eLIFE
elifesciences.org

Figures and figure supplements

Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms

Martial Marbouty, et al.

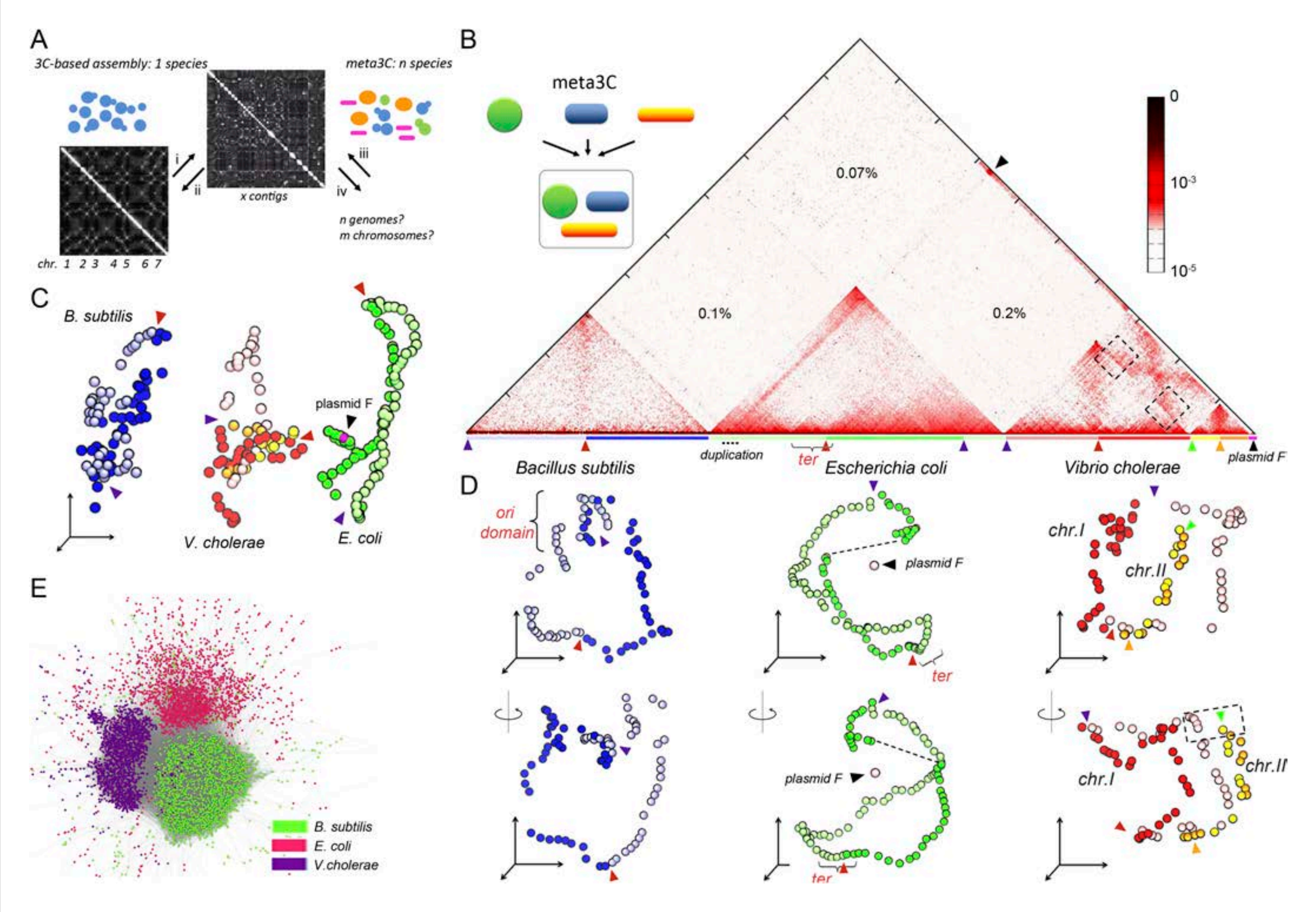
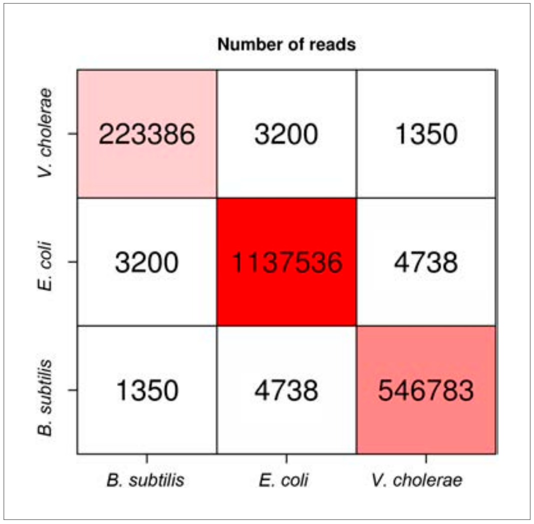


Figure 1. meta3C experiment on a controlled mix of bacterial species. (**A**) Schematic representation of the principle of a meta3C experiment. For a single species, one can (i) generate a genome-wide contact map but also (ii) use the genomic contact data to reorder the contact matrix of a poorly assembled genome (top) in order to scaffold its contigs into a more likely structure (bottom left). For a mixture of species, one can similarly generate a contact map directly from the mix (iii) and then use it to characterize the genomes of the species contained in the mixture (iv). (**B**) Chromosomal contact map of a mixture of three bacteria. The darker the shade in the matrix, the higher the contact frequency (color scale in log). Blue and green arrowheads: origins of replication of the chromosomes. Red and orange arrowheads: termination of replication. Each chromosome arm is represented with a color code. The percentages of interactions that occurred between different species are indicated within the matrix. We detected frequent interactions between a F plasmid sequence (black arrowhead) and a large duplicated region of the *E. coli* chromosome (dotted line). (**C**) 3D reconstruction of the entire contact matrix. Each replichore of the four chromosomes is represented with a different color, with the colored arrowheads indicating *Ori* positions. Black arrowhead: F′ plasmid and *E. coli* chromosome when contacts with the duplication are taken into account; see ***Figure 1—figure supplement 2*** for more details. (**D**) Two different views of the 3D reconstruction of each of the three bacteria taken individually. *Ter* and *Ori* are indicated with colored arrowheads (see above). *E. coli* and *B. subtilis* domains are also shown. The F′ plasmid is positioned according to its 3D contacts with the *E. coli* genome, without taking into account the duplication (dotted line within the *E. coli* genome). (**E**) Networks of interactions between the different contigs for the mix of three bacteria represented with a force-directed graph-drawing algorithm. Each node represents a contig from the de novo assembly. Each link in gray represents at least one 3C contact. Each color corresponds to one community detected by the Louvain algorithm.
DOI: 10.7554/eLife.03318.003



Number of reads

	B. subtilis	E. coli	V. cholerae
V. cholerae	223386	3200	1350
E. coli	3200	1137536	4738
B. subtilis	1350	4738	546783

Figure 1—figure supplement 1. Numbers of intra-specific and inter-specific (chimeric) pairs of reads from the meta3C experiment performed on the bacterial mix.

DOI: 10.7554/eLife.03318.004

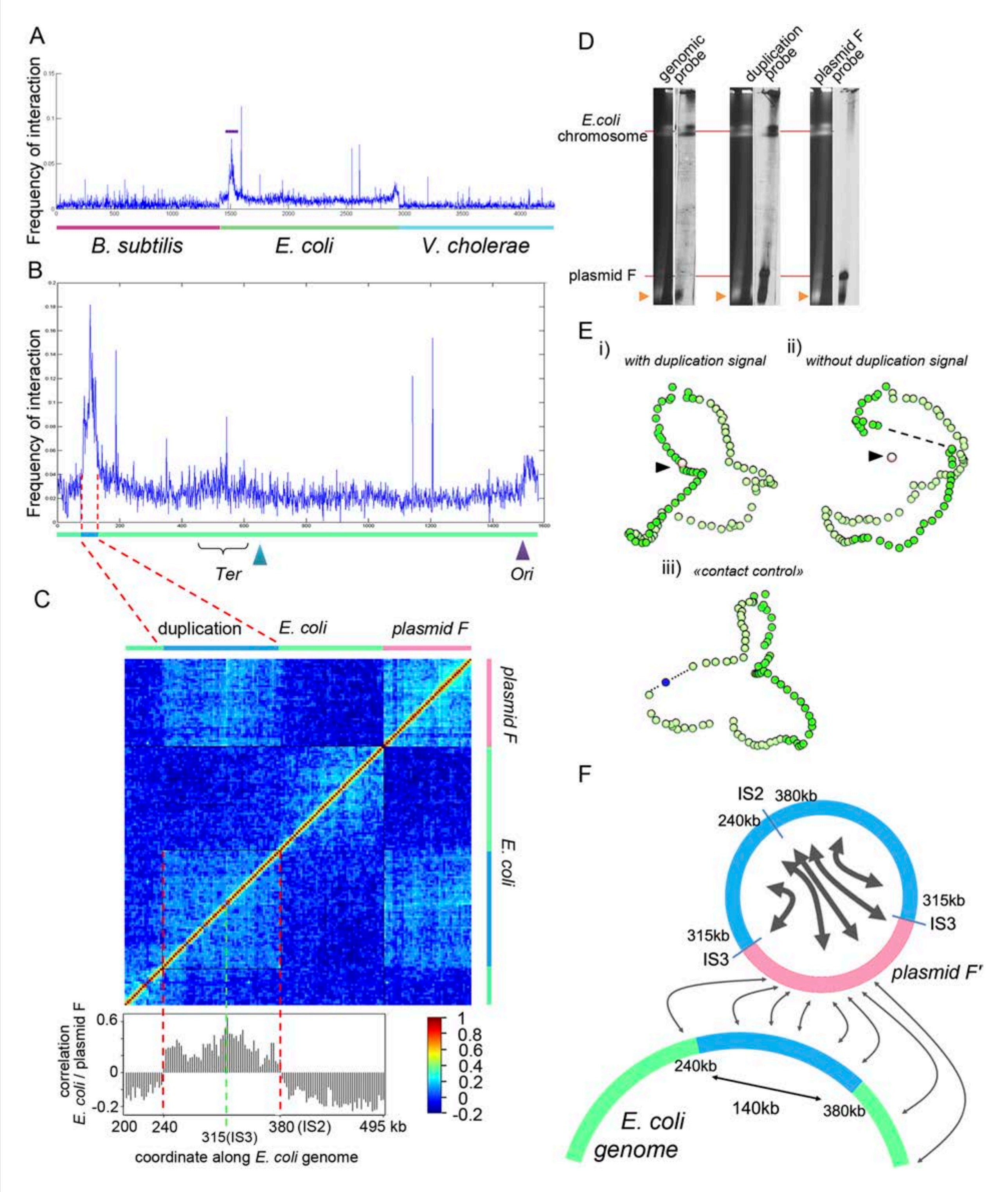


Figure 1—figure supplement 2. Analysis of the organization of the F′ plasmid in the *E. coli* strain used in this study. (**A**) Frequencies of interaction between the F′ plasmid and the three bacterial genomes. (**B**) Contact frequencies between the F′ plasmid and the *E. coli* genome, normalized by the read coverage to take into account the replication. The dotted region encompassing a duplication with frequent contacts with the F′ plasmid corresponds to the region of interest analyzed in **C**. (**C**) Correlation analysis of the contact frequency between the F′ plasmid (in pink) and the duplicated region (in blue) along the *E. coli* genome (in green). A strong shift in the correlation score occurred for a region indicated with the red dotted lines, indicating that the plasmid F′ is in frequent contact with this region but not beyond it (and therefore must be carrying one copy of the duplicated

Figure 1—figure supplement 2. Continued on next page

Figure 1—figure supplement 2. Continued

region). This analysis allowed to position duplication breakpoints at coordinates ~240,000:380,000 along the chromosome. (**D**) PFGE analysis of the *E. coli* genome and corresponding Southern blots hybridized with probes from the genome (duplicated and non-duplicated region) and from the F′ plasmid (see ***Table 2*** for the coordinates of the probe). Orange triangle: migration front of the degraded DNAs of both *E. coli* genome and plasmid. (**E**) 3D reconstruction of genome-wide contact maps of the *E. coli* genome when (i) the duplicated region is considered as a single copy region of *E. coli* genome, (ii) the duplicated region is removed from the analysis, but not the plasmid F′, and (iii) a region of a size similar to the duplicated region is removed from the analysis and replaced with a small DNA segment of the same size as the F′ plasmid. (**F**) Schematic representation of the contacts between the F plasmid sequence (pink segment) and duplicated segments (blue) of the *E. coli* genome (green) that accounts for all the observations in panels **A**–**E**.

DOI: 10.7554/eLife.03318.005

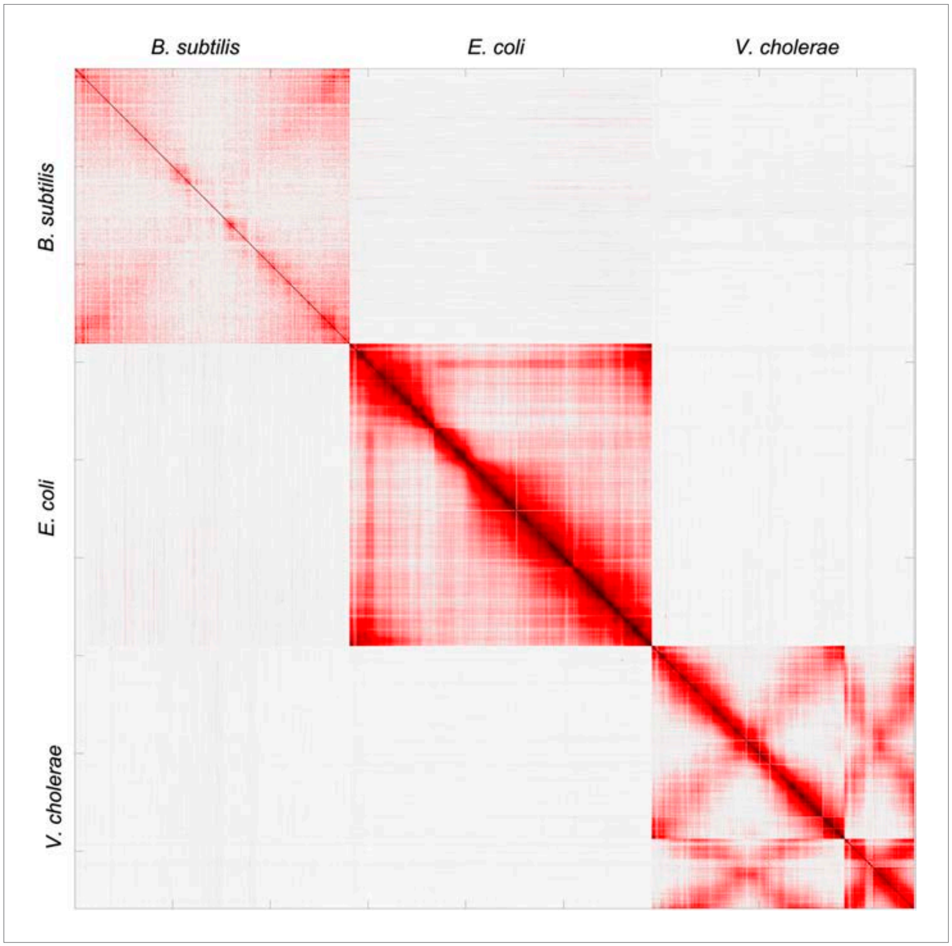


Figure 1—figure supplement 3. Pearson correlation matrix of the meta3C bacterial experiment.

DOI: 10.7554/eLife.03318.006

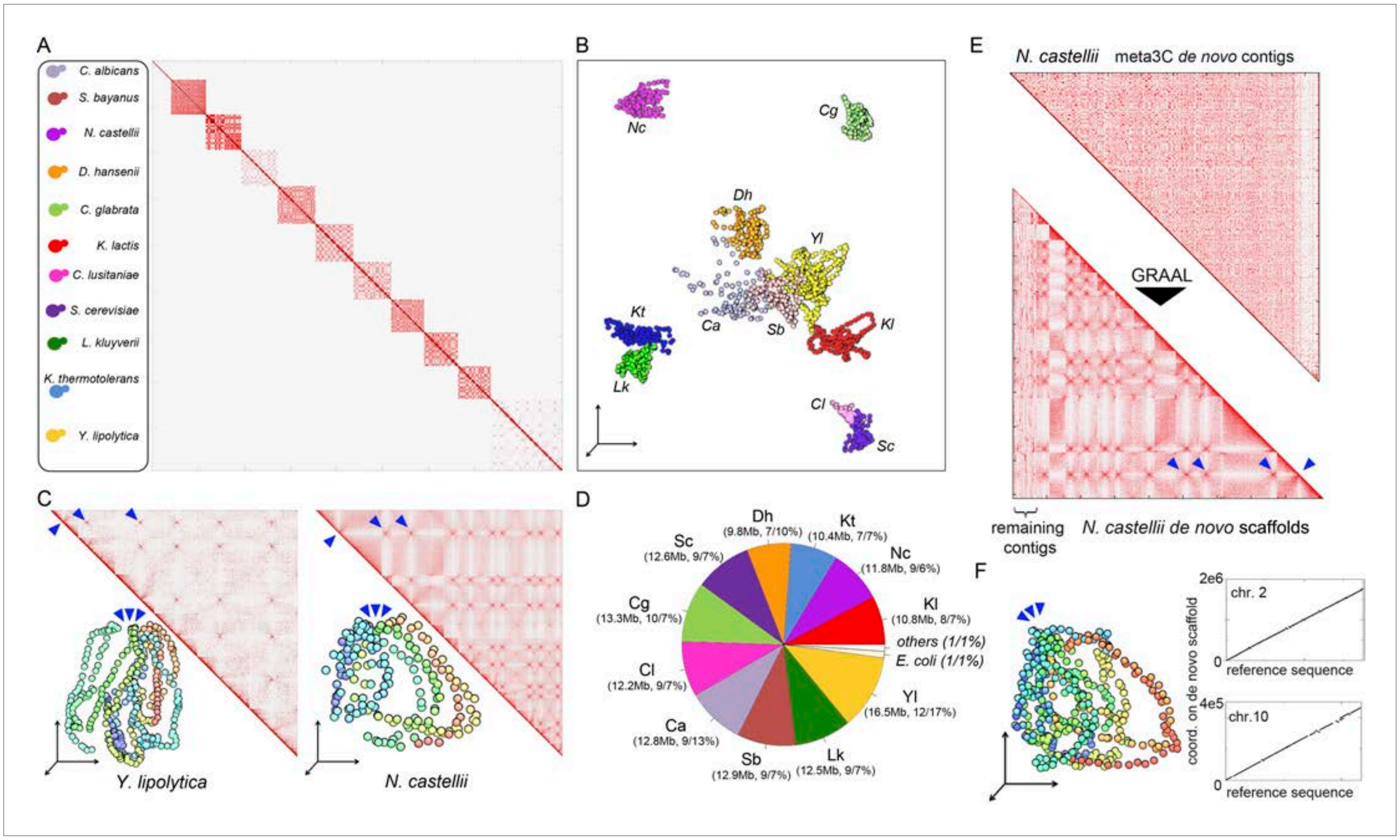
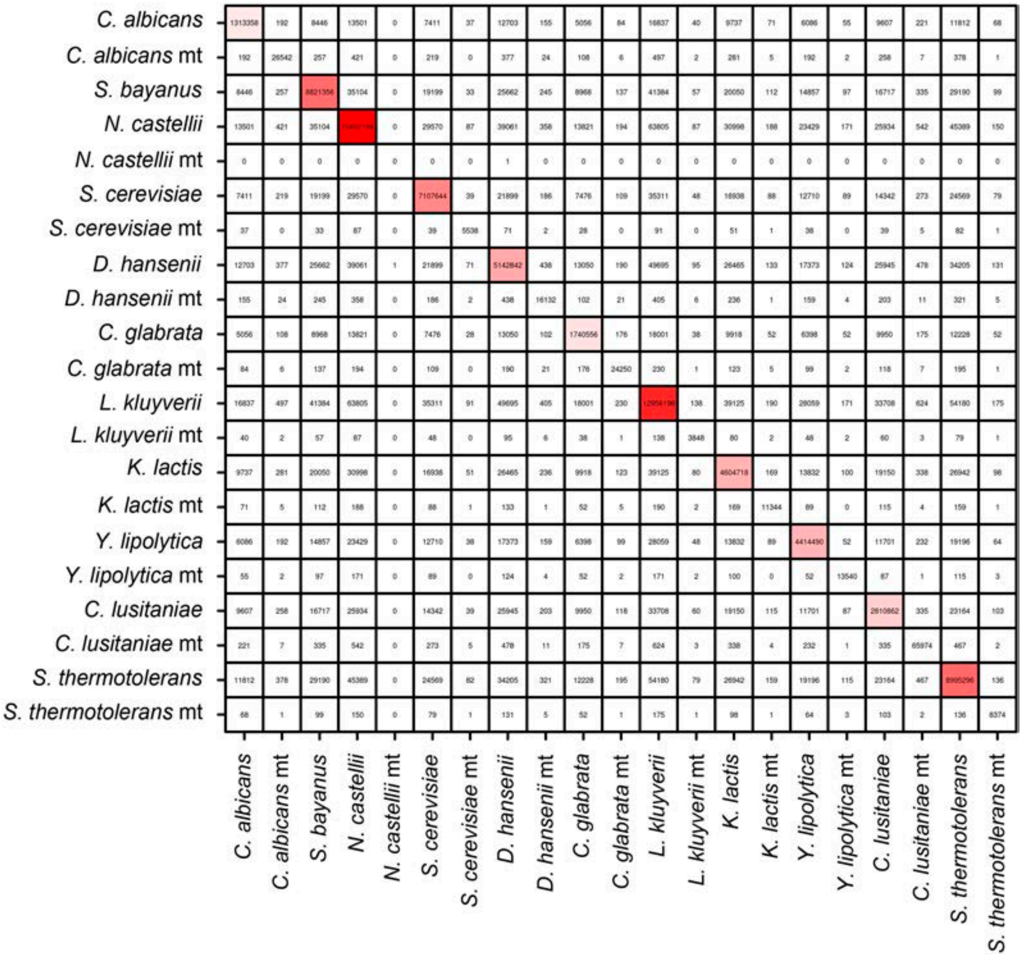


Figure 2. meta3C experiment on a controlled mix of yeasts species. (**A**) meta3C contact map of the mix of eleven species. (**B**) 2D projection of the 3D reconstruction of the entire meta3C contact matrix. Each genome occupies an isolated position in space (the 2D projection induces a visual overlap for some species). The color code is the same as for the schematic yeasts on the left panel in (**A**). (**C**) Close-up on the contact maps of three species, with the 3D representation of the matrix in *vis-à-vis*. (**D**) Quantification of the assembly performed using meta3C reads. The first number indicates the amount of total DNA in the community. The two numbers that follow indicates the proportion of the contigs of each community in regards to the total assembly (% of total kb, % of total number of contigs). (**E**) Top: contact map of the contigs present within the community containing mostly sequences from *N. castellii*. The bottom contact map corresponds to the maps recovered following the GRAAL scaffolding. 11 large scaffolds were retrieved, in near-perfect agreement with the known number of chromosomes of this species. Blue triangles: inter-scaffold signal corresponding most likely to the 10 centromeric interactions (***Marie-Nelly et al., 2014b***). (**F**) Corresponding 3D structure of the *N. castellii* de novo meta3C assembly combined with GRAAL processing. The collinearity between two of the resulting superscaffolds and their corresponding reference chromosome sequences is represented in the right panels.

DOI: 10.7554/eLife.03318.013



	C. albicans	*C. albicans* mt	*S. bayanus*	*N. castellii*	*N. castellii* mt	*S. cerevisiae*	*S. cerevisiae* mt	*D. hansenii*	*D. hansenii* mt	*C. glabrata*	*C. glabrata* mt	*L. kluyverii*	*L. kluyverii* mt	*K. lactis*	*K. lactis* mt	*Y. lipolytica*	*Y. lipolytica* mt	*C. lusitaniae*	*C. lusitaniae* mt	*S. thermotolerans*	*S. thermotolerans* mt
C. albicans	1313358	192	8446	13501	0	7411	37	12703	155	5056	84	16837	40	9737	71	6086	55	9607	221	11812	68
C. albicans mt	192	26542	257	421	0	219	0	377	24	108	6	497	2	281	5	192	2	258	7	378	1
S. bayanus	8446	257	8821356	35104	0	19199	33	25662	245	8968	137	41384	57	20050	112	14857	97	16717	335	29190	99
N. castellii	13501	421	35104	[illegible]	0	29570	87	39061	358	13821	194	63805	87	30998	188	23429	171	25934	542	45389	150
N. castellii mt	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
S. cerevisiae	7411	219	19199	29570	0	7107644	39	21899	186	7476	109	35311	48	16938	88	12710	89	14342	273	24569	79
S. cerevisiae mt	37	0	33	87	0	39	5538	71	2	28	0	91	0	51	1	38	0	39	5	82	1
D. hansenii	12703	377	25662	39061	1	21899	71	5142842	438	13050	190	49695	95	26465	133	17373	124	25945	478	34205	131
D. hansenii mt	155	24	245	358	0	186	2	438	16132	102	21	405	6	236	1	159	4	203	11	321	5
C. glabrata	5056	108	8968	13821	0	7476	28	13050	102	1740556	176	18001	38	9918	52	6398	52	9950	175	12228	52
C. glabrata mt	84	6	137	194	0	109	0	190	21	176	24250	230	1	123	5	99	2	118	7	195	1
L. kluyverii	16837	497	41384	63805	0	35311	91	49695	405	18001	230	12956196	138	39125	190	28059	171	33708	624	54180	175
L. kluyverii mt	40	2	57	87	0	48	0	95	6	38	1	138	3848	80	2	48	2	60	3	79	1
K. lactis	9737	281	20050	30998	0	16938	51	26465	236	9918	123	39125	80	4604718	169	13832	100	19150	338	26942	98
K. lactis mt	71	5	112	188	0	88	1	133	1	52	5	190	2	169	11344	89	0	115	4	159	1
Y. lipolytica	6086	192	14857	23429	0	12710	38	17373	159	6398	99	28059	48	13832	89	4414490	52	11701	232	19196	64
Y. lipolytica mt	55	2	97	171	0	89	0	124	4	52	2	171	2	100	0	52	13540	87	1	115	3
C. lusitaniae	9607	258	16717	25934	0	14342	39	25945	203	9950	118	33708	60	19150	115	11701	87	2810862	335	23164	103
C. lusitaniae mt	221	7	335	542	0	273	5	478	11	175	7	624	3	338	4	232	1	335	65974	467	2
S. thermotolerans	11812	378	29190	45389	0	24569	82	34205	321	12228	195	54180	79	26942	159	19196	115	23164	467	8995296	136
S. thermotolerans mt	68	1	99	150	0	79	1	131	5	52	1	175	1	98	1	64	3	103	2	136	8374

Figure 2—figure supplement 1. Number of intra-specific and inter-specific (chimeric) pairs of reads in the meta3C contact map. Mitochondrial genomes behaved as separate entities.

DOI: 10.7554/eLife.03318.014

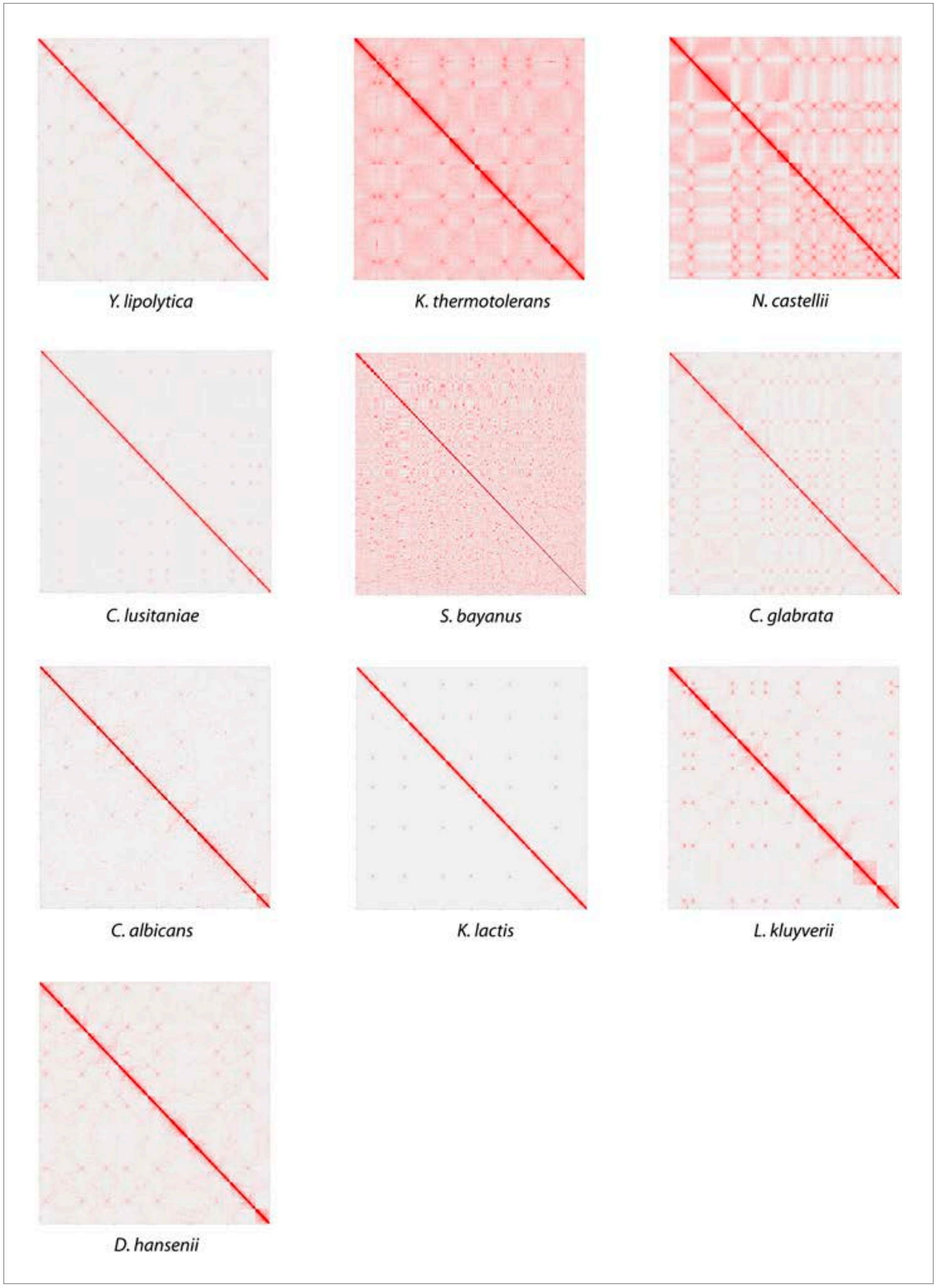


Figure 2—figure supplement 2. Contact matrices of the genomes of *Y. lipolytica*, *K. thermotolerans*, *N. castellii*, *C. lusitaniae*, *S. bayanus*, *C. glabrata*, *C. albicans*, *K. lactis*, *L. kluyveri*, and *D. hansenii*.

DOI: 10.7554/eLife.03318.015

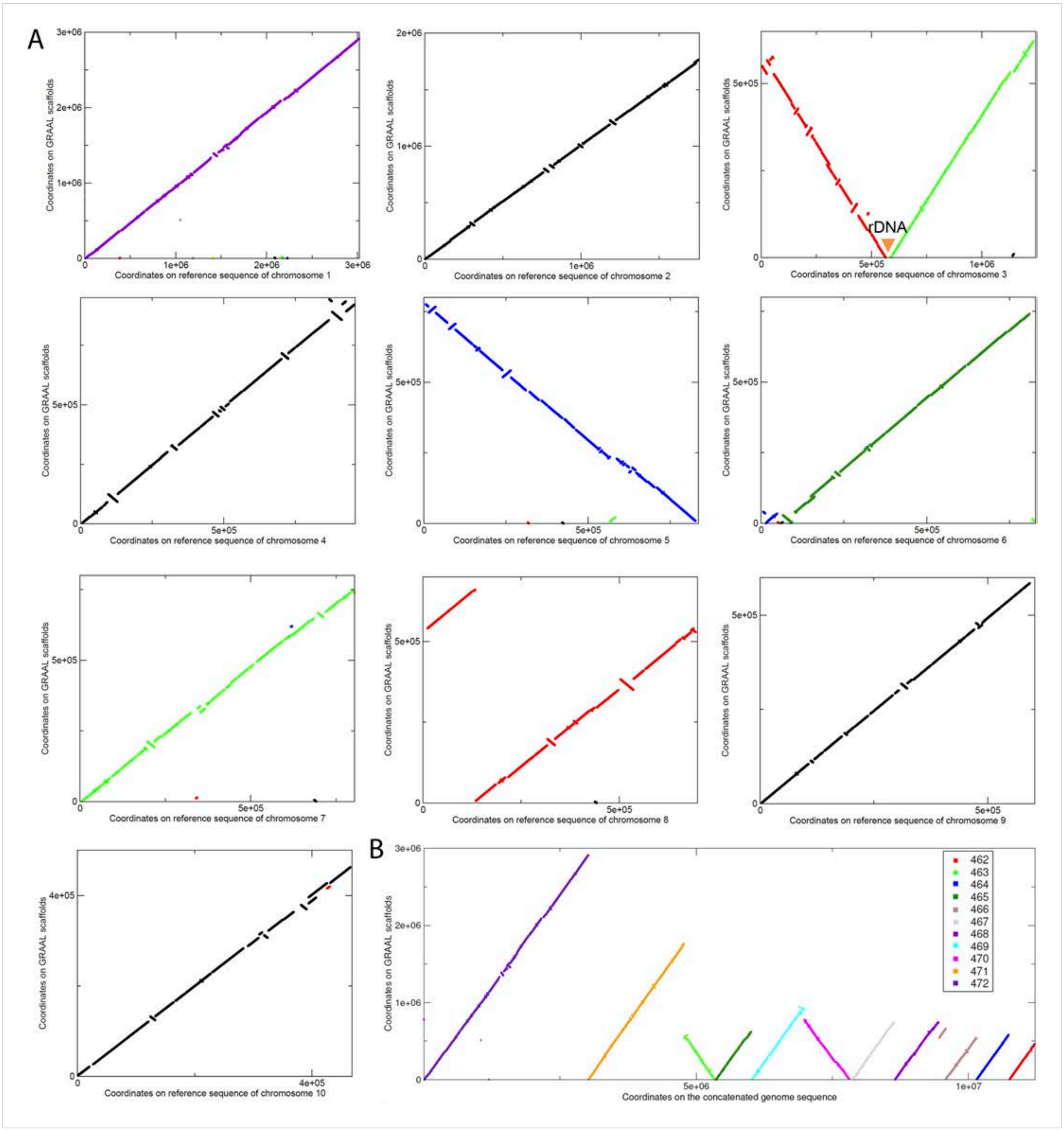


Figure 2—figure supplement 3. Scaffolding using GRAAL of *N. castellii* meta3C contigs. (**A**) Comparison between the scaffolds generated by GRAAL and the reference sequences of the *N. castellii* chromosomes: y-axis: coordinates along the new scaffolds. x-axis: coordinates along the reference chromosome. The position of the rDNA cluster is indicated. (**B**) The eleven largest scaffolds are aligned against the concatenated reference genome, covering 94.5% of the total sequence.

DOI: 10.7554/eLife.03318.016

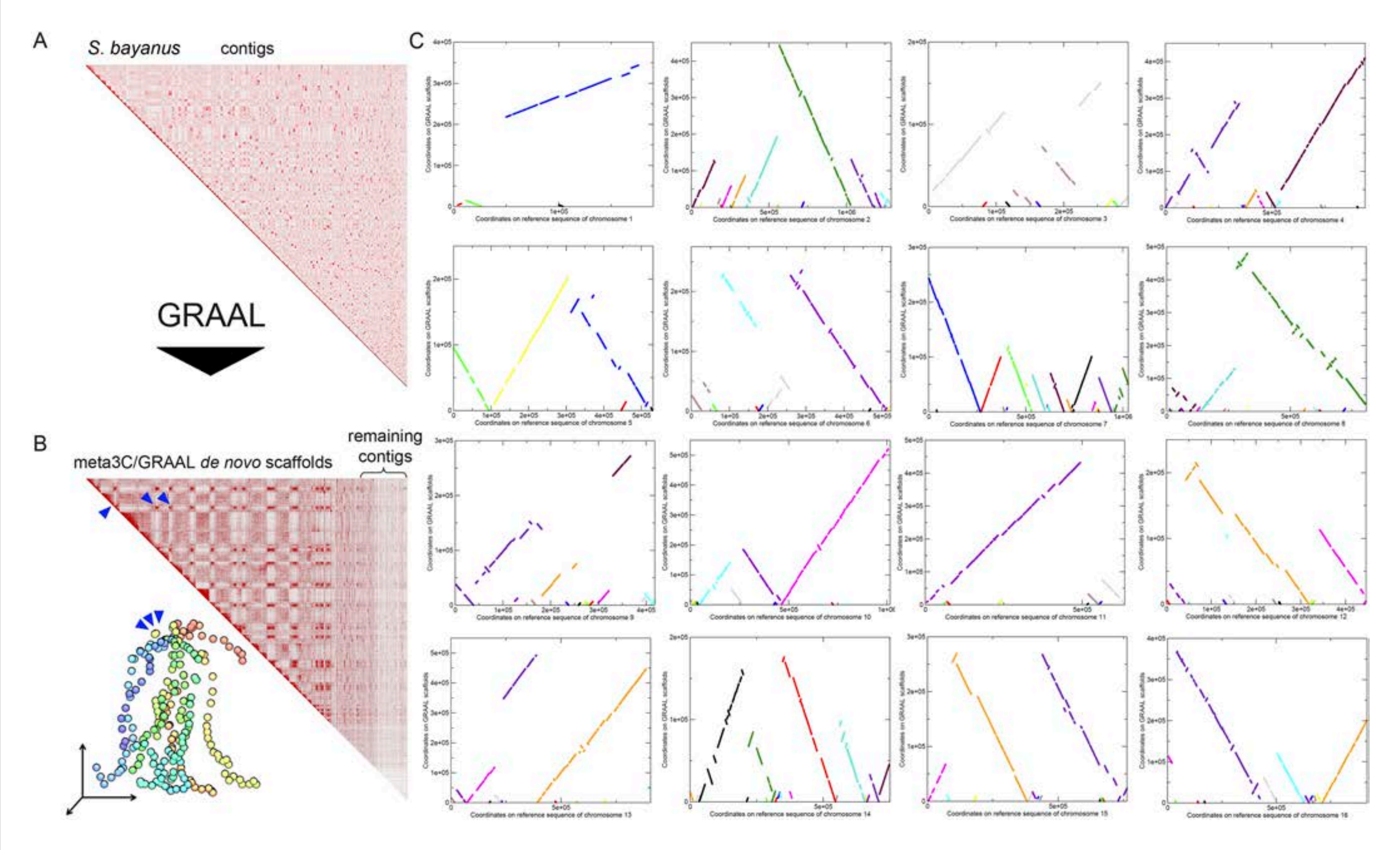


Figure 2—figure supplement 4. Scaffolding using GRAAL of *S. bayanus* meta3C contigs. (**A**) Contact map obtained from a draft genome assembly of *S. bayanus*. (**B**) Contact map and corresponding 3D structure of the *S. bayanus* de novo meta3C assembly combined with GRAAL processing. Blue triangles: inter-scaffold signal corresponding most likely to the 10 centromeric interactions. (**C**) Comparison between the scaffolds generated by GRAAL and the reference sequences of *S. bayanus* chromosomes. The main scaffold is indicated in bold, whereas the other ones correspond to other scaffolds and are shown here for information only.

DOI: 10.7554/eLife.03318.017

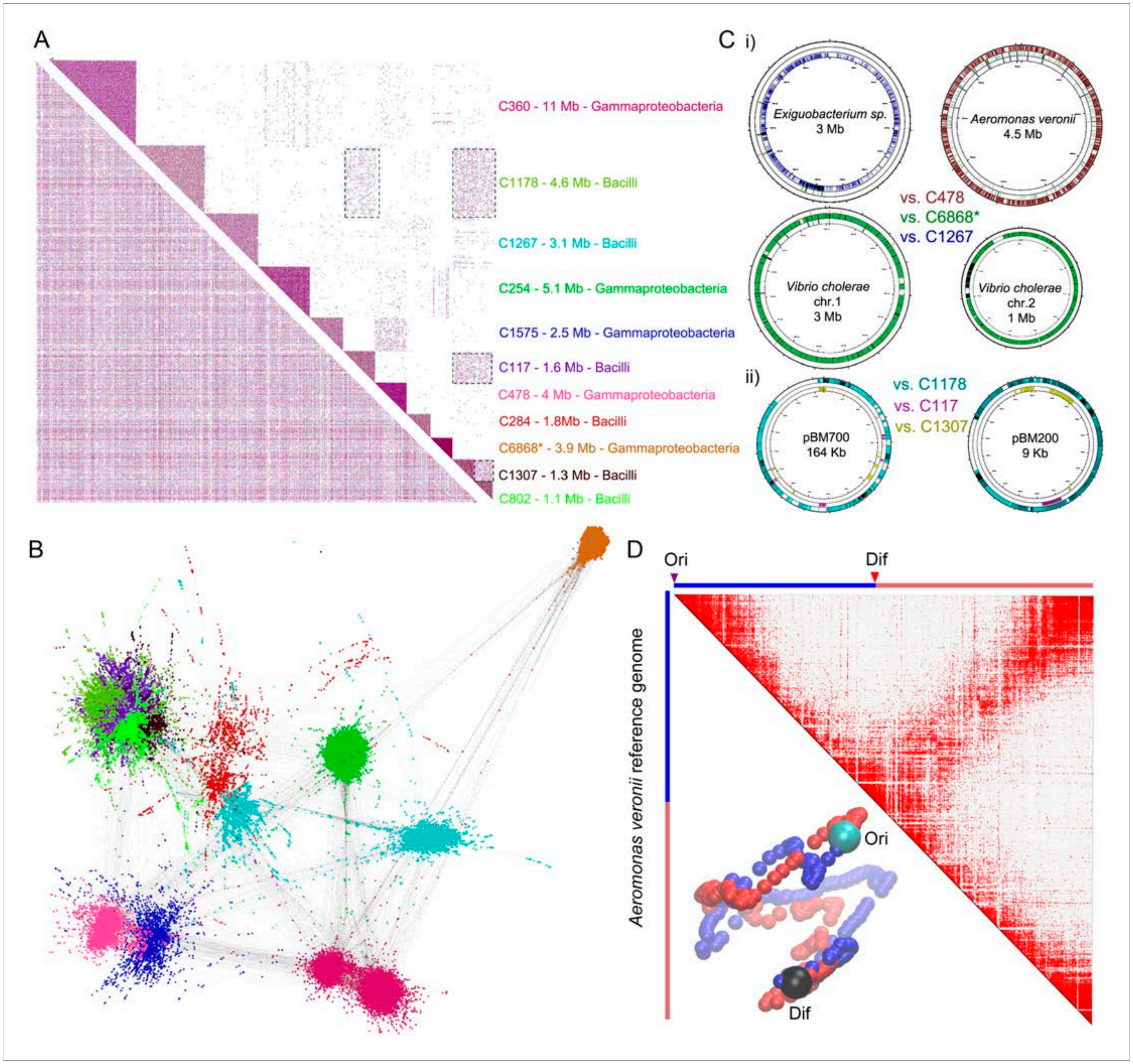


Figure 3. meta3C analysis of a complex environmental sample. (**A**) meta3C contact map of the largest 11 communities of contigs in the matrix before (bottom left) and after (upper right) clustering. Each square corresponds to a community grouping contigs that exhibit significantly more contact with each other than with other communities. For each community, an index, the sum of the sequences, and candidate classes are indicated on the right side of the matrix. Dotted square underlines inter-community contact enrichments. C6868*: *V. cholerae*. (**B**) Illustration of the interactions between the 11 largest communities of contigs using a force-directed graph drawing algorithm Force Atlas 2 (***Jacomy et al., 2014***). Each node corresponds to a contig (or a chunk of a contig) and each link represents at least one meta3C interaction. The colors correspond to the communities identified by the Louvain algorithm and described in **A**. (**C**) i: For three communities, the contigs were mapped against the closest reference genomes identified (Exiguobacterium sp. AT1b, *A. veronii* B565, *V. cholerae* N16961, plasmids pBM700 and pBM200). To illustrate the specificity of the approach, each community was mapped against each of those genomes (the color code is the same as in **A** and the order from the outer circle to the inner circle is indicated in the middle). ii: Similarly, the three Bacilli communities highlighted in color were mapped against two plasmid sequences. (**D**) Genomic contact map of an unknown species (most likely *A. veronii* or a close relative) generated by mapping the reads present in the community 478 against the genome of *A. veronii* (bin size = 10 kb). The corresponding 3D structure is indicated next to the contact map.

DOI: 10.7554/eLife.03318.023

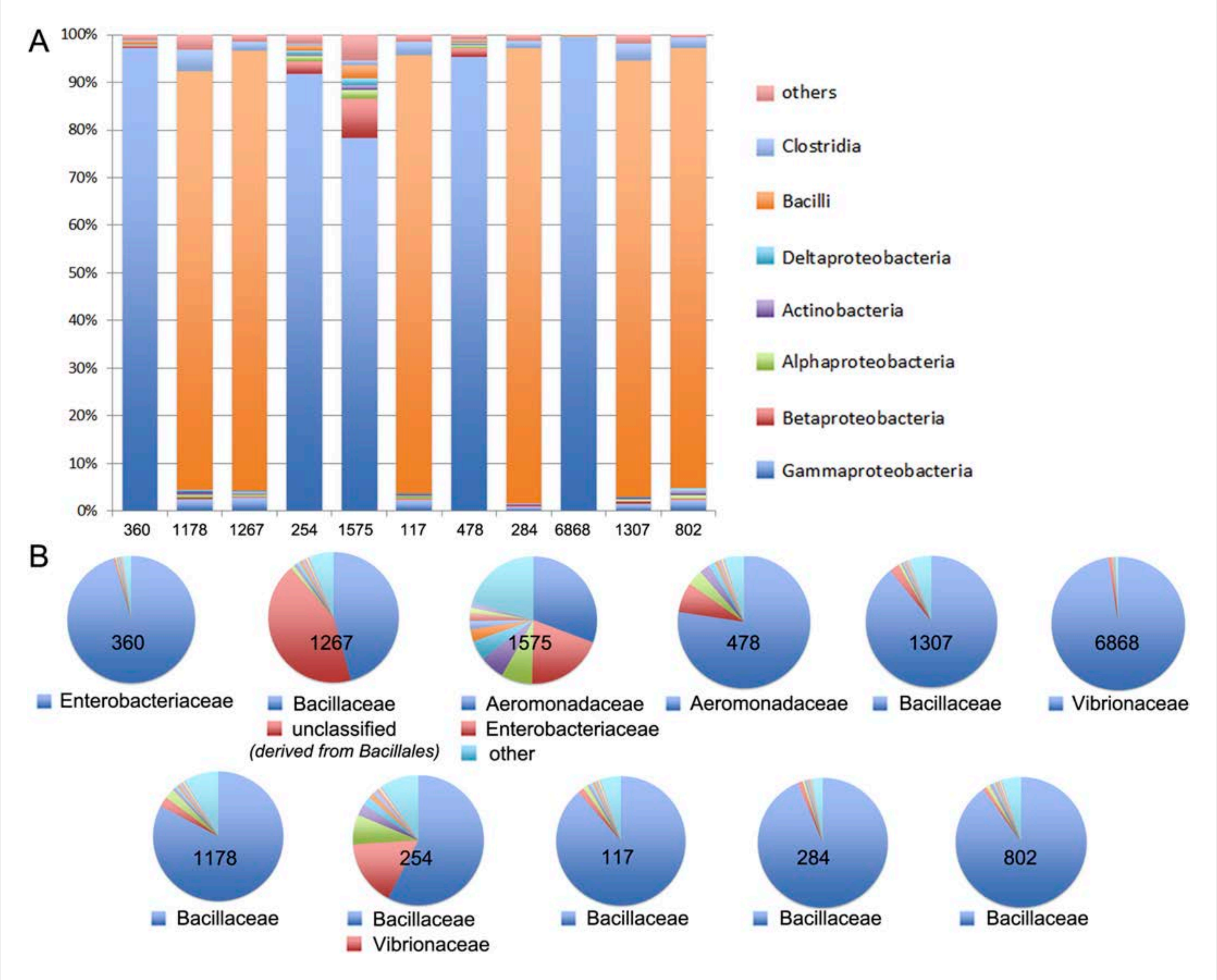


Figure 3—figure supplement 1. MG-RAST analysis of the 11 largest meta3C communities. The similarity between the sequences present in these communities and those of known species was to annotate them. (**A**) Homogeneity regarding class-level annotations; (**B**) homogeneity regarding family-level taxonomy.

DOI: 10.7554/eLife.03318.024

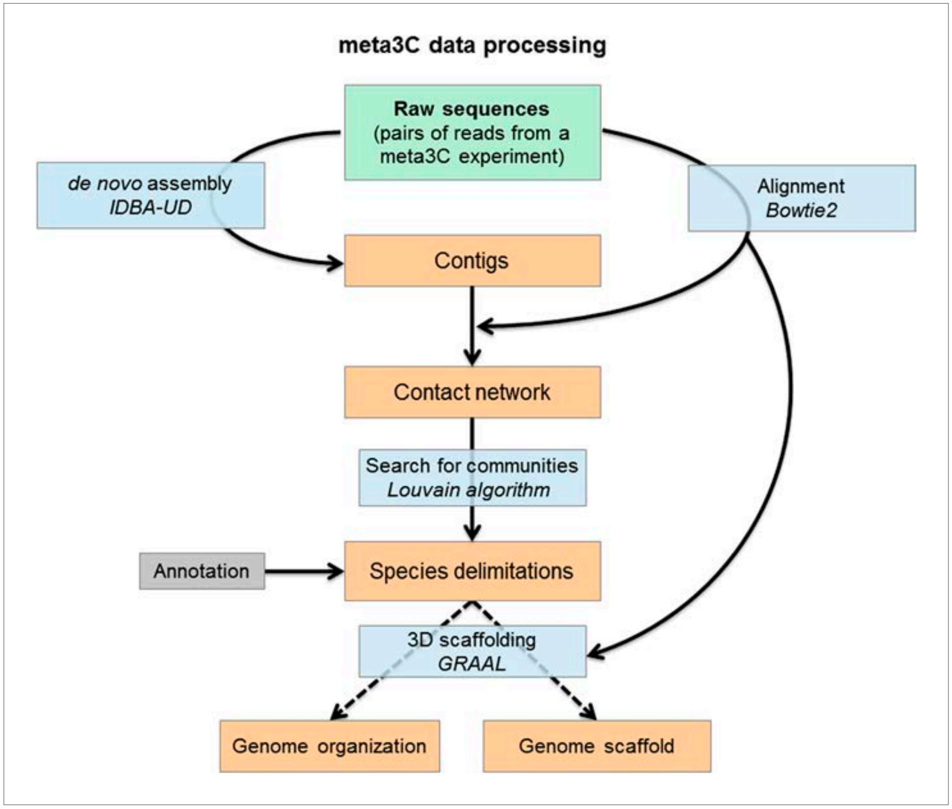


Figure 4. Flowchart representing the computational analysis steps of a meta3C experiment. First, the reads from the sequenced meta3C library are assembled into contigs. The meta3C contact information between the contigs is then used to generate a network of the contacts of all contigs against each other. This contact network is searched for so-called 'communities' (by analogy with social network analysis) using the Louvain algorithm. The significant communities can be annotated by NCBI and correspond principally to sequences from individual species. Ultimately, the contact information can be used to reorder the DNA segments within each community and thus generate the 3D contact map of the corresponding species. In this process contigs can be reassembled by software such as GRAAL, thereby decreasing the percentage of chimeric fragments in the assembly and improving its continuity. DNA segments originating from other species are put aside automatically during this process given their lack of 3C contacts with the rest of the genome under reassembly.

DOI: 10.7554/eLife.03318.026