Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Complete de novo assembly of yeast genomes using enzyme-free, dense optical genome mapping

Fredrik Westerlund

fredrikw@chalmers.se

Chalmers University of Technology https://orcid.org/0000-0002-4767-4868

Luis Leal-Garza

Chalmers University of Technology

Hanna Zachrisson

Chalmers University of Technology

Albertas Dvirnas

Chalmers University of Technology

Gaurav Goyal

Quantum Biosystems Inc.

Ikenna Obi

Umeå University https://orcid.org/0000-0003-0364-8964

Nasim Sabouri

Umeå University https://orcid.org/0000-0002-4541-7702

Tobias Ambjörnsson

Lund University

Article

Keywords:

Posted Date: May 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9608945/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

Large repetitive elements and complex chromosomal organization continue to challenge accurate assembly of eukaryotic genomes. Optical genome mapping (OGM) provides long-range genomic information by imaging individual DNA molecules. While most OGM protocols rely on sparse enzymatic labeling that restricts resolution in poorly labeled or structurally complex regions, dense labeling strategies generate continuous fluorescence intensity profiles that offer a complementary representation of genome structure.

Here we extend our Dense Optical Genome Mapping Assembly (DOGMA) pipeline, previously used for bacterial genomes, to eukaryotic genomes using a competitive binding (CB)-based dense OGM protocol that produces continuous intensity profiles, reflecting local AT/GC-content, along individual DNA molecules. We analyzed long DNA molecules from two model eukaryotes, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, with compact, yet structurally rich genomes. CB-based OGM captures chromosome-spanning molecules and large repetitive arrays (> 500 kbp), while also revealing genome-scale compositional features that influence barcode similarity and assembly behavior. By explicitly accounting for these features inherent to eukaryotic genomes, DOGMA reconstructs genome-wide optical maps while avoiding collapse of non-adjacent repetitive loci. The resulting assemblies identify structural variations and repetitive arrays that are incompletely represented in reference genomes. These results establish a framework for scalable dense OGM to increasingly complex genomes.

Introduction

Accurate and complete assembly of eukaryotic genomes remains a fundamental challenge in modern genomics. Despite substantial advances in sequencing technologies, many assemblies continue to suffer from unresolved gaps, collapsed repeats, and structural ambiguities. These limitations are particularly pronounced in regions characterized by long repetitive elements, segmental duplications, and complex chromosomal architectures[1]. While short-read sequencing platforms provide base-level accuracy, they inherently lack long-range information. Third generation long-read technologies extend read lengths into tens of kilobases, but remain susceptible to coverage biases, elevated error rates, and difficulties in resolving highly repetitive or low-complexity regions[2]. As a result, even well-studied eukaryotic genomes often contain poorly resolved regions such as centromeres, telomeres, ribosomal DNA (rDNA) arrays, and subtelomeric regions.

Optical Genome Mapping (OGM) has emerged as a powerful complementary approach for capturing long-range genomic structure. By imaging individual DNA molecules, hundreds of kilobases to megabases in length, OGM provides direct access to genome-wide structural information that is difficult to obtain through sequencing alone. OGM has proven particularly valuable for characterizing large structural variants, copy number changes, and repeat expansions, and has been successfully applied in both research and clinical settings[3–5]. Commercial implementations, such as those by Bionano Genomics, rely on enzymatic labeling of specific sequence motifs, producing sparse fluorescent tag

patterns (on average 9–15 per 100 kbp) along DNA molecules[6]. While this approach has demonstrated clinical utility, the relatively low label density limits its ability to resolve certain genomic regions, especially those with low motif frequency or structural complexity, including centromeres and large repeats.

As an alternative to sparse labeling, dense labeling strategies that create continuous intensity profiles have emerged. Some approaches employ DNA methyltransferases with short (4-nt) recognition motifs, resulting in closely spaced labels whose fluorescence signals overlap due to the diffraction limit[7]. Other dense OGM strategies avoid enzymatic labeling altogether, instead using AT/GC-specific non-covalent binding proteins or synthetic dyes, such as the AT-specific TAMRA polypyrrole[8,9].

A distinct enzyme-free dense labeling strategy developed in our laboratory is based on competitive binding (CB)[10]. CB-based OGM exploits the competitive interaction between YOYO-1, a non-specific fluorescent DNA bis-intercalator, and netropsin, a non-fluorescent molecule with binding specificity for AT-rich regions[10]. This interaction produces a continuous fluorescence intensity profile along each DNA molecule, where the signal strength reflects the underlying GC content. While this method cannot resolve individual fluorophores, and hence sequence sites, it generates information-dense intensity profiles (barcodes) that offer a fundamentally different representation of genomic structure compared to motif-based OGM. These barcodes can then be aligned to reference genomes or used for *de novo* genome assembly[11–13]. Since the fluorescence signal in CB-based OGM arises from YOYO-1 labeling of the DNA backbone, a dye already widely used in OGM and single-molecule assays[14], CB can be implemented without major modifications to existing instrumentation. This compatibility creates opportunities to integrate CB-based OGM with complementary labeling strategies, enabling multimodal analysis of genomic features along individual DNA molecules.

To exploit the full potential of CB-based OGM, we have recently developed DOGMA (Dense Optical Genome Mapping Assembly), a computational pipeline designed for *de novo* assembly of genomes using CB-derived barcodes[13]. DOGMA was initially validated on *Escherichia coli*, a bacterial genome characterized by limited redundancy, low large-scale compositional variation, and relatively uniform sequence organization, providing an ideal test for overlap accuracy and assembly contiguity.

Extending DOGMA to eukaryotic genomes introduces a complex set of challenges. In contrast to *E. coli*, even relatively small eukaryotic genomes contain complex features, including extensive repetitive regions and large-scale compositional domains that severely impede both sequence- and OGM-based assembly methods[1,15]. Repetitive elements often exceed sequencing read length and occur at multiple loci, making it difficult to assign reads or barcodes to unique genomic positions. Large-scale compositional domains, such as extended AT-/GC-rich regions or gradual changes in AT/GC content, further reduce barcode uniqueness by producing similar intensity profiles across long genomic regions. In OGM, both types of features can generate similar barcode patterns at distant loci, leading to ambiguous placement during alignment and assembly. Examples of these features include rDNA arrays,

centromeres, telomeres, and subtelomeric regions, which are structurally complex, highly repetitive, and often appear at multiple loci in a genome[16–19]. As a result, these regions are systematically underrepresented or misassembled in reference genomes.

In this study, we extend DOGMA to assembly of eukaryotic genomes using *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* as model systems. These two yeasts provide complementary test cases for CB-based OGM due to their compact, yet structurally rich, genomes and distinct chromosomal architectures. In *S. cerevisiae*, regions such as the rDNA array on chromosome XII are associated with genome stability, cellular aging, and nucleolar organization and are notoriously difficult to resolve[20–22]. Likewise, in *S. pombe*, similar arrays located near both ends of chromosome III pose challenges during DNA replication and remain incompletely resolved by sequencing-based approaches, consistent with their repetitive structure[23,24]. Similar repetitive structures are widespread in larger eukaryotic genomes.

DOGMA is here used to generate *de novo* OGM assemblies of eukaryotic genomes that capture key structural features, including telomeric regions and large repetitive arrays, while explicitly examining regions where assembly contiguity breaks down. Reference-based analyses are used for validation and biological interpretation, without guiding the assembly process itself. These results demonstrate the feasibility of dense OGM assembly in eukaryotes and the biologically driven challenges that must be addressed when extending OGM-based assembly methods beyond bacterial genomes.

Results

To perform OGM, we stained genomic DNA samples and stretched them by confinement in nanofluidic devices. Details of the experimental and computational workflows are described in Materials and Methods. In brief, DNA molecules were extracted from yeast cells enclosed in agarose plugs, labeled by competitive binding of YOYO-1 and netropsin, extended in nanochannels, and imaged by fluorescence microscopy, producing continuous intensity profiles along individual DNA molecules. Images were then processed to extract one-dimensional fluorescence intensity profiles, referred to as barcodes (Fig. 1a). The extracted barcodes underwent quality control and repetitive region screening prior to assembly (Fig. 1b). Assembly was carried out using a multi-step implementation of the DOGMA pipeline[13], grouping barcodes into bargroups (contiguous optical maps created from barcodes) based on similarity and contextual consistency (Fig. 1c). Assembled bargroups were then aligned to reference genomes for validation and structural comparison (Fig. 1d). Finally, consensus bargroups were generated and compared to each reference chromosome (Fig. 1e).

Reference-based alignment

A total of 1,102 *S. pombe* and 551 *S. cerevisiae* CB-labeled DNA molecules were imaged and converted into barcodes. To evaluate genome-wide coverage prior to the *de novo* assembly, the barcodes were aligned to theoretical reference barcodes derived from the corresponding reference genomes (see Methods)[25,26]. The alignment was performed using a full-length alignment strategy, in which each query

barcode is aligned in its entirety to the reference, with the Pearson cross-correlation (CC-value) scoring method (**Fig. 2**, see Methods for details).

Genome-wide alignments were obtained for all three *S. pombe* chromosomes and all sixteen *S. cerevisiae* chromosomes **(Fig. 2)**. Reference-based coverage exceeded 98% for both genomes, with mean coverage depths of 29.5x for *S. pombe* and 16.4x for *S. cerevisiae*, indicating that the dataset provides near-complete genome representation for downstream assembly.

Despite the general overall uniformity in coverage, several genomic regions, particularly in *S. pombe*, displayed elevated coverage. Given the random sampling of DNA inherent to OGM, pronounced local enrichment is not expected. Elevated coverage is more likely to reflect alignment ambiguity, where barcodes from distinct loci exhibit high full-length similarity and are therefore assigned to alternative genomic positions. Such ambiguity can arise when regions at different loci share similar intensity profiles, or when sequences are incompletely represented in the reference assembly, leading to misplacement.

A complementary pattern was observed in specific regions of the *S. cerevisiae* genome, where reduced coverage was observed. This pattern is consistent with a limitation of the full-length alignment strategy, which requires query barcodes to be shorter than the reference. Molecules approaching or exceeding chromosome length are therefore unable to align optimally, and are instead placed at alternative loci, contributing both to apparent depletion at their true origin and enrichment elsewhere.

As an additional specificity control, barcodes from each species were aligned against the reference genome of the other species **(Fig. 2)**. While a majority of the barcodes (76.4%) aligned preferentially to their correct species, a subset exhibited comparable or higher global similarity to regions in the reference of the other species. After inspecting cross-species matches (**Fig. 2**, insets), we noticed that they occurred mainly around specific loci for each genome, creating alignment ambiguity hotspots. We also observed that cross-species matches were enriched for barcodes characterized by either long-range (200-300 kbp) low-frequency intensity trends or extended repetitive regions, both of which reduce barcode uniqueness and increase alignment ambiguity.

Overall, the reference-based analysis indicated that global alignment provides reliable genome-wide coverage estimates, but that it is susceptible to alignment ambiguity under specific structural conditions. To clarify the origin of these effects and assess their impact on assembly, we next examined experimental and theoretical barcode properties influencing misalignment prior to initializing the *de novo* assembly.

Structural features driving alignment ambiguity in eukaryotic OGM

Analysis of the barcode dataset revealed three features that strongly influence similarity scoring and assembly behavior: chromosome-spanning molecules, repetitive barcode components, and long-range

intensity trends. Each of these properties can affect similarity scores and influence the grouping during assembly.

Chromosome-spanning molecules are captured by OGM

Using the full-length alignment framework, molecules exceeding the theoretical chromosome length cannot be properly aligned, leading to misplacement or exclusion. This effect was particularly evident for the smaller chromosomes in *S. cerevisiae*, where full-length molecules exceed the corresponding theoretical barcode length.

Individual barcodes, up to 800 kbp in size, fully spanned ten of the sixteen *S. cerevisiae* chromosomes (examples in **Fig. 3a**), demonstrating the presence of chromosome-scale molecules in the dataset. This observation highlights the capacity of CB-based OGM to capture entire eukaryotic chromosomes within single intact DNA molecules. For sufficiently small chromosomes, and given limited DNA fragmentation, assembly may therefore be simplified through local alignment (see Methods), which identifies the highest-scoring subregions between barcode and reference without requiring full-length matching.

Repetitive barcode components cause alignment ambiguity

A subset of barcodes exhibited broad alignment across multiple genomic loci and were identified as being enriched in repetitive features. The identified repetitive units ranged between 9 and 10.5 kb in size, consistent with known repetitive regions in *S. cerevisiae* and *S. pombe*, respectively. This repetitive DNA is known to be underrepresented or collapsed to just two copies per locus in current reference assemblies[26,27].

When repetitive regions are not masked, their repetitive components dominate the alignment, producing visible misalignments to incorrect loci **(Fig. 3b.i)**. To address this issue, and to incorporate repetitive DNA into the alignment and assembly more effectively, we developed a repetitive-element detection algorithm (see Supplementary Methods 2) that identifies and masks repetitive elements. The algorithm also classifies barcodes as fully repetitive, partially repetitive (when a subregion of the barcode has at least five contiguous repetitive units), and non-repetitive.

After masking the repetitive segments, the remaining non-repetitive portion of the barcode can be used to determine the correct genomic locus. As illustrated in **Fig. 3b.ii**, masking resolves the ambiguity and enables consistent alignment of the same molecules to a single genomic locus. Since the pair-wise similarity scoring of the assembly process is similar to the scoring system from the alignment process, it is expected that the resolution of alignment ambiguity via masking of repetitive regions can improve the bargroup formation during DOGMA assembly as well.

Long-range GC-dependent intensity gradients influence barcode similarity

The reference-based analysis revealed recurrent cases in which barcodes from distinct genomic loci exhibited high global similarity. Inspection of these barcodes suggested the presence of gradual, long-

range intensity variations in addition to the typical CB-based OGM barcode features. To better understand the origin of this behavior, we searched for long-range intensity trends in the theoretical reference barcodes.

Smoothing of theoretical reference barcodes using a moving-average filter revealed pronounced long-range intensity trends spanning hundreds of kilobases in several locations **(Fig. 3c)**, reflecting genome-scale compositional domains that can be detected by dense CB-based labeling. In *S. pombe*, these features were particularly evident, with gradually decreasing GC-content extending ~300 kbp toward chromosome ends and characteristic low-GC motifs spanning ~200 kbp at centromeric regions. While local barcode structure is defined by short-scale intensity variations (5–20 kbp), these long-range trends introduce an underlying "low-frequency" barcode component across much larger genomic distances.

Consistent with this organization, experimental barcodes sharing similar long-range intensity trends frequently exhibited ambiguous alignments, including preferential placement at incorrect loci or, in some cases, within the reference genome of the other species **(Fig. 2)**.

These results demonstrate that CB-based OGM captures biologically meaningful genome-scale compositional information that is largely inaccessible to sparse labeling approaches that just label specific motifs. While these long-range intensity trends pose challenges for similarity-based alignment and assembly, they also provide direct insight into large-scale chromosomal organization, highlighting both a limitation and a unique strength of dense OGM.

To validate that the elevated coverage observed in **Fig. 2** arises from intrinsic genome-scale signal organization, rather than coverage limitations, intragenomic self-similarity analysis was performed across the *S. pombe* chromosomes (Supplementary Fig. 4). Chromosomal regions ≥150 kbp exhibiting CC-value ≥0.75 were considered highly similar. This analysis revealed discrete similarity "hotspots" shared between chromosomes, consistent with shared long-range signal features captured in CB-based OGM data **(Fig. 3c)**. These long-range features can dominate global similarity measures despite divergent local structure, providing a mechanistic explanation for the coverage inflation and reduced alignment confidence observed in the reference-based analysis.

Together, these findings indicate that structural properties intrinsic to eukaryotic genomes, rather than insufficient coverage, represent the primary constraints on assembly using dense OGM.

Molecule characteristics before and after quality control processing

Mean molecule lengths prior to quality control (QC) were 337.8 kbp for *S. pombe* and 359.7 kbp for *S. cerevisiae*. The QC procedures included stitching molecules spanning multiple imaging fields and masking barcode regions dominated by repetitive features (see Methods). Repetitive segments were masked rather than removed, preserving the non-repetitive parts for downstream analysis. Following QC, the mean molecule lengths remained largely unchanged (337.6 kbp for *S. pombe* and 366.0 kbp for *S.*

cerevisiae), indicating that filtering primarily addressed problematic features without introducing size-related bias.

A second reference-based alignment was performed to evaluate the QC (Supplementary Fig. 3). This alignment implemented local barcode-to-reference matching to accommodate chromosome spanning molecules, repetitive masking to account for repetitive regions, and a ΔCC threshold to filter barcodes with alignment ambiguity. Increasing the ΔCC threshold improved organism-specific alignment accuracy while reducing the fraction of retained barcodes, illustrating the trade-off between specificity and molecule retention **(Fig. 3d)**. An optimal threshold of ΔCC = 0.03 maximized correct alignments while maintaining a near complete estimated coverage of both genomes.

Bargroup formation and consensus barcode construction

Having characterized the structural features influencing similarity metrics, we next assembled QC-filtered and repeat-masked barcodes into bargroups using the DOGMA pipeline. Repetitive elements were identified and masked from barcodes prior to assembly. Assembly without masking produced bargroups containing misaligned repetitive and non-repetitive segments **(Fig. 4a.i)**. Masking allowed the non-repetitive parts of barcodes to drive the assembly, while still maintaining the repetitive content for post-assembly analysis. This strategy enabled correct assembly of molecules that would otherwise be mis-grouped or left ungrouped, while also resolving incorrect groupings driven by repetitive segments **(Fig. 4a.ii)**.

To mitigate grouping driven primarily by shared long-range intensity trends rather than local structural similarity, we introduced a bargroup scoring metric based on signal-to-noise characteristics (see Methods). This metric prioritizes bargroups with well-defined local features consistent among their constituting barcodes **(Fig. 4b.i)** and penalizes bargroups dominated by smooth, low-frequency signal components with constituting barcodes that differ in their short-frequency features **(Fig. 4b.ii)**.

Bargroup assembly applied pairwise similarity constraints incorporating limited stretching variation, positional tolerance in pixel coordinates, and statistical thresholds on correlation significance (see Methods). The assembled bargroups were subsequently aligned to the corresponding reference genomes using a local bargroup-to-reference alignment strategy to assess genome-wide coverage and contiguity. Each bargroup consensus was represented as an averaged barcode intensity profile derived from its constituent barcode members **(Fig. 4c)**.

Overall, approximately 20% and 30% of the barcodes used as input for the bargrouping process contributed to bargroup formation from *S. pombe* and *S. cerevisiae*, respectively, while the rest remained ungrouped despite passing QC. Ungrouped molecules were retained for downstream coverage analysis and gap resolution.

Genome-wide assemblies reveal chromosomal extensions

Genome-wide coverage by assembled bargroups closely mirrored the high reference-alignment coverage observed prior to assembly, indicating that most chromosomal regions were represented within contiguous optical maps. Across multiple chromosomes, bargroups spanned megabase-scale intervals, demonstrating robust large-scale assembly performance.

Following initial assembly, bargroups were aligned to the reference genomes using a local bargroup-to-reference alignment strategy. Adjacent bargroups with consistent placements were subsequently merged based on local bargroup-to-bargroup similarity to generate chromosome-scale assemblies. The resulting assemblies retained only bargroups satisfying the meanCC, stdCC, and SNR quality criteria defined during bargroup formation (Supplementary Table 3).

The final assemblies generated for both *S. pombe* and *S. cerevisiae* included highly repetitive regions typically underrepresented in sequence-based assemblies **(Fig. 5)**. Some of the assembled chromosomes show minor extensions (>6 kbp) relative to the reference chromosomes in the end(s) of each chromosome. These can be explained as a product of our theoretical reference model, which omits the intensity of the first and last 6 pixels, assuming them to be uncertain.

Local breaks in contiguity were observed at *S. pombe* chromosome I from position 4466 kbp to 4787 kbp and chromosome II from position 3208 kbp to 3211kbp (Supplementary Fig. 5). To distinguish assembly limitations from absence of coverage, we examined high-confidence barcodes that aligned strongly to the reference, but did not group into bargroups. Inclusion of these molecules revealed that regions lacking assembled contigs were nevertheless supported by individual barcodes, resulting in near-complete genome-wide molecular representation.

Large repetitive structures were resolved in both organisms. In *S. cerevisiae*, a large repetitive block interrupted the reference assembly of chromosome XII near 479 kbp. This region extended ≥550 kbp and consisted of ~9 kbp repeat units, corresponding to ≥60 tandem copies. In *S. pombe*, extended repeat arrays were observed at both ends of chromosome III beyond the length represented in the current reference assembly, spanning at least 1 Mb in total (≥500 kbp from each telomeric end). Individual repeat units measured ~10.5 kbp, consistent with rDNA repeats[19,27] and indicating the presence of ≥50 tandem copies at each end.

Structural variations (SV) in the form of smaller insertions were also detected in chromosomes such as *S. cerevisiae* chromosome VIII and III. We were able to manually curate a ~30 kbp insertion in chromosome VIII (position 206 kbp). A similar-sized insertion on chromosome III between positions 100 kbp and 200 kbp was also identified, but its insertion point was not precisely determined. SV detection is not streamlined into our pipeline, making it non-trivial to pinpoint exact insertion points.

Discussion

This work demonstrates that dense optical genome mapping (OGM) assembly using the DOGMA pipeline can be extended from bacterial to eukaryotic genomes. In doing so, it reveals constraints and

opportunities that are largely invisible in simpler systems, such as bacterial genomes, and require targeted methodological adaptations. Applying DOGMA to *S. cerevisiae* and *S. pombe* demonstrates that OGM-based assembly of eukaryotic genomes is constrained less by genome size and more by repeat composition and long-range genome organization.

For *E. coli*, DOGMA achieved nearly complete genome reconstruction with minimal preprocessing, reflecting the relative simplicity and low redundancy in bacterial genomes[28]. The transition to *S. cerevisiae* and *S. pombe* revealed a fundamentally different assembly landscape. Yeast genomes introduced an increase in barcode ambiguity due to repetitive genomic features such as rDNA arrays and subtelomeric repeats[21,22,25,29].

By explicitly identifying and isolating repetitive signal components, DOGMA can assemble repeat-rich regions without collapsing non-adjacent loci and subsequently reintegrate them using unique flanking context. Assembly of these repetitive regions with OGM is solely limited to the size of the isolated DNA to ensure the collection of data including repetitive regions and their adjacent non-repetitive loci. This demonstrates that repeat content is not a fundamental limitation of dense optical genome assembly; rather, the critical factor is the ability to detect, classify, and appropriately weight repetitive and long-range signal features during assembly.

Another key insight of this study is that long-range intensity trends, spanning hundreds of kilobases, contribute substantially to barcode similarity across distinct loci. In *S. pombe*, these trends manifest as chromosome GC gradients and centromeric low-GC domains, which are consistently detectable in both theoretical and experimental barcodes. While such trends can bias global similarity metrics and promote spurious alignments, they also reflect large-scale chromosomal organization that is inaccessible to sparse-label OGM approaches. Accounting for this signal component will therefore be essential for future OGM assembly frameworks.

Together, these findings position dense OGM as both an assembly tool and a probe of large-scale genome organization. While further methodological refinement will be required to fully exploit long-range signal information in even more complex genomes, the complete yeast assemblies presented here establish an important foundation for extending dense OGM to larger eukaryotic systems and to biologically and clinically relevant repetitive loci.

From a methodological perspective, the combination of repeat-aware masking and characterization of long-range intensity profiles provides a generalizable strategy for extending dense, CB-based, OGM beyond yeast. Many of the challenges encountered here, such as repetitive arrays and long-range compositional structure, are even more common in mammalian genomes, where repetitive regions constitute over half of the human genome and remain systematically underrepresented or collapsed in many reference assemblies[30,31]. Despite recent advances, such as telomere-to-telomere assemblies, large repetitive loci remain difficult to characterize robustly with sequencing alone, particularly when copy number, structural heterogeneity, or long-range organization are biologically relevant[30]. In this

context, dense OGM offers a complementary view, capturing continuous intensity profiles, that are orthogonal to sequence-based information, over hundreds of kilobases. While evaluation in human genomes lies beyond the scope of this study, the framework presented here suggests a path toward resolving medically relevant repeat structures, including large satellite arrays, rDNA clusters, and disease-associated macrosatellites[3,17].

An additional advantage of CB-based, dense OGM is that the intensity variation signal arises from YOYO-1 labeling of the DNA backbone, a fluorophore already used in many OGM and single-molecule assays to simply detect the DNA backbone. Therefore, the incorporation of competitive binding does not require additional excitation sources, emission filters, or major instrumentation changes, making this signal layer effectively available in already existing platforms. This creates opportunities to combine CB-based OGM with other labeling strategies, such as enzymatic motif labeling, damage mapping[32], or epigenetic assays[14], while retaining long-range genomic context. In this way, CB-based OGM is not limited to structural reconstruction, but can serve as a scaffold for integrating multiple molecular readouts along individual DNA molecules.

In conclusion, this study demonstrates that dense, continuous optical genome mapping can be extended from bacterial to eukaryotic genomes when eukaryotic structural constraints are appropriately considered. By combining repeat-aware masking with local alignment assembly, DOGMA reconstructs genome-wide optical maps in yeast without collapsing non-adjacent loci. Importantly, CB-based OGM preserves large-scale compositional features of chromosomes, enabling analysis of genomic structure beyond strictly local sequence similarity. Together, these results establish DOGMA as a robust framework for eukaryotic optical genome assembly and lay the groundwork for applying dense OGM to increasingly complex genomes.

Methods

Yeast cultivation and DNA extraction

Experiments were performed using *Saccharomyces cerevisiae* strain BY4742 and *Schizosaccharomyces pombe* strain NS112 (*leu1-32 his3-D1 ura4-D18 ade6-M210*). The protocol for DNA extraction was adapted from previous work[33]. In short, *S. cerevisiae* was grown overnight and used to inoculate a culture that was grown at 30°C for 6 to 8 hours to reach an OD600 of 0.1. Afterwards, 6 x10^8 cells were harvested and frozen at -20°C until the DNA extraction was performed. For *S. pombe*[34], 50 mL exponentially growing cells at initial concentration of 0.5 million cells/mL were grown in Edinburgh minimal media for 16 h at 30°C until they reached 14 million cells/mL. Cells were harvested by centrifugation at 3000 rpm for 5 mins and washed twice with 50mM cold EDTA. Cells were frozen at -20°C prior to DNA extraction. The extraction of DNA from both types of yeast cells was performed using the CHEF Yeast Genomic DNA Plug Kit from BIO-RAD. In short, 1 mL of agarose plugs were created for each strain by combining 6 x10^8 cells with 625 µL CSB (Cell Suspension Buffer) and 375 µL 2% CCA

(CleanCut agarose). Agarose plugs were then treated with Lyticase, RNase, and Proteinase K, and washed several times according to the CHEF Yeast Genomic Plug Kit protocol.

DNA was then extracted from the agarose plugs. This is done by first incubating in 1x CutSmart (New England Biolabs) solution at 70°C for 10 minutes, followed by incubation at 42°C for 10 minutes. This was followed by a 42°C incubation with 2 U agarase for 2 hours (New England Biolabs). DNA was then quantified by using Qubit dsDNA Broad Range Assay kit (Invitrogen).

CB-based OGM lab procedures

Densely labelled DNA molecules were obtained by a single-step competitive binding (CB) based staining reaction, as described previously[33].The CB staining reactions were performed by adding 4-6 μM DNA (in bases), 0.4-0.6 μM YOYO-1, and 120-180 μM netropsin in a 10:1:300 ratio to a 0.5X TBE (Tris-Borate EDTA) 10 μL solution. DNA from bacteriophage λ (48,502 bp bought from Roche) was added to make up ~30% of the DNA mass in the staining reactions, to serve as an internal size reference. Finally, an incubation at 50°C for 30 minutes was followed by a 15-fold dilution with MQ-water and 3 μL of β-mercaptoethanol (BME) to reduce photodamage.

Nanofluidic experiments

Stained DNA molecules were stretched for imaging using a nanofluidic device whose fabrication is described elsewhere[34–36]. These devices are designed to have two sets of two loading wells, where each set is connected via a separate microchannel. Both microchannels connecting the loading wells are connected to each other by 120 parallel nanochannels, each having a cross section of 100 x 150 nm^2 and extending 500 μm in length. To ensure uniform conditions, the chips underwent pre-conditioning with 0.033X TBE buffer supplemented with 2% BME v/v prior to sample loading. DNA loading was done by adding 10 μL of stained sample into one of the loading wells, while the others contained the same buffer used for pre-conditioning, but no DNA. DNA molecules were then driven first into the microchannels and then the nanochannels by pressure-driven N_2 flow. Stretched and stained DNA molecules in the nanochannels were then imaged by using an inverted fluorescence microscope (Zeiss Axio Observer D.1) equipped with a 63x oil immersion objective (NA = 1.46, Zeiss), a Colibri 7 light source (Zeiss), and a Photometrics Evolve EMCCD camera. Systematic acquisition of videos of up to 100 frames was carried out for each molecule, utilizing an exposure time of 100 ms and a FITC filter set (Zeiss).

Barcodes processing

Barcodes were generated from videos using a previously developed custom MATLAB software called 'Ildev'[37,38]. In short, DNA molecules are detected in each time frame within a movie. The detected DNA molecules are then turned into kymographs, where each row of pixels represents the intensity along the DNA molecule in a single frame. The kymographs are aligned to compensate for small thermal fluctuations of the DNA inside the nanochannels. Aligned kymographs are then processed by a custom

quality control (QC) algorithm to merge molecules spanning multiple fields of view (multi-FoV) and to remove molecules that show a significant change in size during imaging (see Supplementary Methods 1). Finally, the filtered and aligned kymographs are time-averaged, rendering a time series-like 1D intensity trace along the extension of the DNA molecule, referred to as a barcode.

Reference alignment

Reference genome sequences were used as an input to previously established methods[11] to *in silico* generate theoretical reference barcodes that can be used for alignment and coverage analysis. For *S. cerevisiae*, the S288C reference genome (R64 assembly), obtained from Saccharomyces Genome Database[26], was used as reference since the BY4742 strain was derived from S288C. For *S. pombe*, the reference genome of strain 972h−, obtained from PombeBase[27], was used as a reference since strain NS112 (*leu1-32 his3-D1 ura4-D18 ade6-M210*) was derived from the 972 strain background.

Genome-wide barcode alignment was performed with a full-length similarity search alignment strategy to estimate coverage across the reference genomes. This approach, implemented in previous work[33], is based on a Pearson correlation coefficient (CC) scoring system to identify the best placement for a query barcode within a given reference. In this strategy, the full query barcode is aligned to a segment of the reference, and therefore requires the query to be shorter than the reference.

To identify individual molecules spanning complete *S. cerevisiae* chromosomes, a local similarity alignment was used in a barcode-to-reference mode. This approach uses a sliding window to identify the highest-scoring subregions between query and reference, allowing partial matches and overhang beyond reference boundaries. Both query and reference must exceed the selected overlap window length. The local alignment is based on a matrix profile framework and is also used in a barcode-to-barcode mode as the core function for pairwise similarity calculations in DOGMA[13,39]. For each chromosome, an overlap window corresponding to 70% of the chromosome length was used. Candidate barcodes were restricted to lengths between 70% and 150% of the corresponding chromosome size to accommodate molecule length variability while ensuring chromosome-scale coverage.

Following assembly, bargroup alignment to the reference genome was performed using local similarity alignment rather than full-length alignment. This choice accounts for assembled contigs that may extend over the edges or exceed the full length of the corresponding reference chromosomes due to the presence of insertions, repetitive or unresolved genomic segments. Local alignments were calculated using an overlap window corresponding to 70% of the non-repetitive portion of each contig.

For bargroup alignment to the reference, chromosome XII of the *S. cerevisiae* genome was split into left and right side after noticing partial alignment due to repetitive regions not included in the reference. The breakpoint for the insertion of repetitive region into chromosome XII was manually determined to be around position 472 kbp.

Repetitive Finder

Since DOGMA uses local barcode-to-barcode alignment for its bargrouping procedure, repetitive regions shared between molecules can dominate the matching process, potentially leading to non-canonical groupings independent of the non-repetitive content elsewhere in the molecules. For this reason, it is important to identify molecules with repetitive regions and mask the repetitive part of them so that the rest of the molecule can be used as a contextual “anchor” in the bargrouping process. This way, bargroups are created based on the similarity of the non-repetitive parts of each barcode, while the repetitive parts are still preserved and can be “unmasked” for downstream analysis.

To identify and mask highly repetitive regions prior to assembly, we developed a custom *Repetitive Finder* algorithm implemented in MATLAB (Supplementary Fig. 2). The method detects periodic patterns in barcode intensity profiles using the autocorrelation function to estimate repeat periodicity, followed by Fourier transform validation. A rolling-window correlation then identifies continuous repetitive regions, with segmentation achieved using the findchangepts function (Supplementary Methods 2). Regions with an average correlation >0.4 across at least five repeat units are marked as repetitive and masked in subsequent assembly steps. These empirically determined thresholds effectively separate highly repetitive barcodes from non-repetitive ones (Supplementary Fig. 2).

DOGMA pipeline implementation and bargroup quality metrics

Assembly was performed using the DOGMA algorithm as previously described[13]. Pairwise barcode comparisons were evaluated using normalized Pearson cross-correlation under constrained alignment parameters, including stretch tolerance, positional tolerance, and statistical significance thresholds defined by the DOGMA null model framework (Supplementary Table 1).

Because repetitive, partially repetitive, and non-repetitive barcode classes exhibit distinct similarity distributions, parameter values within the null-model-based p-value framework were optimized separately for each class using the parametrization.m script within the DOGMA package (Supplementary Table 2). This parametrization is recommended to be run once per experimental setup, since parameters are greatly affected by pixel size and effective PSF of the set up.

To enable incorporation of repetitive regions, assembly was performed in multiple stages reflecting barcode class. Independent DOGMA assemblies were first carried out on non-repetitive and partially repetitive masked barcodes. Repetitive barcodes were then assembled in a subsequent stage using the partially repetitive assemblies as anchoring context. This staged strategy prevents repetitive segments from driving initial grouping while allowing their incorporation once non-repetitive context has been established. Finally, all resulting bargroup contigs were merged through an additional DOGMA-based local assembly step to generate unified bargroups spanning all barcode classes. Consensus bargroups were constructed by averaging aligned barcode intensity profiles following stretch normalization. To mitigate grouping driven by long-range low-frequency trends and to ensure that only structurally consistent and information-rich groups were retained, three complementary quality metrics were evaluated for each bargroup.

Internal similarity was first assessed using pairwise Pearson cross-correlation coefficients (CC) computed between all barcodes within each bargroup and their consensus. The mean of these values (meanCC) was required to exceed 0.7, ensuring overall coherence among members. The standard deviation of the CC values (stdCC) was also evaluated to measure internal consistency; bargroups with stdCC > 0.05 were discarded, as high dispersion indicates heterogeneous membership despite acceptable mean similarity.

Finally, a signal-to-noise ratio (SNR) was computed to quantify the strength of the consensus structure relative to variability among its members. For a bargroup with N members, let $B_i(x)$ denote the intensity at pixel x of barcode i, and $\bar{B}(x) = \frac{1}{N}\sum_{i=1}^{N} B_i(x)$ the consensus intensity. The SNR was defined as in equation (1) where the numerator reflects variation across the consensus profile, and the denominator reflects the average per-pixel variability among individual barcodes. Bargroups with SNR < 1.0 were excluded.

$$\mathrm{SNR} = \frac{\mathrm{std}_x[\bar{B}(x)]}{\mathrm{mean}_x[\mathrm{std}_i[B_i(x)]]} \quad (1)$$

Only bargroups satisfying all three criteria were retained for downstream assembly.

Barcodes that remained unassembled but exhibited high-confidence reference alignments were subsequently anchored to the theoretical genome map to maximize genome-wide representation.

Declarations

Acknowledgements

We thank Dr. Mikael Molin for providing the Saccharomyces cerevisiae strain BY4742 and for valuable advice on yeast cultivation. The nanofluidic devices used in this work were fabricated using Chalmers MyFab cleanroom facility.

Funding sources

F.W. discloses support for the research of this work from the Swedish Childhood Cancer fund (Barncancerfonden, MT2022-0003) and the Swedish Cancer Foundation (Cancerfonden, 24 3885 Pj). Work in N.S. lab was supported by Knut and Alice Wallenberg foundations (KAW.2021.0173). T.A. and the rest of the authors declare no relevant funding.

Author Contributions

Conceptualization: L.M.L.G., G.G., F.W.

Methodology: L.M.L.G., A.D.

Software: L.M.L.G., A.D.

Validation: L.M.L.G.

Formal analysis: L.M.L.G.

Investigation: L.M.L.G., H.Z.

Resources: S.K.K., I.O., N.S.

Data curation: L.M.L.G., H.Z.

Writing – original draft: L.M.L.G., F.W.

Writing – review & editing: All authors

Visualization: L.M.L.G.

Supervision: F.W., T.A.

Project administration: L.M.L.G., F.W.

Funding acquisition: F.W.

Competing interests

The authors declare no competing interests.

Data and Code availability

The datasets generated and analyzed during this study, together with the custom MATLAB code used for barcode processing and assembly, are available to reviewers via a private Google Drive link provided at submission (https://drive.google.com/drive/folders/1aEM9YhjKS1UgG0qovO57dkADrqzxVKeG?usp=sharing). Upon acceptance, these materials will be made publicly available through a permanent repository (e.g. Zenodo) with assigned DOIs. Source data including raw microscopy images, processed barcode data and assembly outputs, are included in these deposits.

References

1. Liao X et al (2019) Current challenges and solutions of de novo assembly. *Quantitative Biology* vol. 7 90–109 Preprint at https://doi.org/10.1007/s40484-019-0166-9
2. Yuan Y, Chung CYL, Chan TF (2051) Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal* vol. 18 –2062 Preprint at https://doi.org/10.1016/j.csbj.2020.07.018 (2020)

3. Dai Y et al (2020) Single-molecule optical mapping enables quantitative measurement of D4Z4 repeats in facioscapulohumeral muscular dystrophy (FSHD). J Med Genet 57:109–120
4. van der Sanden B et al (2025) Optical genome mapping enables accurate testing of large repeat expansions. Genome Res 35:810–823
5. Weissensteiner MH et al (2017) Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. Genome Res 27:697–708
6. Dremsek P et al (2021) Optical genome mapping in routine human genetic diagnostics—its advantages and limitations. Genes (Basel). 12
7. Grunwald A et al (2015) Bacteriophage strain typing by rapid single molecule analysis. Nucleic Acids Res. 43
8. Park J et al (2019) Single-molecule DNA visualization using AT-specific red and non-specific green DNA-binding fluorescent proteins. Analyst 144:921–927
9. Lee S et al (2018) TAMRA-polypyrrole for A/T sequence visualization on DNA molecules. Nucleic Acids Res 46
10. Nyberg LK et al (2012) A single-step competitive binding assay for mapping of single DNA molecules. Biochem Biophys Res Commun 417:404–408
11. Nilsson AN et al (2014) Competitive binding-based optical DNA mapping for fast identification of Bacteria - Multi-ligand transfer matrix theory and experimental applications on Escherichia coli. Nucleic Acids Res 42
12. Müller V et al (2020) Cultivation-Free Typing of Bacteria Using Optical DNA Mapping. ACS Infect Dis 6:1076–1084
13. Dvirnas A et al (2025) DOGMA: de novo assembly of densely labelled optical DNA maps using a matrix profile approach. PLoS ONE 20:e0335633
14. Jeffet J, Margalit S, Michaeli Y, Ebenstein Y (2021) Single-molecule optical genome mapping in nanochannels: Multidisciplinarity at the nanoscale. *Essays in Biochemistry* vol. 65 51–66 Preprint at https://doi.org/10.1042/EBC20200021
15. Hartley G, O'neill RJ (2019) Centromere repeats: Hidden gems of the genome. *Genes* vol. 10 Preprint at https://doi.org/10.3390/genes10030223
16. Klein SJ, O'Neill RJ (2018) Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Research* vol. 26 5–23 Preprint at https://doi.org/10.1007/s10577-017-9569-5
17. Yang J, Li F (2017) Are all repeats created equal? Understanding DNA repeats at an individual level. *Current Genetics* vol. 63 57–63 Preprint at https://doi.org/10.1007/s00294-016-0619-x
18. Smirnov E, Chmúrčiaková N, Liška F, Bažantová P, Cmarko D (2021) Variability of human rDNA. *Cells* vol. 10 1–14 Preprint at https://doi.org/10.3390/cells10020196

19. Oizumi Y et al (2021) Complete sequences of Schizosaccharomyces pombe subtelomeres reveal multiple patterns of genome variation. Nat Commun 12
20. Egidi A, Di Felice F, Camilloni G (2020) Saccharomyces cerevisiae rDNA as super-hub: the region where replication, transcription and recombination meet. *Cellular and Molecular Life Sciences* vol. 77 4787–4798 Preprint at https://doi.org/10.1007/s00018-020-03562-3
21. D'Alfonso A, Micheli G, Camilloni G (2024) rDNA transcription, replication and stability in Saccharomyces cerevisiae. *Seminars in Cell and Developmental Biology* vols 159–160 1–9 Preprint at https://doi.org/10.1016/j.semcdb.2024.01.004
22. James SA et al (2009) Repetitive sequence variation and dynamics in the ribosomal DNA array of Saccharomyces cerevisiae as revealed by whole-genome resequencing. Genome Res 19:626–635
23. Aquiles Sanchez J, Kim S-M, Huberman JA (1998) Ribosomal DNA Replication in the Fission Yeast, Schizosaccharomyces Pombe. Exp Cell Res 238
24. Sabouri N, McDonald KR, Webb CJ, Cristea IM, Zakian VA (2012) DNA replication through hard-to-replicate sites, including both highly transcribed RNA Pol II and Pol III genes, requires the S. pombe Pfh1 helicase. Genes Dev 26:581–593
25. Wood V et al (2002) The genome sequence of Schizosaccharomyces pombe. Nature 415:871–880
26. Cherry JM et al (2012) Saccharomyces Genome Database: The genomics resource of budding yeast. Nucleic Acids Res 40
27. Harris MA et al (2022) Fission stories: using PomBase to understand Schizosaccharomyces pombe biology. Genetics 220
28. Sela I, Wolf YI, Koonin EV (2016) Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 113, 11399–11407
29. Pasero P, Marilley M (1993) Size variation of rDNA clusters in the yeasts Saccharomyces cerevisiae and Schizosaccharomyces pombe. Mol Gen Genet 236:448–452
30. Nurk S et al The complete sequence of a human genome. *Science (*(1979)). 4453 (2022)). 4453 (2022)
31. Lander S et al (2001) Initial Sequencing and Analysis of the Human Genome International Human Genome Sequencing Consortium* The Sanger Centre: Beijing Genomics Institute/Human Genome Center. *NATURE* vol. 409 www.nature.com
32. Detinis Zur T et al (2025) Single-molecule toxicogenomics: Optical genome mapping of DNA-damage in nanochannel arrays. DNA Repair (Amst). 146
33. Müller V et al (2019) Enzyme-free optical DNA mapping of the human genome using competitive binding. Nucleic Acids Res 47
34. Persson F, Tegenfeldt JO (2010) DNA in nanochannels-directly visualizing genomic information. Chem Soc Rev 39:985–999
35. Frykholm K, Müller V, Kk S, Dorfman KD, Westerlund F (2022) DNA in nanochannels: theory and applications. *Quarterly Reviews of Biophysics* vol. 55 Preprint at

https://doi.org/10.1017/S0033583522000117

36. Sriram KK, Persson F (2024) and F. J. and B. J. P. and T. J. O. and W. F. Fluorescence Microscopy of Nanochannel-Confined DNA. in *Single Molecule Analysis: Methods and Protocols* (ed. Heller Iddo and Dulin, D. and P. E. J. G.) 175–202Springer US, New York, NY. 10.1007/978-1-0716-3377-9_9
37. Nyblom M et al (2023) Strain-level bacterial typing directly from patient samples using optical DNA mapping. Commun Med 3
38. Dvirnas A, Lin Y-L (2021) Ildev.v.0.5.3. Preprint at https://doi.org/10.5281/zenodo.5718208
39. Yeh M et al (2016) C.-C. Barcelona,. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets. in *IEEE 16th International Conference on Data Mining (ICDM)* 1317–1322 10.1109/ICDM.2016.0179

Figures

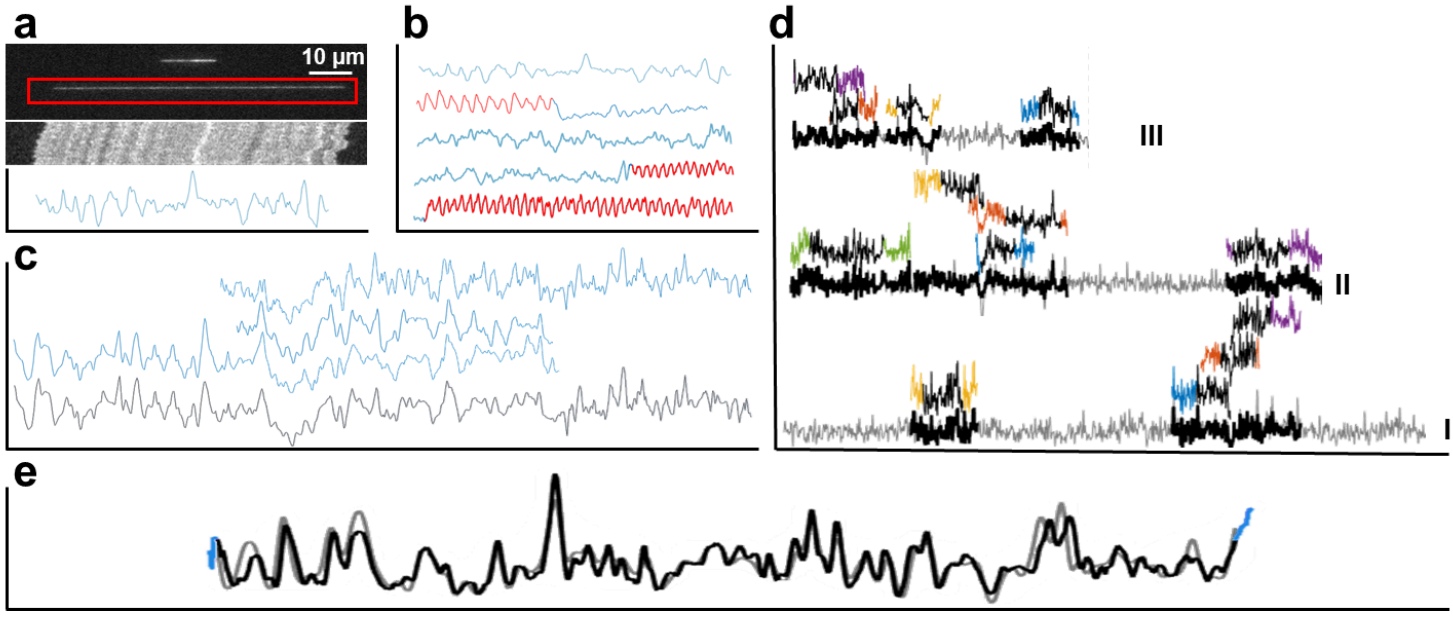


Figure 1

Workflow overview. a CB-labeled DNA molecules are stretched in nanochannels, imaged and detected (red box) to generate continuous fluorescence intensity profiles, referred to as barcodes. **b** Extracted barcodes undergo quality control and repetitive region (red) detection. **c**Multi-step assembly with the DOGMA pipeline generates barcode contigs (bargroups, gray) from individual barcodes (blue). **d** Bargroups (colored) are aligned to reference genomes (gray) for validation and structural comparison. Roman numerals indicate chromosome numbers. **e** Consensus (black) is generated from bargroups and aligned to each reference chromosome (gray). Segments present in the reference but not in the consensus are highlighted in yellow and segments present in the consensus but not in the reference in blue.

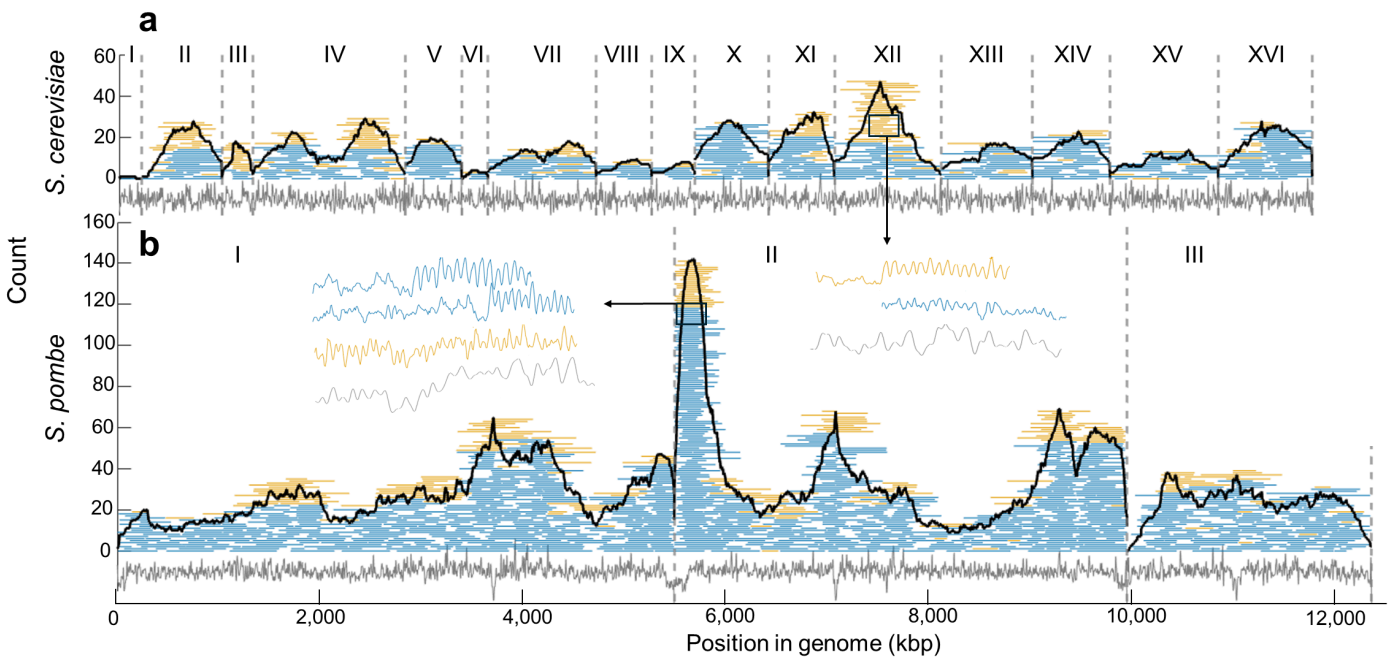


Figure 2

Alignment to reference genomes. Alignment of all barcodes to the reference genomes of **a** *S. cerevisiae BY4742* and **b***S. pombe 972h-*, with chromosomes indicated as Roman numerals. Matches to the correct species are represented in blue and to the incorrect species in yellow. The theoretical optical maps are shown in gray. Black lines represent coverage depth. Insets highlight the most enriched genomic loci for each organism.

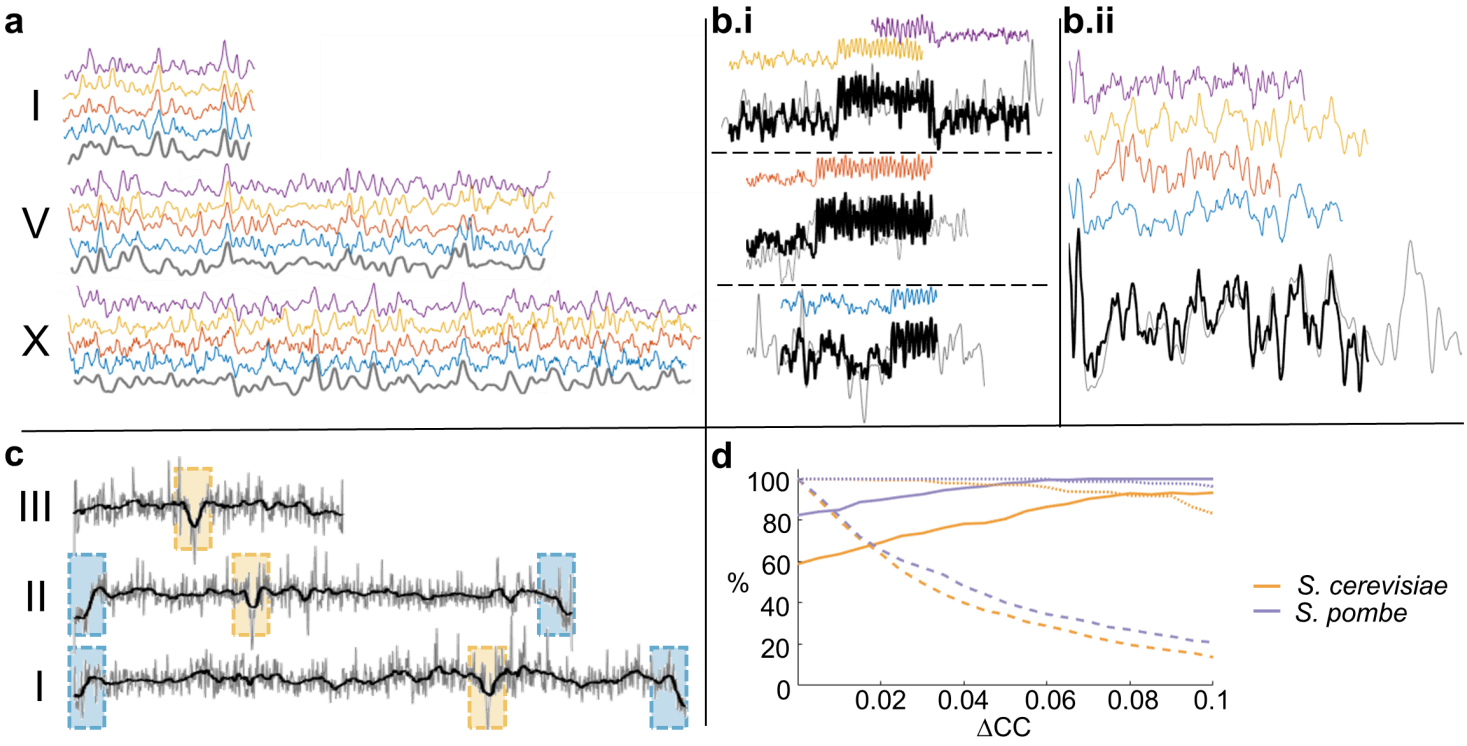


Figure 3

Challenges in eukaryotic dense OGM assembly. a Representative barcodes (colored) spanning entire reference chromosomes (gray) in *S. cerevisiae*. **b.i, b.ii** Examples of partially repetitive barcodes (colored) aligned to the reference genome (gray) before **b.i** and after **b.ii** masking of repetitive regions. The consensus barcode of each alignment is shown in black for comparison. **c** Theoretical barcodes for the three *S. pombe* chromosomes (in gray) and their smoothened trends (black). Orange boxes mark centromeres, showing conserved long-range low-intensity motifs, and blue boxes indicate regions with intensity profile gradients, due to extended GC-content gradients toward chromosome ends. **d** Percentage of barcodes aligning to the correct organism (solid lines), percentage of retained barcodes (dashed lines) and percentage of estimated coverage (dotted lines) after using a local barcode-to-reference alignment of repeat-masked barcodes under varying thresholds of ΔCC (difference in CC-value between the first- and second-best match for each molecule) for *S. cerevisiae (orange)* and *S. pombe (purple)*.

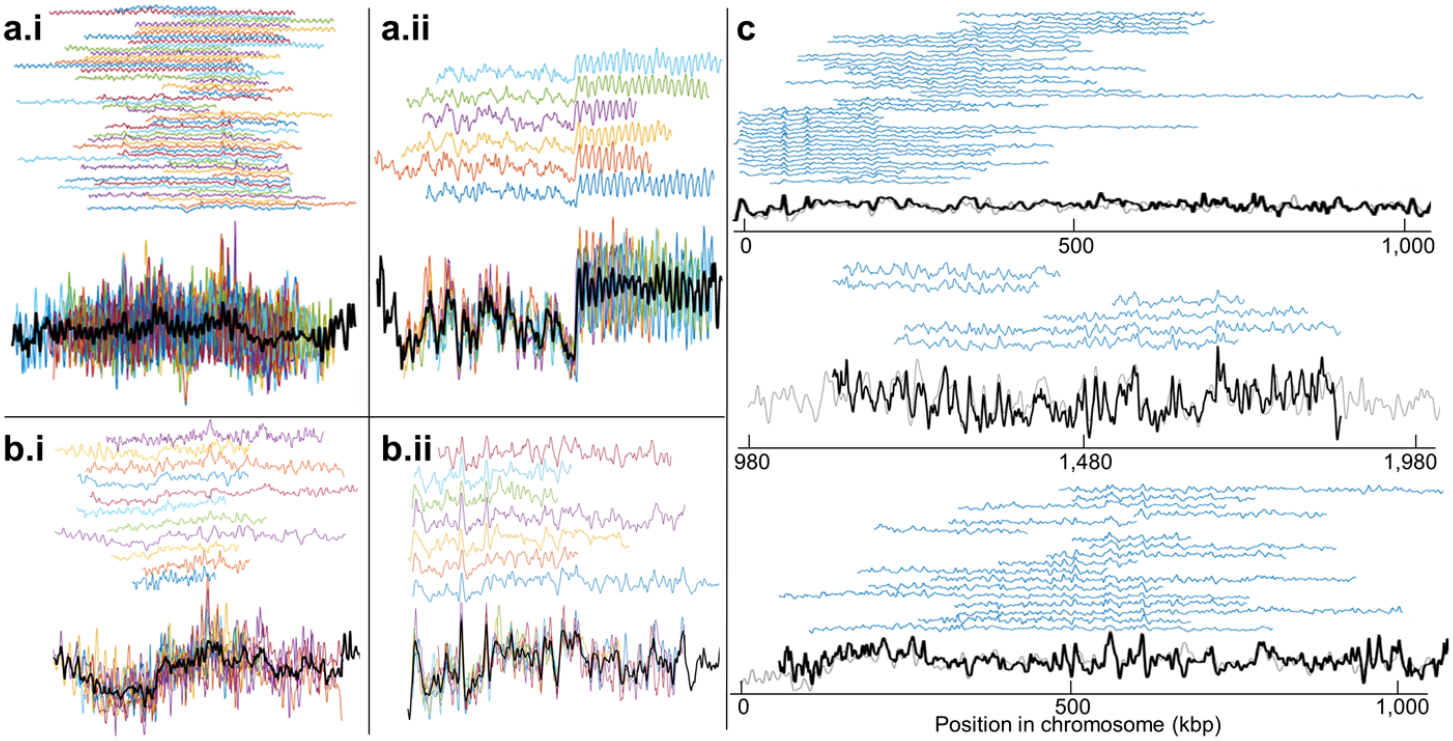


Figure 4

Assembly of bargroups. Example bargroup assembled **a.i** without masking and **a.ii** after masking of repetitive regions. Constituent barcodes are shown in color and the resulting bargroup consensus in black. **b**Examples of bargroups illustrating different signal-to-noise characteristics. **b.i**A bargroup that fails (SNR=0.9) the SNR criterion and **b.ii** a bargroup that passes (SNR=1.4) the SNR criterion. **c** Example of assembled bargroups aligned to the reference genome. Consensus bargroups (black) constructed from multiple barcode members (blue) are aligned post-assembly to their best matching regions along

the theoretical reference barcodes (gray) in chromosomes I, I and II (from top to bottom respectively of *S. pombe*).

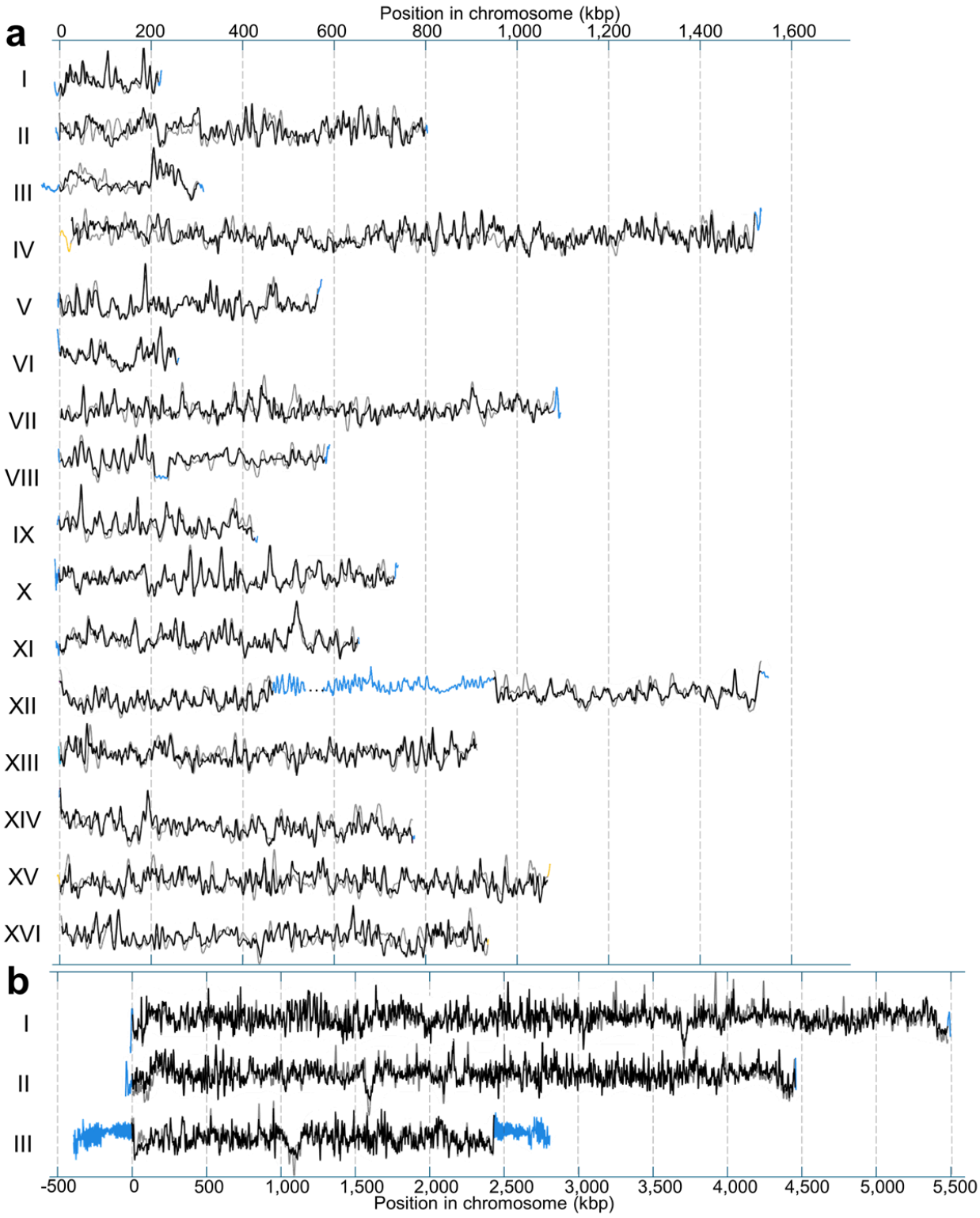


Figure 5

Full genome assemblies. a All assembled *S. cerevisiae*chromosomes (black) aligned to the reference (gray). **b** All assembled *S. pombe* chromosomes (black) aligned to the reference (gray). Chromosomal

extensions relative to the reference are highlighted in blue and regions not covered by the assembly are highlighted in yellow. **a** and **b** have different size scales for their x-axes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- YeastassemblySIsubmission.docx
- SupplementaryDatacode.zip