045.1), *N. castellii* (GCF_000237345.1), *D. hansenii* (GCF_000006445.1), and *K. capsulata* (Morales *et al*., in revision) using bowtie 2 in local and very sensitive modes (Langmead and Salzberg, 2012). Only pairs of reads with a Mapping Quality above 30 were retained, and contact reads mapping on the same fragment were discarded (Cournac *et al.*, 2012). PCR duplicates were also removed. All rejected reads (except PCR duplicates) were included in a pool of "leftover reads". These filtering steps generally remove 15 to 20% of the initial set of raw reads. After alignment of individual reads on the reference genome, we built a 2D histogram, where the value of each 2D bin (pixel) indicates how many reads fall into the corresponding pair of genomic segments. The genomic partition defining these segments was based on the restriction enzyme cutting sites, rather than on constant genomic intervals. For the *S. cerevisiae*, the DpnII restriction enzyme leads to a contact matrix M_0 of size $m_0 \times m_0$, with $m_0 = 35914$. At this resolution, however, the contact matrix is very sparse, and hence noisy. The signal-to-noise ratio can be improved at the expense of genomic resolution by binning the reads into larger genomic intervals. We therefore considered three additional matrices, M_k *(k =1,2,3)* obtained by summing non-overlapping blocks of $3^k \times 3^k$ pixels. For *S. cerevisiae*, these matrices have genomic bins of $R_1 = 1{,}233 \pm 1{,}095$bp (mean ± standard deviation of 3 restriction fragment [RF]), $R_2 = 3{,}696 \pm 1{,}919$bp (9 RF) and $R_3 = 11{,}034 \pm 3{,}455$bp (27 RF), and size $m_1 = 9{,}712$, $m_2 = 3{,}240$ and $m_3 = 1{,}086$, respectively (Figure 2A-C, respectively).

For small genomic bin sizes *R*, the limited signal-to-noise ratio of these matrices can complicate the identification of contact frequency enrichments. Computing a correlation matrix, as initially done in (Lieberman-Aiden, van Berkum, *et al*., 2009), allows to strongly increase the contrast of contact patterns. We then computed a new matrix *C* from *M*, where *C(i,j)* is defined as the Pearson correlation coefficient of the rows *i* and *j* of *M*:

$$C(i,j) = \frac{\sum_{l=1}^{m} \left(M(i,l) - \overline{M}(i)\right)\left(M(j,l) - \overline{M}(j)\right)}{\sqrt{\sum_{l=1}^{m} \left(M(i,l) - \overline{M}(i)\right)^2} \sqrt{\sum_{l=1}^{m} \left(M(j,l) - \overline{M}(j)\right)^2}} \quad (1)$$

And

$$\overline{M}(i) = \frac{1}{m} \sum_{l=1}^{m} M(i,l)$$

is the average value of row *i* of matrix *M*. Note that in order to remove the chromosome length bias, the correlation was computed from the interchromosomal parts of the matrix (Figure 2D).

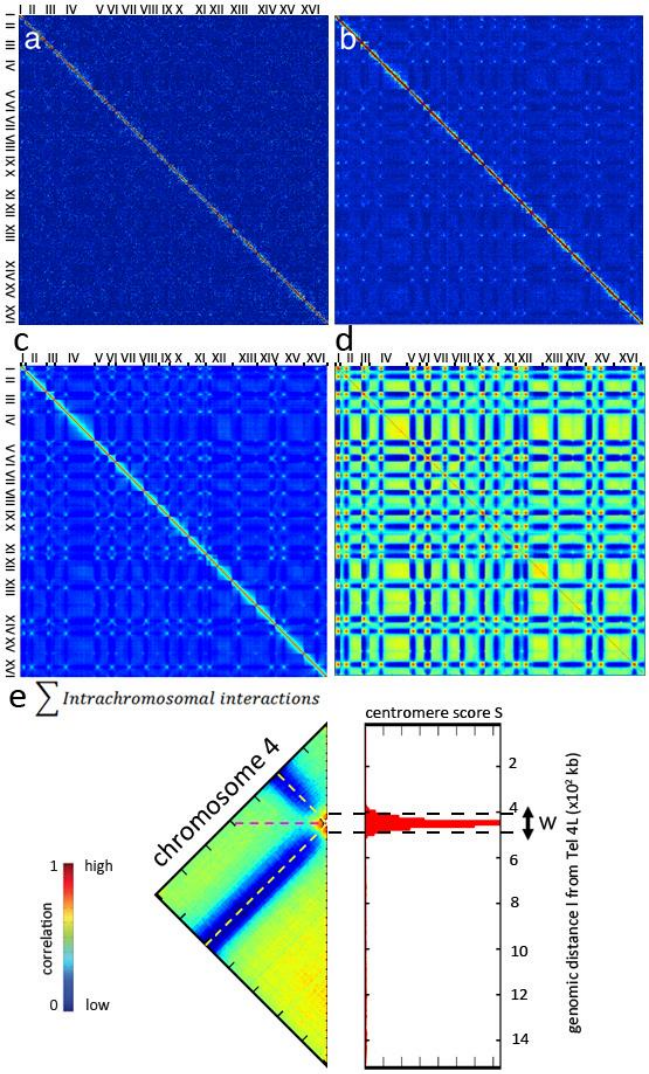


Fig. 2. Contact frequency matrices M_1, M_2 and M_3, for *S. cerevisiae*, at three levels of genomic resolution: (**A**) R_1 = 1,233 bp (3RF), (**B**) R_2 = 3,696 bp (9 RF), (**C**) R_3 = 11,034 bp (27 RF). The 16 chromosomes of *S. cerevisiae* are labeled from I to XVI. The strong diagonal is due to intrachromosomal contacts. Note the peaks corresponding to contacts between centromeres from different chromosomes. (**D**) Correlation matrix for *S. cerevisiae*: each element of the matrix is the Pearson coefficient between the vectors i and j of the matrix of contacts (bin size of 3,696 kb). (**E**) Zoom on the intra-chromosomal correlation map of chromosome 4. The centromere score *S(l)* for each bin *l* is plotted along the sub-matrix (scale bar = 100kb). The peak of this distribution defines the center of a 40kb window *w* likely to contain the centromere.

2.2 Rough pre-localization of centromeric regions from *cis* contacts

In the correlation matrix *C*, the blocks corresponding to intrachromosomal (*cis*) contacts within pericentromeric regions exhibit a characteristic "cross" shape pattern (see diagonal in Figure 2D and 2E). This pattern can be explained by the clustering of centromeres near the spindle pole body (SPB) and the polymer brush-like organization of chromosomes in this region, whereby the two chromosome arms are stretched out away from the SPB (Wong *et al.*, 2012). As a result, the centromere is sequestered away from other loci along the chromosome, leading to a depletion of contacts along the yellow dotted lines in Figure 2E, while loci on opposite arms located at similar genomic distances from the centromeres tend to be in proximity, leading to contact enrichments along the "anti-diagonal" (pink dotted line in Figure 2E). We took advantage of this pattern for the automated identification of centromeres by defining a "centromere score" as:

$$S(l) = \frac{\frac{1}{2l-1} \sum_{i=1}^{2l-1} C(l, 2l-i)}{\frac{1}{p} \sum_{j=1}^{p} C(l,j)} \quad \text{for} \quad l = 1,2,\ldots E\left(\frac{p+1}{2}\right) \quad (2) \quad \text{and}$$

$$S(l) = \frac{\frac{1}{2(p-l)+1} \sum_{i=2l-p}^{p} C(l, 2l-i)}{\frac{1}{p} \sum_{j=1}^{p} C(l,j)} \quad \text{for} \quad l = E\left(\frac{p+1}{2}\right)+1,\ldots,p-1,p \quad (3)$$

where *p* is the number of rows of the submatrix and *E(x)* denotes the largest integer $\leq x$. Thus, for each genomic bin *l*, *S(l)* is the ratio of the average correlation along the anti-diagonal passing through *C(l,l)* and the average correlation along the row *l* of *C*. The 'centromere score' *S(l)* is expected to be largest for *l* near the actual position of the centromere (Figure 2E). Note that for acrocentric chromosomes, the peak of *S* can differ significantly from the true centromere position. Therefore, for each chromosome *k*, we used the location of this maximum, $l_0 = \arg\max S(l)$ to define a genomic interval

$$[i_L(k); i_R(k)] = [l_0 - 20\text{kb}; l_0 + 20\text{kb}] \quad (4)$$

along the chromosome that we expect to contain the centromere. The size of the interval is arbitrary and depends on the size of the chromosome: it must be kept within the chromosome boundaries, and must be large enough that the Gaussian fit can be applied correctly (see below). For *S. cerevisiae* we used window sizes of 40kb. A more accurate localization of the centromere is performed in the next step, as described below.

2.3 Refined estimation of centromere position from *trans* contacts

In principle, the position of a given centromere could be obtained using only the *cis* contact submatrix for the corresponding chromosome, or alternatively using only the *trans* contact submatrix involving one other chromosome. However, since contact matrices are histograms obtained

from a limited number of reads, they are subject to Poisson noise, which imposes a fundamental limit to the localization accuracy (much as in single molecule localization, see e.g. 18). For improved localization accuracy, we therefore took advantage of the redundancy provided by the N_{chr}-1 distinct *trans* contact patterns available for each of the N_{chr} centromeres (N_{chr} = 16 for *S cerevisiae*). This approach was applied both on the raw contact matrix *M* and on the normalized matrix *N*, obtained as described in (Cournac *et al.*, 2012). Briefly, this normalization step aims to correct for experimental biases affecting the transformation from ligation product counts into contact frequencies (a different approach as the procedure described in (Yaffe and Tanay, 2011). The procedure employs an iterative algorithm that enforces all rows and columns to have unit sum, i.e. it ensures that $\sum_{i=1}^{m} N(i,j) = 1$ for all $j = 1..m$ and $\sum_{j=1}^{m} N(i,j) = 1$ for all $i = 1..m$, where *m* is the size of the matrix. For details, see (Cournac *et al.*, 2012), or (Imakaev *et al.*, 2012) for a related approach. The normalization has the overall effect to increase the contrast of the contact data, and to attenuate noise in the raw data (Figure 3A).

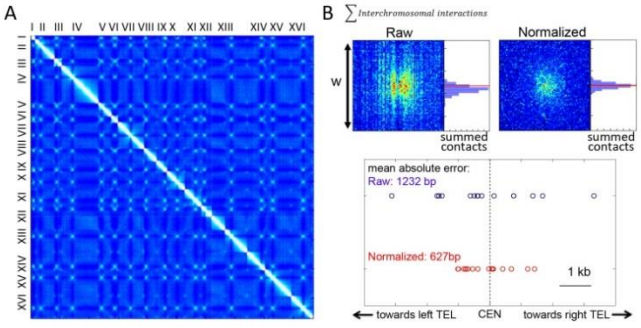


Fig. 3. (**A**) Normalized contact frequency matrix N_1 for *S cerevisae*. (**B**) Summed *trans*-contact matrices corresponding to the submatrix of size *w* (Figure 2E) for chromosome 2. On the right side of each sub-matrix we plot the distribution of the centromere localizations obtained using bootstrapping. The true centromere position is indicated with a red line. The diagram below represents the distribution of the 16 centromere positions as estimated from raw and normalized data (blue and red circles, respectively). x-axis: distances along the chromosome, centered on the position of the centromere (scale bar = 1kb).

Specifically, for each chromosome, we carved out N_{chr}-1 submatrices of size 40 kb x 40 kb corresponding to *trans* contacts and defined by the $[k_0 - 20\text{kb}; k_0 + 20\text{kb}]$ intervals obtained above (if necessary, the size of this matrix was reduced to that of the smallest interval, such that all submatrices had the same size). Note that in computing the superposed matrix we did not use the intrachromosomal contact data because of the bias for acrocentric chromosomes mentioned above. These submatrices were then summed, yielding a single "superposed" contact matrix A_k (for the centromere of chromosome *k*; Figure 3B):

$$A_k = \sum_{\substack{l \in [1, N_{chr}] \\ l \neq k}} M\left(i_L(l)...i_R(l), i_L(k)...i_R(k)\right) \quad (5)$$

For normalized data, *M* is simply replaced by *N* (Figure 3B). The next step consists in projecting this summed contact matrix into a 1D profile $F_k(i) = \sum_{j=1}^{p} A(i,j)$

As apparent from Figure 3B**,** normalization typically produces a less noisy profile, allowing more accurate identification of the centromere-related peak. Finally, in order to accurately estimate the centromere position, we implemented a Gaussian fitting procedure similar to that commonly used for single molecule localization (Ober *et al.*, 2004). Specifically, we used an iterative algorithm that aims to minimize the mean squared difference:

$$H(a,b,i_c,\sigma) = \sum_i \left[F_k(i) - G(i;a,b,i_c,\sigma)\right]^2 \quad (6)$$

between *F* and the Gaussian function:

$$G(i;a,b,i_c,\sigma) = a\exp\left(-\frac{(i-i_c)^2}{2\sigma^2}\right) + b \quad (7)$$

where *a*, *b*, i_c and σ are the parameters to be fitted, i.e. we seek:

$$\left(\hat{a},\hat{b},\hat{\imath}_c,\hat{\sigma}\right) = \arg\min H(a,b,i_c,\sigma) \quad (8)$$

Thus the final estimated position of the centromere for chromosome *k* is given by $\hat{\imath}_c$.

Application of this procedure to our normalized *S. cerevisiae* contact data and comparison with the genomic annotation revealed that the centromeres could be localized with a mean absolute error of only 627 bp (1232 bp without normalization) - demonstrating that this functionally important locus can be accurately located from the contact data alone (Figure 3B).

2.4 Confidence intervals and effect of coverage and normalization on localization accuracy

In order to provide a robust roadmap for future studies, we next quantified centromere localization accuracy and how it is affected by coverage (i.e. sequencing depth), binning, and the normalization procedure.

We used a bootstrapping approach to estimate confidence intervals of the computed centromere localization and to examine the influence of coverage. Specifically, we simulated many contact frequency matrices with an expected total number of reads either equal to, or smaller than the experimentally obtained matrix *M* (which for *S. cerevisiae* totals $N_{reads,Sc}$ = 21,457,086). To do this, we generated N_{bs} = 500 contact matrices $M_{bs,k}$, $k = 1..N_{bs}$ where $M_{bs,k}(i,j)$ is a random integer value drawn from a Poisson distribution of density $\lambda(i,j) = fM(i,j)$, where $f \leq 1$ indicates the coverage relative to the original matrix. Thus the expected total number of contacts in $M_{bs,k}$ is fN_{reads}. We then used each of the random contact matrices $M_{bs,k}$, to compute an independent estimate of the centromere positions.

Centromere position confidence interval. For *f = 1*, the distribution of these estimates provides a measure of the uncertainty with which the centromere positions have been determined from the original contact data. We compared the distribution of localization errors for the 16 centromeres of *S. cerevisiae* to the normal distribution of mean 0 and variance given by the bootstrap samples. The two distributions cannot be distinguished by a Kolmogorov-Smirnov test (p=0.12; Figure 4A). This suggests that the confidence intervals determined by the bootstrap estimates correctly reflect actual localization uncertainties.

Effect of coverage, normalization and binning. To examine the effect of coverage (or sequencing depth), which determines the total number of reads involved in contacts in the matrix, we applied the bootstrapping method to a range of *f* smaller than 1 (*i.e.* to experimental matrices where contact events have been down-sampled), specifically: *f = 0.8, 0.6, 0.4, 0.2, 0.1, 0.01, 0.001*. For each value of *f*, we computed the centromere position error from the N_{bs} samples (relative to the ground truth) and the mean over the 16 centromeres and the N_{bs} samples. Figure 4B plots this mean error as function of the mean number of reads in the bootstrapped samples (fN_{reads}). As expected, the localization accuracy generally improves with coverage, provided that the contact data are binned at adequate genomic resolution and that the qualities of the libraries are equivalent. Also, normalization improves localization accuracy for high coverage ($N_{reads} > 2.10^6$), but gives much poorer results for low coverage, where the raw data

should be preferred and provide more consistent accuracy. This result underlines the complexity of 3C contact data analyses and the need to take into account sequencing depth, binning, and bias correction. However, it also shows that a ~1M reads contact map is sufficient to identify with high accuracy centromeric positions. The graph provides a tool to determine the likely optimal choice of binning and normalization options for the DpnII enzyme applied to a yeast genome. Using bins smaller than ~3kb does not significantly affect localization accuracy of normalized data (Figure 4B).

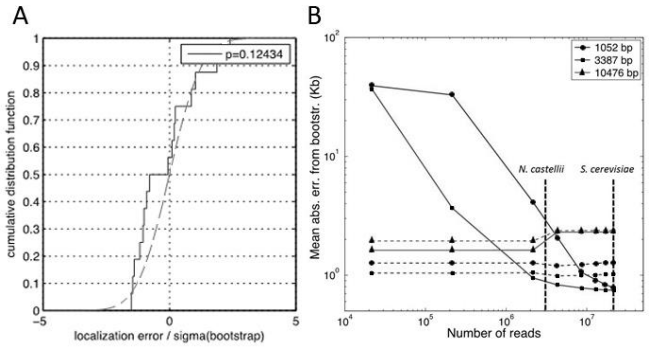


Fig.4. (**A**) The distribution of localization errors for the 16 *S. cerevisiae* chromosomes normalized by the standard deviation of corresponding bootstrap estimates (cumulative distribution, solid curve) is consistent with a normal distribution (dotted curve). (**B)** Effect of coverage, normalization and binning on localization accuracy. The mean absolute localization error for the 16 *S. cerevisiae* centromeres is plotted as function of the number of reads for normalized (solid curves) and raw contact data (dashed curves) and for three resolutions (bin size median indicated in legend).

2.5 Identification of rDNA loci in chromosome contact matrices

We next proceed to show that contact matrixes can also allow the characterization of ribosomal gene clusters. First, a contact matrix of *S. cerevisiae* was generated where the bins containing the two rDNA repeats of the reference genome were removed (chromosome 12 region comprised between 450,000 and 470,000 of the reference genome was removed, encompassing the two rDNA units positioned between coordinates 451575 and 468931 and reflecting the 100 – 200 repeats of total rDNA). The pair-end reads were remapped on this modified genome (including the mitochondrial DNA; bin size = 10kb). We then selected in the pool of "leftover reads" all the pairs where one mate maps unambiguously on the genome (mapping quality above 30), while the other mate does not eliminating reads containing unknown bases (N). These unmapped sequences were blasted on a sequence dataset containing yeast ribosomal sequences retrieved from the NCBI server (parameters: blast2 -p blastn -e 2e-30) to keep only highly significant hits. The corresponding mates were then mapped along the genome divided into bins (Figure 5). The peak in the distribution was clearly apparent on chromosome 12 (~10,000 hits compare to an average of 20 along the rest of the genome). Zooming in the distribution along an unmodified genome (bin size = 10bp) clearly show that the peaks lie within chromosome 12 region comprised between 450,000 – 490,000 bp, adjacent to the position of the ribosomal gene cluster (Figure 5B; note the 5S variants lying between 470,000 and 490,000 positions of the cluster per se do not allow high quality mapping of the reads and appear also as a blank area along the genome).

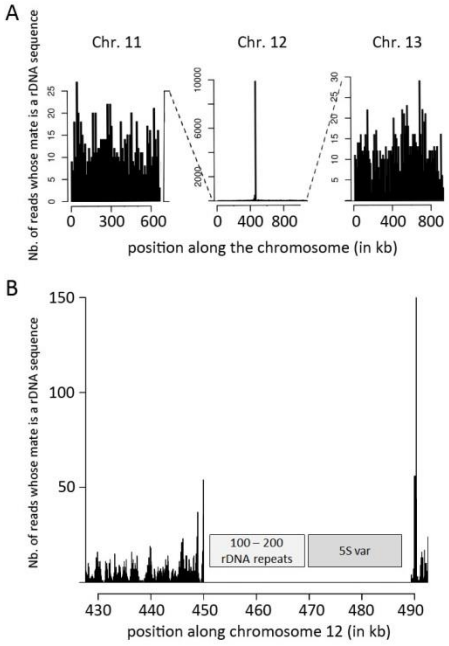


Fig. 5. Characterization of rDNA position in *S. cerevisiae* from contact matrix. (**A**) Distribution of the reads whose one mate maps unambiguously along the genome and the other maps in a rDNA sequence. The scale bar of the y-axis is different for chromosome 12 to adjust to the increase in contacts observed in the region around ~450kb. (**B**) Distribution of interactions a rDNA sequence and another sequence in the region of chromosome 12 around position 450,000 kb (bin size: 10bp). The genome used for the distribution encompasses the rDNA repeat cluster position (grey squares).

3 RESULTS

3.1 Discriminating true centromeres among computational predictions in *Kuraishia capsulata*

Centromeres of yeast species have mainly been characterized so far through computational analysis approaches aiming at identifying landmarks in the sequence likely to predict centromeres positions with high confidence level. A classical approach consists in searching for the consensus sequences specific to most *Saccharomycetaceae* species studied so far, and related to the point centromeres of the well-studied yeast *Saccharomyces cerevisiae* (Souciet *et al.*, 2009; Gordon *et al.*, 2011). These centromeres are very compact (125bp) and present a consensus sequence composed of three centromere DNA elements (CDEI, II, III; 20). CDE I and III present a strong consensus core region and flank CDEII which is characterized by a strong AT rich content but a high sequence variability (>90%). Not all yeast species exhibit such distinguishable point centromeres, and an alternative analysis searching for local composition-bias in low-GC content was recently described (Lynch *et al.*, 2010). Drop in GC content is likely to reflect a drop in recombination frequency, and coincide with the positions of centromeres experimentally characterized in *Yarrowia lypolitica* (Lynch *et al.*, 2010), as well as putative centromeric regions made of clusters of retrotransposon Tdh5 in *Debaryomyces hansenii*, for instance.

The genome of the nitrate assimilating yeast *K. capsulata* has been recently sequenced and assembled into seven chromosomes (Morales *et al.*, in press). A search for CDEI and III consensus

sequences failed to identify putative centromeres. However, a search for low-GC content regions as described in (Louis *et al.*, 2012; Lynch *et al.*, 2010) led to the characterization of nine putative centromeric regions (with chromosome four containing three; Morales et al., in press; Table 1). In order to confer an experimental validation of these results, and see if we could discriminate between ambiguous sequences, we performed a genomic 3C experiment on *K. capsulata* and sequenced the resulting library. The quality of the matrix was relatively poor despite an important coverage ($N_{reads,Kc}$ = 16,446,227; Figure 6A), as seen by the "flatness", or lack of contrasts, of the matrix and as quantified by the ratio between mitochondrial and genomic DNA interactions (AC and RK, personal communication). Despite the apparent noise, each chromosome still exhibits a discrete region presenting a strong enrichment in interactions with the corresponding other chromosomal regions, similar to centromeric behavior in *S. cerevisiae* (pink arrowheads, Figure 6B). We followed the procedure described above and characterized for each of the seven chromosomes *cis*-contact matrixes the genomic intervals w containing centromeric regions. Given the low coverage in informative reads of the matrix, we opted for a binning of 2,191kb (9 RF) and assessed from the analysis above that little if any improvement would result from SCN normalization. We proceeded to superpose the *trans*-submatrices containing the centromeric regions defined from the *cis*-contacts (Figure 6B). A Gaussian fit was applied as described, and the coordinates of centromere positions along with the precision calculated (Figure 6B; Table 1). Quite remarkably, the regions identified experimentally through this approach overlapped exactly with those obtained after computational analysis for the six chromosomes exhibiting a single, unambiguous putative centromere position (Table 1). In addition, the region identified on chromosome four as the centromere overlapped with only one of the three putative positions identified from the composition bias analysis, allowing the annotation of this position as the true centromere. This analysis indicates that careful handling of a contact matrix can successfully back up computational annotation, experimentally confirming and disambiguating weak predictions.

3.2 Identification of centromeres in *Naumovozyma castellii*

We then turned to *N. castellii*, an organism in which centromeric regions remained elusive to date, despite thorough and repeated investigation using different computational approaches (Gordon *et al.*, 2011; Cliften *et al.*, 2006). We built a genomic 3C library of *N. castellii* CBS 4309 strain and generated the corresponding contact matrix (N_{chr} = 10; $N_{reads,Nc}$ = 3,265,947 contacts; Figure 6C). Following through the procedure described above, we characterized for each of the ten chromosomes cis-contact matrixes the genomic intervals containing centromeric regions (Figure 6D). From the *S. cerevisiae* analysis, we estimated that the optimal binning for a 3M reads raw contact matrix to 3,696 kb bins (9 RF) and that normalization through SCN was likely to improve the results (Figure 4B). Therefore, we generated this matrix and proceeded to superpose the *trans*-submatrices containing the centromeric regions defined from the *cis*-contacts. The Gaussian fit was applied as described, and the coordinates of centromere positions along with the precision calculated (Table 1).

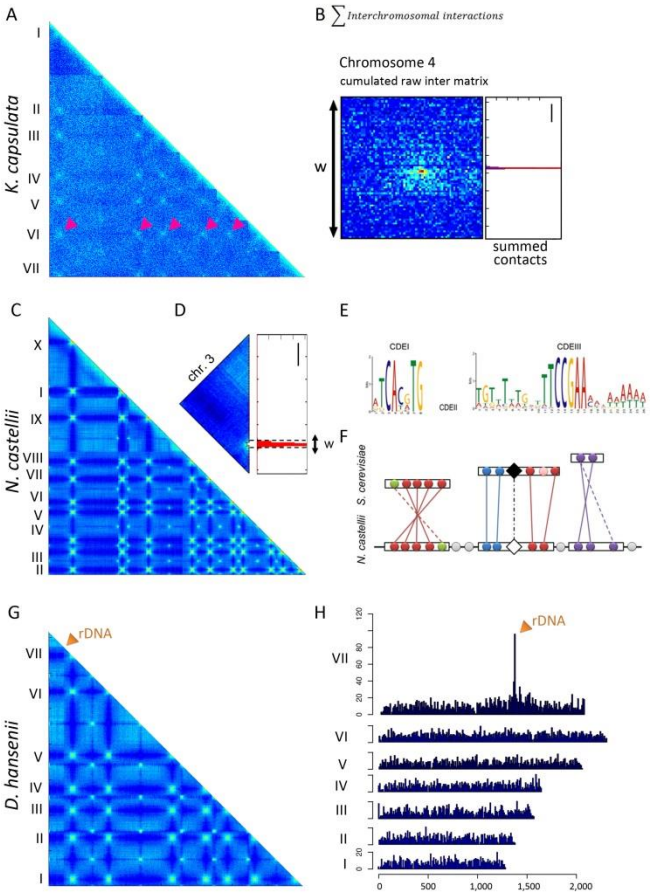


Fig. 6. Identification of genomic features in three different species. (**A**) Correlation matrix of K. capsulata (seven chromosomes). The pink arrowheads indicate some of the punctual inter-chromosomal contacts observable in the matrix. (**B**) Left: zoom on cumulated inter-chromosomal contacts of chromosome four around the region w likely to contain the centromere. Right: distribution of the sum of interchromosomal contacts along chromosome four (in blue). The red line indicates the peak of the distribution and the region containing the centromere (scale bar = 10kb). (**C**) Correlation matrix of *N. castellii* (10 chromosomes). (**D**) Zoom on intra-chromosomal contacts of chromosome 3. The peak of this distribution defines the center of a 20kb window *w* likely to contain the centromere (scale bar = 100kb). (**E**) Identification of CDEI and CDEIII consensus sequences in a mix of intergenic regions from *Lachancea* centromeric regions and *N. castellii* centromere sequences predicted from the 3C data. The signal identified corresponds only to *Lachancea* sequences. (**F**) Schematic representation of synteny conservation between a centromeric region of *N. castellii* (bottom line) and *S. cerevisiae* (three upper blocks). Grey circles: genes between syntenic blocks (in black rectangles). Full colored circles: conserved genes. Black diamonds: known centromere. Empty diamond: predicted centromere. (**G**) Correlation matrix of *D. hansenii* (seven chromosomes). The centromere position can be assessed from the matrix by the "break" it generates in the matrix. (**H**) Distribution along the genome of reads with one pair mate maps with a good quality score and the other does not but correspond to ribosomal DNA. The peak observed on chromosome seven marks the position of the rDNA cluster.

N. castellii, although positioned within a clade of species that present the CDE consensus sequences characteristic of point centromeres, is an intriguing exception in this regards. We hypothe-

sized that CDE sequences may have escaped from former investigations because of important divergence of the consensus sequence, and performed a computational analysis focusing on the centromeric region identified through genomic 3C. First, the intergenic sequences of the coding DNA sequences (CDS) found within these coordinates were recovered and submitted to the motif finder algorithm MEME (Bailey and Elkan, 1994) under the zoops (zero or one motif per sequence) or the oops (only one motif per sequence) modes. No significant motif could be identified from these first analyses. In order to guide the motif finder program, intergenic sequences from *N. castellii* were included into a set of 63 intergenic regions known to contain centromeres from 8 other yeast species from the *Lachancea* clade. These regions were used as a validation of the CDE I and III detection approach. MEME was then able to identify CDEI and CDEIII consensus sequences (Figure 6E) but all of these motifs corresponded to centromere regions of *Lachancea* species, whereas no CDEI and only a very weak CDEIII signal was observed for *N. castellii* regions (and no signature of a CDEII region was found upstream of CDEIII).

For an independent verification, and to test whether the centromeric regions of *N. castellii* have retained their ancestral positions, we analyzed the synteny conservation of pericentromeric regions between *N. castelii* and two related species, *S. cerevisiae* and *Zygosaccharomyces rouxii*. To do so, we defined synteny blocks between pericentromeric regions, encompassing 10 protein-coding genes in the two related species and the genome of *N. castellii*. We then looked if the coordinates of the conserved synteny blocks in the genome of *N. castellii* overlapped the coordinates of the centromeric regions defined from the contact map (Figure 6B; Drillon *et al.*, 2014). Four out of 10 centromeres in *N. castellii* belonged to the first category. Of the remaining six, two additional centromeric regions in *N. castellii* were found to lie right next to a synteny breakpoint with the genome of either *S. cerevisiae* or *Z. rouxii*, and therefore were also compatible with ancestral centromeric locations. For the four remaining centromeric regions, we found that multiple rearrangements having occurred in these regions have hidden the evolutionary relatedness between these regions. In summary, at least six out of ten centromeric regions in *N. castellii* correspond to orthologous regions in other species that contain point centromeres. Therefore the majority of the centromeres in *N. castelii* have retained their ancestral positions since they diverged from their last common ancestor with *S. cerevisiae* and *Z. rouxii*. This demonstrates that if all ten consensus centromeric motifs have evolved beyond recognition in *N. castellii*, centromeres positions are conserved for at least 6 of them. It is likely that extending this approach to more closely related species with recognizable point centromeres will unveil more synteny links and increase this number.

Interestingly, Gordon et al. (2009) had sought without success for consensus centromere sequences at putative ancestral centromeric locations in *N. castellii*. Here, we show that the centromere function remains nevertheless linked to these ancestral positions for at least six out of ten chromosomes although CDE regions are not identifiable within these regions. This suggests that the centromeric binding proteins and/or the mechanisms involved have evolved significantly in this lineage. Interestingly, and perhaps not coincidently, RNA interference is also conserved in this species.

3.3 Identification of ribosomal DNA locus in *Debaryomyces hansenii*

The genome of *D. hansenii*, a cryotolerant and osmotolerant marine yeast important in the agro-food industry, lacks annotation of the ribosomal DNA locus and has centromeres predicted through computational analysis (Lynch *et al.*, 2010). We generated a genomewide contact matrix of the seven chromosomes ($N_{reads,Dh}$ = 7,020,925 contacts, bins of 3.2kb corresponding to 9 RF; Figure 6G). First, we confirmed the position of the centromeres that were predicted through a genomic computational approach (Table 1). We then search for ribosomal DNA locus (Material and Methods) and found a peak on chromosome G in the distribution of reads along the genome for which the other mate corresponds to rDNA (Figure 6H). By zooming in the distribution, the position of the ribosomal DNA cluster of *D. hansenii* was identified at 1,354,000bp (Figure 6H). This region corresponds to an intergenic region containing a pseudogene and a gap, according to the published reference genome (Deha2G::1,353,661-1,356,925, www.genolevure.org). This region was blasted on the NCBI database, revealing two small (75bp) regions matching with ribosomal DNA at positions 1,354,446 and 1,355,863. We therefore inferred the position of a large, unique ribosomal DNA cluster within this window on chromosome G, ruling out the hypothesis regarding the existence of three intrachromosomal clusters in this genome.

We then compared the chromosomal location of the rDNA between the genomes of *D. hansenii* and two other genomes for which rDNA location is known (*Pichia stipitis* and *Yarrowia lipolytica*; Dujon *et al.*, 2004; Jeffries *et al.*, 2007) using SynChro (Drillon *et al.*). No synteny conservation could be found between these 3 genomes. In addition, we checked if any rDNA annotation could be retrieved from the genomes of 11 species belonging to the CTG clade at the locus corresponding to *D. hansenii* rDNA using the CGOB database (Fitzpatrick *et al.*, 2010). No such information was present in the database. In conclusion, no indication of synteny conservation of the rDNA locus between *D. hansenii* and other yeast species could be identified, consistent with the hypothesis that rDNA is mobile in the *Candida* clade (Proux-Wera *et al.*, 2013).

Overall, we showed that genome-wide chromosome conformation capture can be used to unveil important functional elements that sometimes escape standard genomic analyses. After validating the procedure on the well-known yeast *S. cerevisiae* genome, we successfully determined centromere positions in *Naumovozyma castellii*, where these coordinates could not be obtained previously. Although computed centromere positions were characterized by conserved synteny with neighboring species, no consensus sequences could be found, suggesting that centromeric binding proteins or mechanisms have significantly diverged. We also applied our approach to choosing among multiple predicted centromere positions in *K. capsulata*, and to identifying rDNA positions in *D. hansenii*. Thus, our study demonstrates how 3C data can be used to complete the functional annotation of eukaryotic genomes with a bioinformatic approach. The sequencing depth necessary to reach this goal does not have to be high (~3M reads proved largely sufficient for *N. castellii* and *S. cerevisiae*). It is likely that our standardized procedures will allow identifying other functional elements from contact data matrixes in the genome of microorganisms, and

potentially in metazoans. Combined with the recent application of 3C to genome assembly, this study confirms the helpfulness of tridimensional information to genomic analysis.

Table 1. Centromeres identification

Centromeres of *Kuraishia capsulata*				
#chr.	Predicted*		3C mean position	Precision (bp)
	Start	End		
1	466903	470996	470602	2426
2	1546884	1551323	1547201	6741
3	910200	913556	911375	3651
	469800	472200	X	
4	1033064	1035337	1034424	851
	476500	478800	X	
5	574146	576900	572890	2686
6	604123	607946	606172	3120
7*	1101261	1106974	1099909	3548

Centromeres of *Naumovozyma castellii*			
#chr.	3C mean position	Precision (bp)	Supported by synteny
1	1047129	681	YES
2	864103	570	YES
3	973309	489	-
4	535959	720	YES
5	576626	1102	-
6	206931	527	YES
7	591720	1055	YES
8	293288	718	-
9	376666	746	-
10	183626	922	YES

Centromeres of *Debaryomyces hansennii*				
#chr.	Tdh5 cluster position		3C mean position	Precision (bp)
	Start	End		
A	333868	350254	346614	7134
B	991276	1007003	1008332	1755
C	975133	993839	979218	4402
D	479577	494999	481966	1186
E	504719	523489	511437	1329
F	1543521	1560082	1548204	2039
G	940014	967691	949618	3330

ACKNOWLEDGEMENTS

We thank Julien Mozziconacci and the members of the RSG and IM labs for fruitful discussions. We thank Jean-Yves Coppee and Caroline Proux from the PF2 at the IP Genopole for technical help regarding sequencing.

Funding: 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] and by Agence Nationale de la Recherche [ANR-09-PIRI-0024] to C.Z. and R.K. H.M-N. is supported by a fellowship from Fondation pour la Recherche Médicale (FRM). MM is the recipient of an Association pour la Recherche sur le Cancer fellowship [20100600373].

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Burton,J.N. *et al.* (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.*, **31**, 1119–1125.

Cliften,P.F. *et al.* (2006) After the Duplication: Gene Loss and Adaptation in Saccharomyces Genomes. *Genetics*, **172**, 863–872.

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

Dekker,J. *et al.* (2002) Capturing Chromosome Conformation. *Science*, **295**, 1306–1311.

Drillon,G. *et al.* (2014) SynChro: a fast and easy tool to reconstruct and visualize Synteny blocks along eukaryotic Chromosomes. *PLoS One, in press*.

Duan,Z. *et al.* (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.

Dujon,B. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.

Fitzgerald-Hayes,M. *et al.* (1982) Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. *Cell*, **29**, 235–244.

Fitzpatrick,D.A. *et al.* (2010) Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser. *BMC Genomics*, **11**, 290.

Gordon,J.L. *et al.* (2009) Additions, Losses, and Rearrangements on the Evolutionary Route from a Reconstructed Ancestor to the Modern Saccharomyces cerevisiae Genome. *PLoS Genet*, **5**, e1000485.

Gordon,J.L. *et al.* (2011) Mechanisms of Chromosome Number Evolution in Yeast. *PLoS Genet*, **7**, e1002190.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Imakaev,M. *et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Jacques,N. *et al.* (2010) Population Polymorphism of Nuclear Mitochondrial DNA Insertions Reveals Widespread Diploidy Associated with Loss of Heterozygosity in Debaryomyces hansenii. *Eukaryot. Cell*, **9**, 449–459.

Jeffries,T.W. *et al.* (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. *Nat. Biotechnol.*, **25**, 319–326.

Kaplan,N. and Dekker,J. (2013) High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.*, **31**, 1143–1147.

Kellis,M. *et al.* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, **428**, 617–624.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

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

Lieberman-Aiden,E., van Berkum,N.L., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–93.

Louis,V.L. *et al.* (2012) Pichia sorbitophila, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization. *G3 GenesGenomesGenetics*, **2**, 299–311.

Lynch,D.B. *et al.* (2010) Chromosomal G + C Content Evolution in Yeasts: Systematic Interspecies Differences, and GC-Poor Troughs at Centromeres. *Genome Biol. Evol.*, **2**, 572–583.

Ober,R.J. *et al.* (2004) Localization Accuracy in Single-Molecule Microscopy. *Biophys. J.*, **86**, 1185–1200.

Oza,P. *et al.* (2009) Mechanisms that regulate localization of a DNA double-strand break to the nuclear periphery. *Genes Dev.*, **23**, 912–927.

Proux-Wera,E. *et al.* (2013) Evolutionary Mobility of the Ribosomal DNA Array in Yeasts. *Genome Biol. Evol.*, **5**, 525–531.

Roy,S. *et al.* (2010) Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.

Saha,S. *et al.* (2002) Using the transcriptome to annotate the genome, Using the transcriptome to annotate the genome. *Nat. Biotechnol. Nat. Biotechnol.*, **20, 20**, 508, 508–512.

Sexton,T. *et al.* (2012) Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell*, **148**, 458–472.

Souciet,J.-L. *et al.* (2009) Comparative genomics of protoploid Saccharomycetaceae. *Genome Res.*, **19**, 1696–1709.

Wang,J. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

Wong,H. *et al.* (2012) A Predictive Computational Model of the Dynamic 3D Interphase Yeast Nucleus. *Curr. Biol. CB*, **22**, 1881–90.

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.