Hicberg : Prediction of omics signals from repeated elements

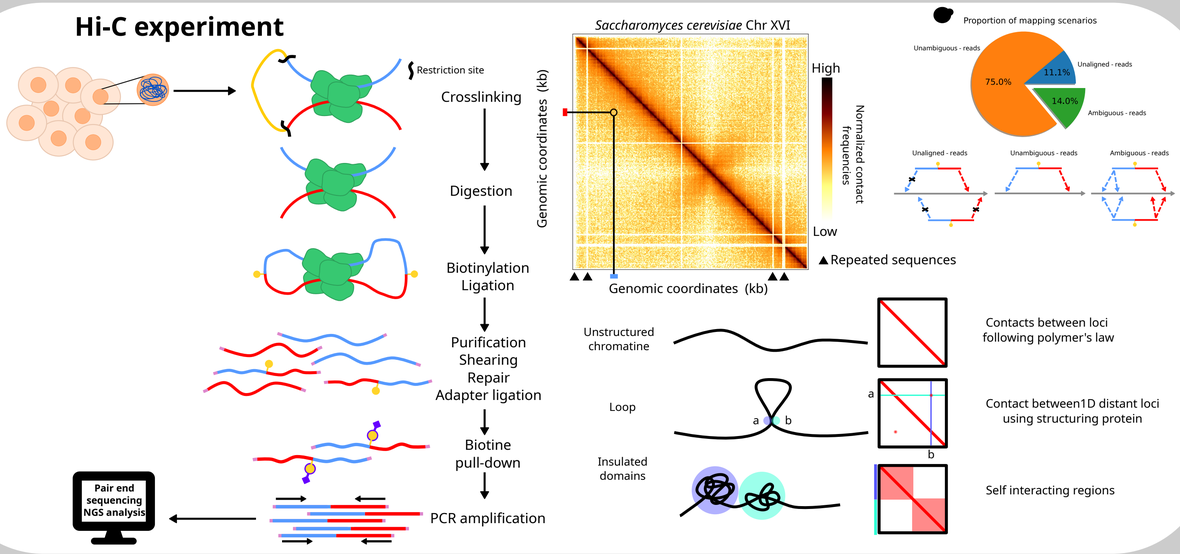
S. Gradit [1, 2], P. Larrous [1], A. Thierry [1], R. Koszul [1], A. Cournac [1]

[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.

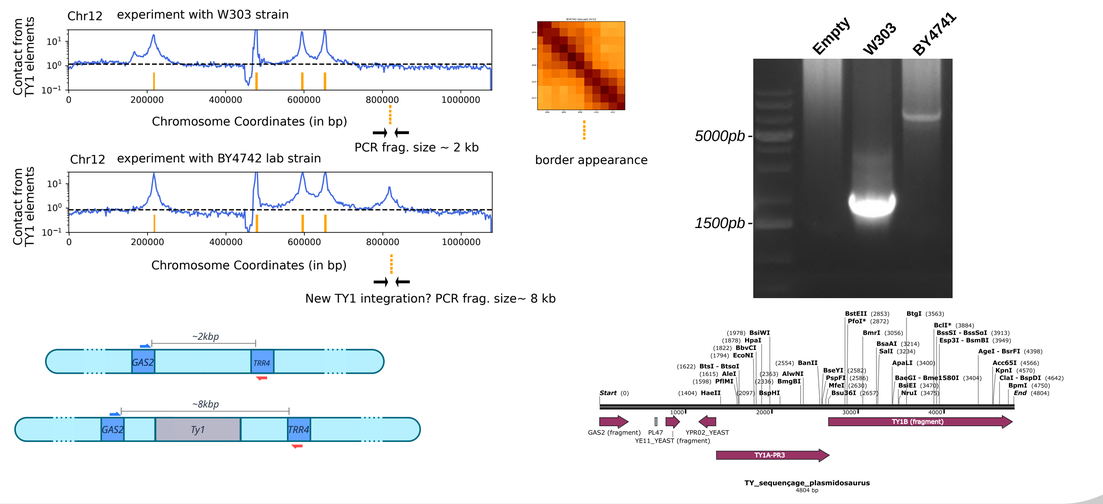
During their evolution, the genomes of micro-organisms can acquire quantities of different repeated elements such as retrotransposons, duplicated genes or tandem repeats. This type of sequence within genomes cannot be processed directly by NGS technologies because they generate short reads that cannot 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 developed Hicberg, an algorithm 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, ...). The model is based on multiple components relative to the polymeric behavior of the DNA and sequencing protocols' features, established on the unambiguous 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 controlled 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 quantity of 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.

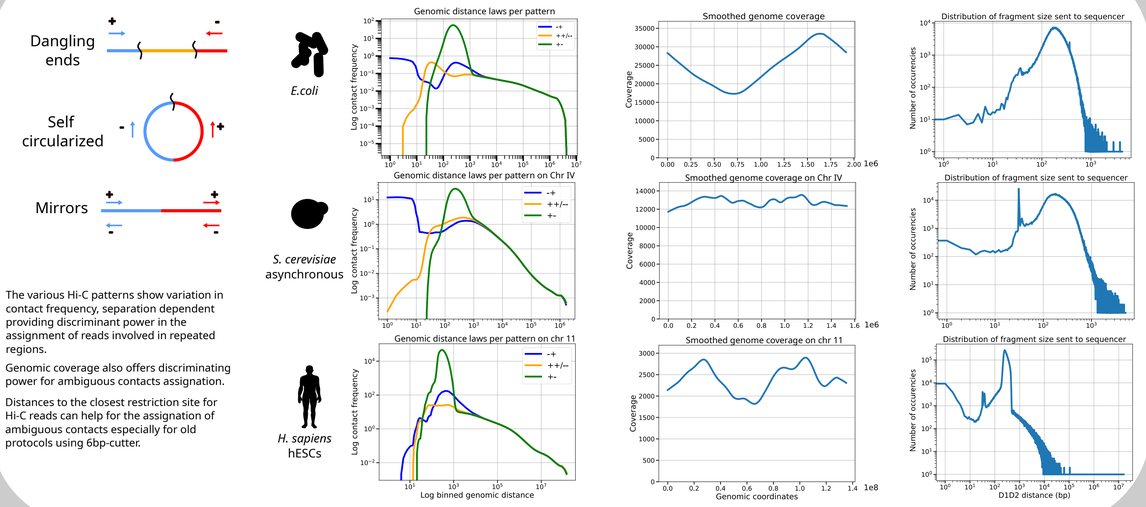
Hi-C experiment



Bench-top comparison of signals generated by Hicberg for the detection of retrotransposons



Leveraging pair-end features to reduce mapping ambiguity

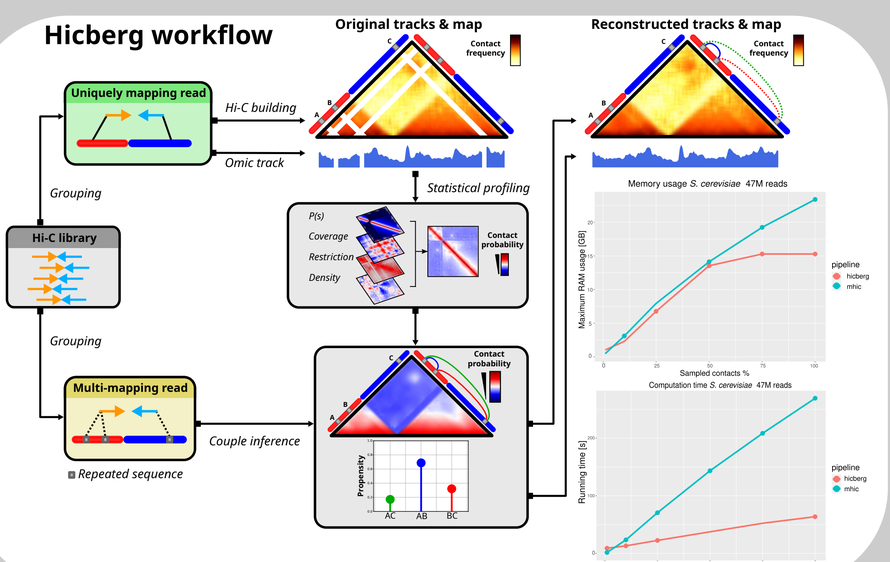


The various Hi-C patterns show variation in contact frequency, separation dependent providing discriminant power in the assignment of reads involved in repeated regions.

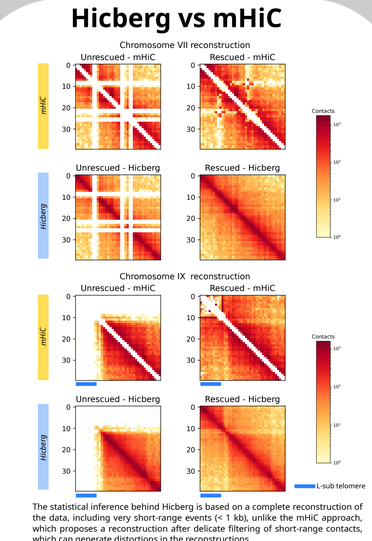
Genomic coverage also offers discriminating power for ambiguous contacts assignation.

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

Hicberg workflow



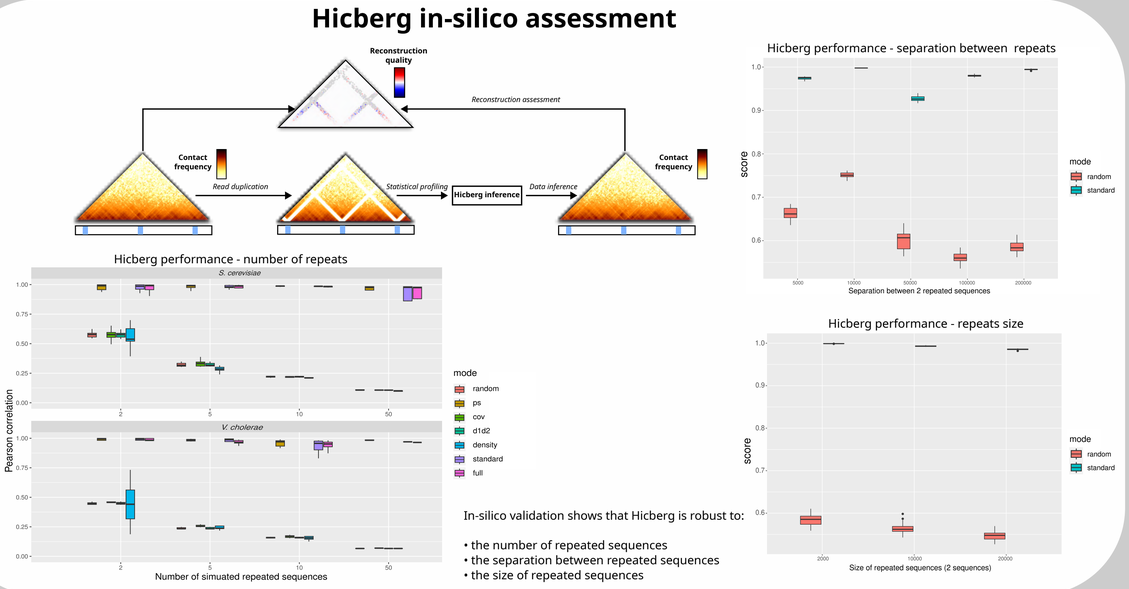
Hicberg vs mHiC



The statistical inference behind Hicberg is based on a complete reconstruction of the data, including very short-range events (< 1 kb), unlike the mHiC approach, which proposes a reconstruction after delicate filtering of short-range contacts, which can generate distortions in the reconstructions.

Zheng and al. Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. 2019

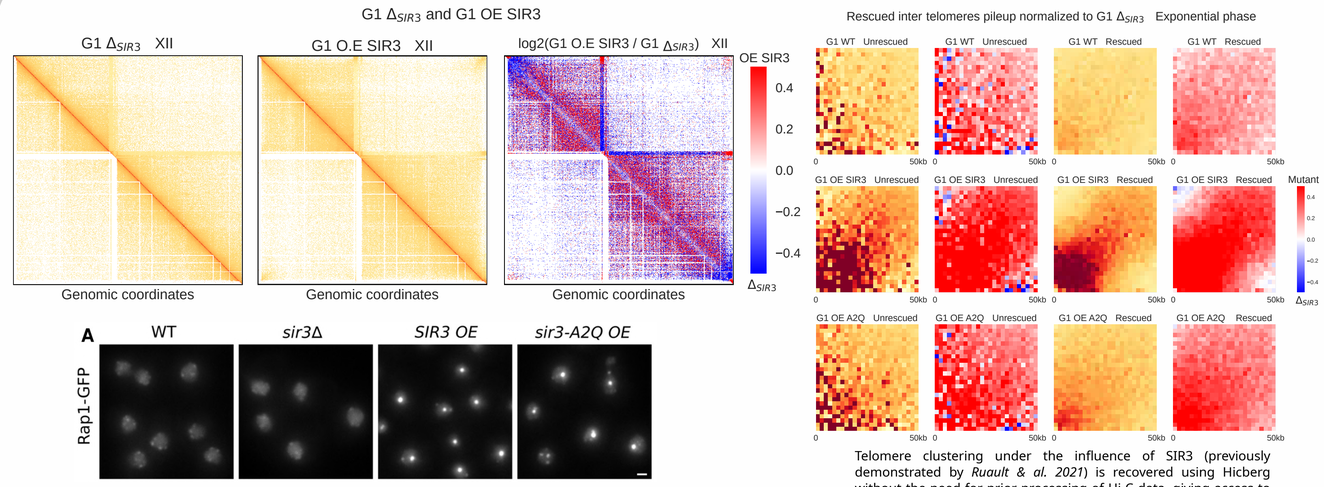
Hicberg in-silico assessment



In-silico validation shows that Hicberg is robust to:

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

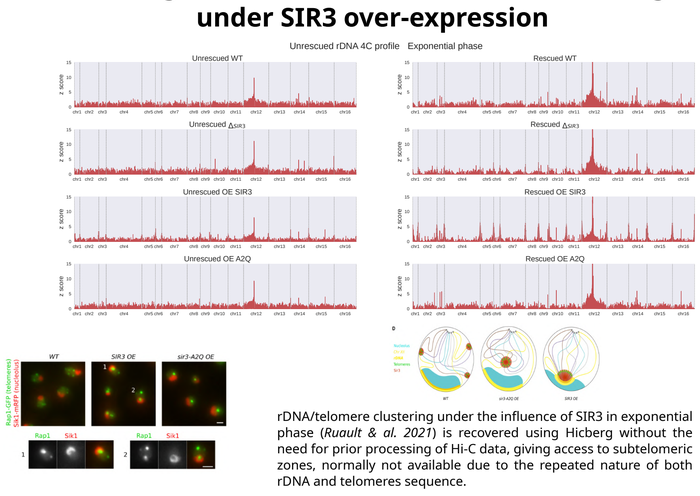
Hicberg retrieves telomere clustering under SIR3 over-expression



Data : Ruault. and 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 areas, normally not determinated due to the repeated nature of the sequences.

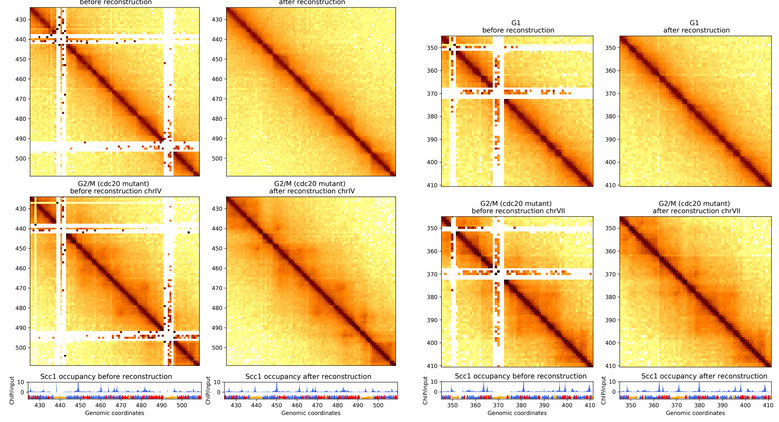
Hicberg retrieves rDNA - telomeres clustering under SIR3 over-expression



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 available due to the repeated nature of both rDNA and telomeres sequence.

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

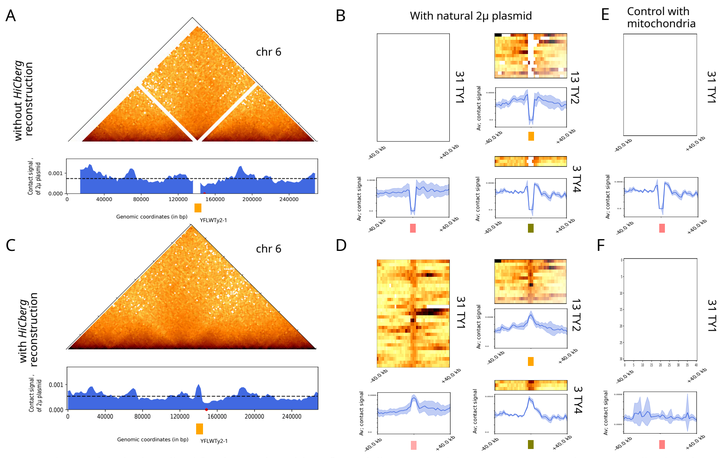
Hicberg retrieves chromatin loops along cell cycle



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 Tys retrotransposon in convergent configuration sites.

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

Girard, F. and al. Anchoring of parasitic plasmids to inactive regions of eukaryotic chromosomes through nucleosome signal. 2023

Conclusion & Perspectives

Hicberg allows 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 or domains (Hi-C).

Hicberg paves the way to the exploration of functional 3D organization of structures involving repeated elements such as telomeres, rDNA, Ty retrotransposons.

New questions can now be addressed through Hicberg :

- the potential interaction of 2 parasitic objects: 2micron plasmid and retrotransposons given the spatial colocation found through Hicberg.

- the involvement of Tys in the positioning and functioning of loop extrusion molecular motors.

- study of the mobility and structural impact of repeated elements during the evolution of *S. cerevisiae* genomes.

References

Dekker J., Rippe K., Dekker M., Kleckner N. Capturing chromosome conformation. 2002.

Lieberman-Aiden E. & al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. 2009.

Cournac A., Koszul R, Mozziconacci J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. 2016.

Forcato M., Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. 2017.

Piazza A. , Bordelet & al. Cohesin regulates homology search during recombinational DNA repair. 2021.

Ruault M. & al. Sir3 mediates long-range chromosome interactions in budding yeast. 2021.

Fulton AM., Adams SE, Mellor J, Kingsman SM, Kingsman AJ. The organization and expression of the yeast retrotransposon, Ty. 1987.

Zheng & al. Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. 2019.

Girard F. & al. Anchoring of parasitic plasmids to inactive regions of eukaryotic chromosomes through nucleosome signal. 2023.