Hicberg : Hi-C based Biological Estimation of Repeated elements in Genomes

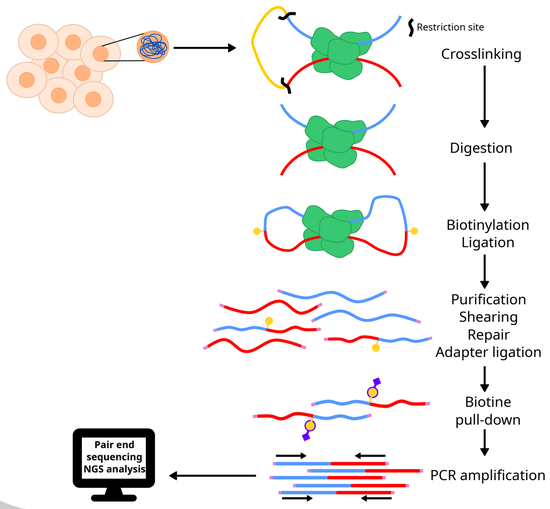
S. Gradit [1,2], R. Koszul [1], A. Cournac [1]

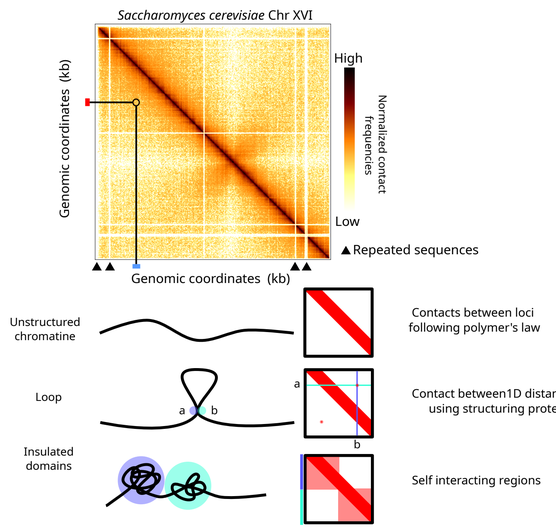
[1] Institut Pasteur, Unité Régulation Spatiale des Génomes, CNRS UMR 3525, 75015 Paris, France.
[2] Ecole Doctorale Complexité du Vivant, Sorbonne Université, ED 515, 75005 Paris, France.

During their evolution, the genomes of micro-organisms can acquire quantities of different repeated elements such as retrotransposons, duplicated genes or tandem repeats. This type of sequence within genomes cannot be processed directly by NGS technologies because they generate short reads that cannot be located unambiguously on reference genomes. This information is filtered out by most current pipelines, leading to incomplete genomics tracks resulting in a significant loss of information on biological functions, processes and genomic structures involving repeated elements. To overcome these limitations we developped Hicberg, an alogorithm that uses statistical inference and pseudo-random generators to predict the positions of repeated sequences' reads from different omics paired-end data (including Hi-C, Mnase-seq, ChIP-seq, ...), based multiple components relative to the polymeric behavior of the DNA and sequencing protocoles' features, estblished on the non unmbiguous part of the tracks. Thus read pairs belonging to repeated sequences can be assigned with robust confidence in genomes filling-in genomic tracks. After development and calibration on a controled test bench Hicberg improves genomic data interpretability of various species, starting with microbial one such as *Saccharomyces cerevisiae.*

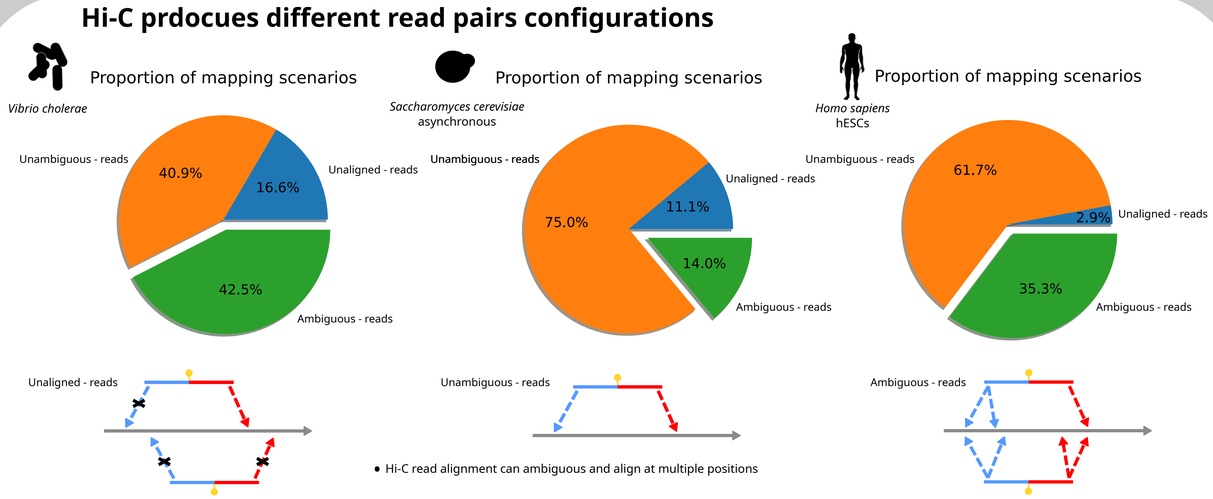
Reconstructions of Hi-C and ChIP-seq genomic tracks with Hicberg revealed how some retrotransposons in this model contribute to the positioning of cohesin, a molecular motor involved in the formation of chromatin loops. A new role of retrotransposon sequences as contact hot points for the elusive yeast 2 micron episomal molecule was also identified.Overall, these results underline the power of the approach to discover new novel molecular relationships, and the interest in applying this tool more widely to larger genomes with greater quantityof repeats. The proposed method can therefore provide an alternative visualization of genomic signals in a wide variety of biological conditions and allow a more comprehensive view of genome organization and plasticity. Importantly, existing dataset can be revisited using this approach to unveil overlook features.

Hi-C experiment



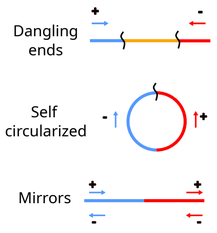


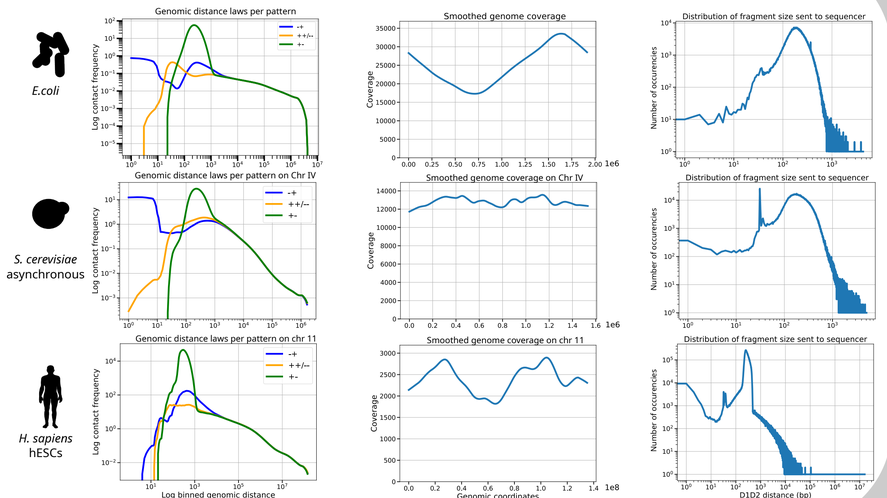
Hi-C prdocues different read pairs configurations



- Hi-C read alignment can ambiguous and align at multiple positions
- Ambiguous reads lead to a loss of information, resulting in white stripes on Hi-C map and a loss of folding information

Leveraging pair-end features to reduce ambiguity



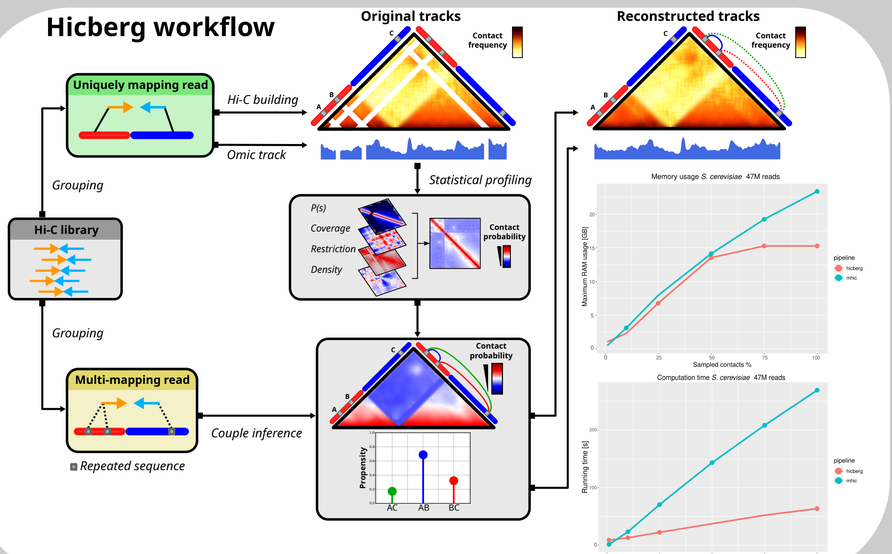


The various Hi-C patterns show variation in P(s) providing discriminant power in the assignement of reads involved in repeated regions

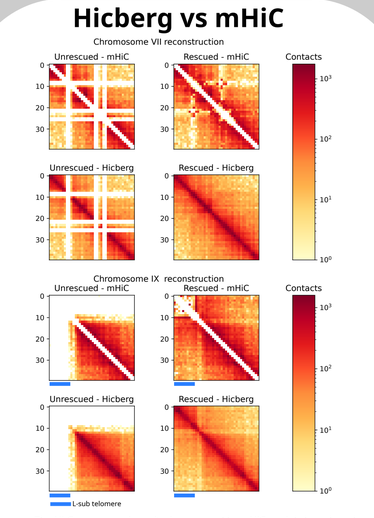
Genomic coverage also offers discriminating power for ambiguous contacts assignation

Distances to the closest restriction site for Hi-C reads can help for the assgination of ambiguous contacts especially for old protocols using 6bp-cutter

Hicberg workflow

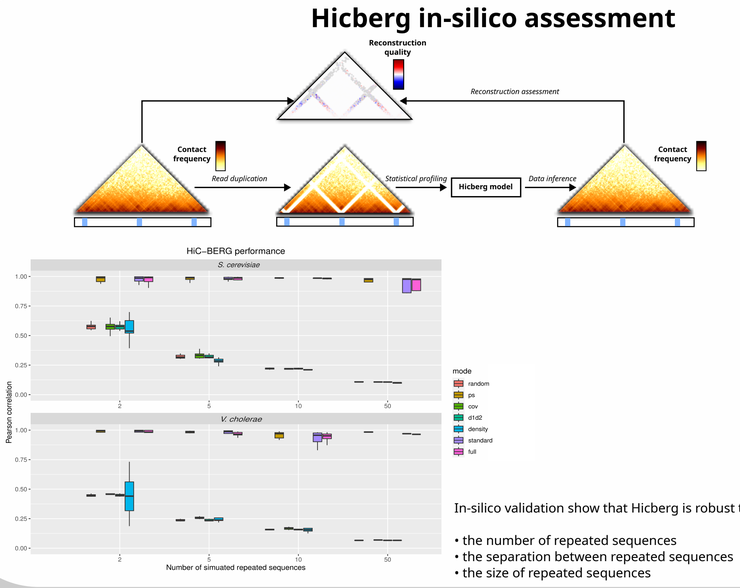


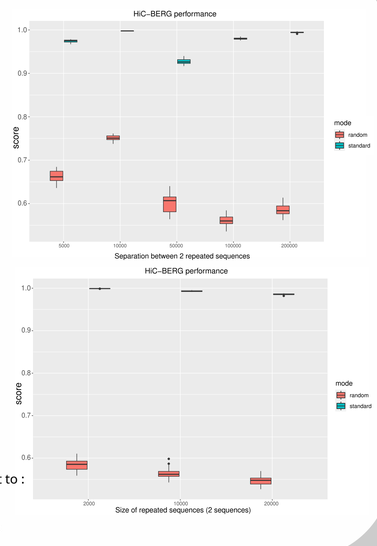
Hicberg vs mHiC



The Hi-C reconstruction criteria proposed by mHiC, mainly based on the otherwise stringent P(s), lead to the appearance of less well-defined structures and artifactual structures.
Hicberg proposes better-defined, non-artifactual structures.

Hicberg in-silico assessment

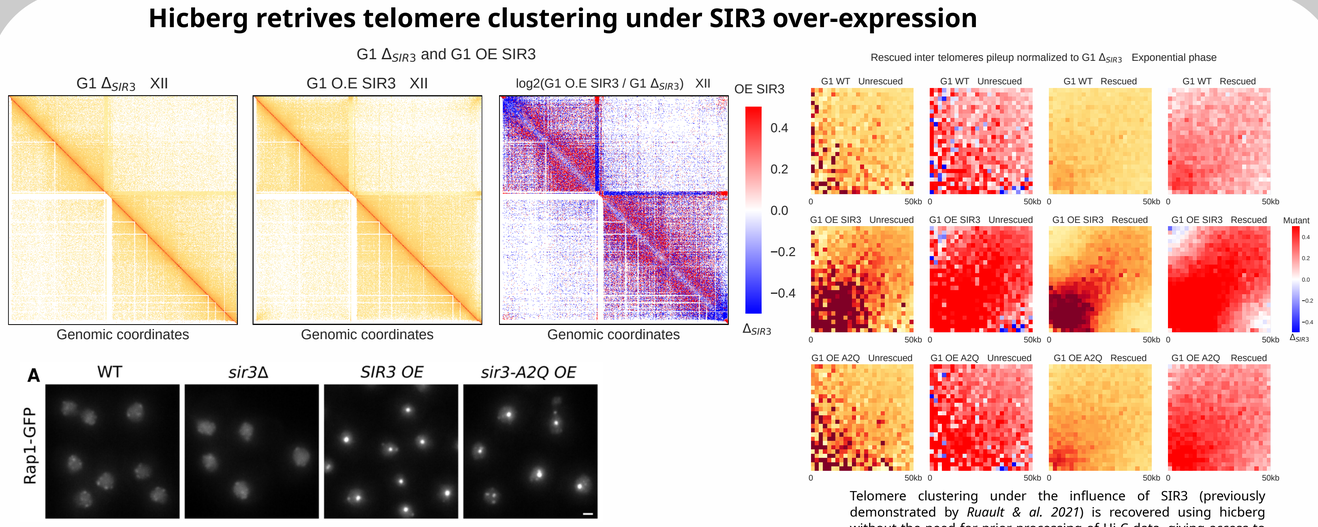




In-silico validation show that Hicberg is robust to :

- the number of repeated sequences
- the separation between repeated sequences
- the size of repeated sequences

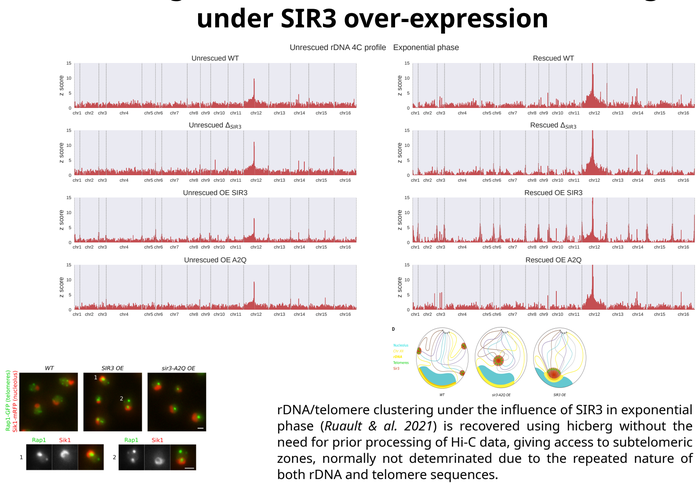
Hicberg retrives telomere clustering under SIR3 over-expression



Data :Ruault. et al. Sir3 mediates long-range chromosome interactions in budding yeast. 2021.

Telomere clustering under the influence of SIR3 (previously demonstrated by *Ruault & al. 2021*) is recovered using hicberg without the need for prior processing of Hi-C data, giving access to subtelomeric zones, normally not detemrinated due to the repeated nature of the sequences.

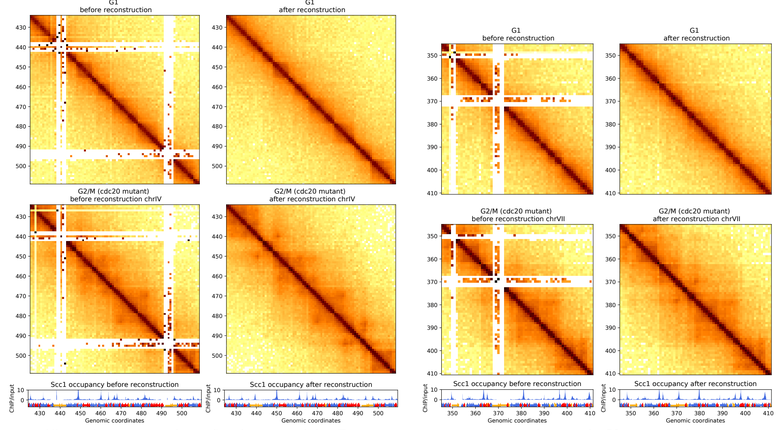
Hicberg retrives rDNA - telomeres clustering under SIR3 over-expression



rDNA/telomere clustering under the influence of SIR3 in exponential phase (*Ruault & al. 2021*) is recovered using hicberg without the need for prior processing of Hi-C data, giving access to subtelomeric zones, normally not detemrinated due to the repeated nature of both rDNA and telomere sequences.

Data : Ruault and al. , Sir3 mediates long-range chromosome interactions in budding. 2021

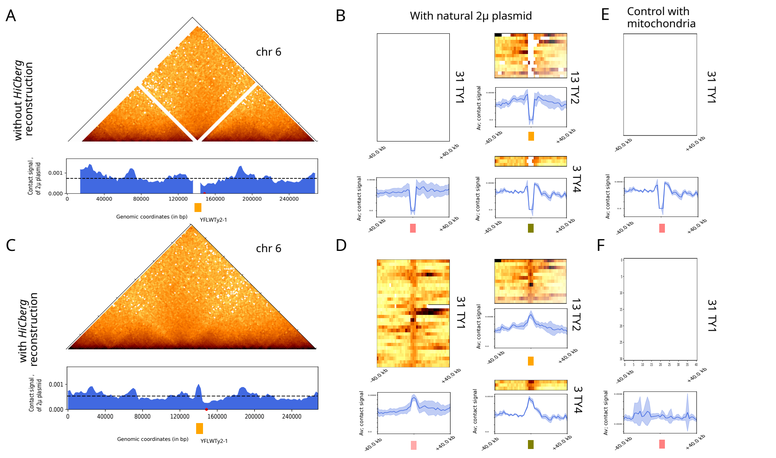
Hicberg retrives coherent chromatin loops along cell cycle



Hicberg's reconstructions of regions containing Tys reveal the presence of cohesin loops in G2 phase, which are absent in G1 as previously demonstrated.
Furthermore, ChIP-seq SCC1 reconstructions reveal cohesin placement at converging genes adjacent to Tys.

Data : Piazza A, Bordelet H. and al. Cohesin regulates homology search during recombinational DNA repair. 2021

Tys as hot spots for 2µ plasmid anchoring



As previously described 2-micron plasmid preferentially contact the yeast genome on low transcribed regions
Hicberg revealed that Tys are preferred contact sites for 2-micron plasmid as they are low level transcribed regions

Conclusion & Perspectives

Hicberg paves the way to the exploration of functional 3D organization of structures involving repeated elements such as *V. cholerae* super integron, *S. cerevisiae* transposable elements, *P. falciparum* telomeric regions, etc.

Currently applied on genomes of ~ 100Mb in size carrying variable repeated sequences involved in pathogenicity (*V. dahliae*, *E. festucae*)

Hicberg allow working with more complete set of data that will also improve the quality of the already visible part of genomes giving more evenly distributed signals at the genome scale

Hicberg may improve Hi-C normalization procedure and specific patterns detections such as peaks (1D signals) and loops/domains (Hi-C).

References

[1] Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002 Feb 15;295[5558]:1306–11.

[2] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326[5950]:289–93.

[3] Cournac A, Koszul R, Mozziconacci J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. Nucleic acids research. 2016;44[1]:245–255.

[4] Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nature methods. 2017;14[7]:679.

[5] Cournac A, Marbouty M, Mozziconacci J, Koszul R. Generation and analysis of chromosomal contact maps of yeast species. Yeast Functional Genomics: Methods and Protocols. 2016;227–245.