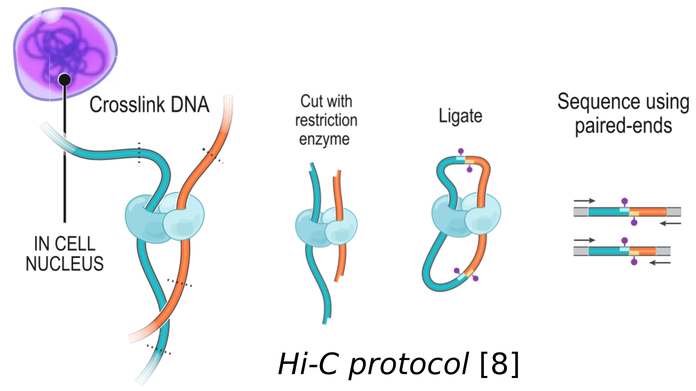
Benchmarking Hi-C scaffolders

[1]Nadège Guiglielmoni, [2]Romain Koszul, [1,3]Jean-François Flot

[1]Service Evolution Biologique et Ecologie, Université libre de Bruxelles, 1050 Brussels, Belgium
[2]Equipe Régulation Spatiale des Génomes, Institut Pasteur, 75015 Paris, France
[3]Interuniversity Institute of Bioinformatics in Brussels - (IB)^2, 1050 Brussels, Belgium

Introduction

Chromosome conformation capture (3C) is a technique to **study the 3D structure of genomes** [1], further developed as **Hi-C**.
The method allows to compute interaction frequencies between every part of a genome, and has revealed that:
▸ intrachromosomal contacts are more frequent than interchromosomal contacts;
▸ interaction frequencies are a function of the distance between two loci in the sequence [2][3].
These principles are used to **scaffold fragmented genomes and obtain chromosome-level assemblies**.
Several tools have been developed for this purpose, such as **3D-DNA** [4], **instaGRAAL** [5][6], **SALSA2** [7].
We propose to evaluate these Hi-C scaffolders on two genomes, *Caenorhabditis elegans* and *Drosophila melanogaster*.

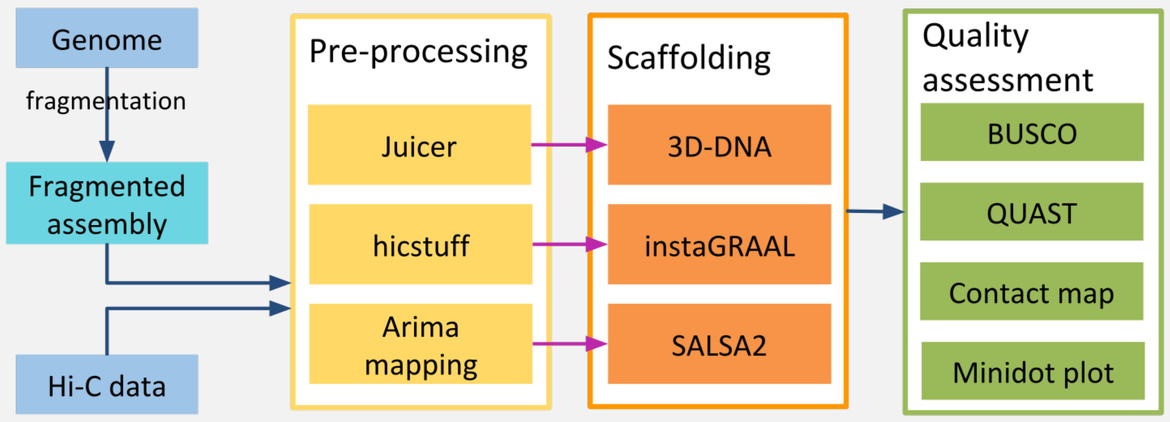


Hi-C protocol [8]

Benchmark method

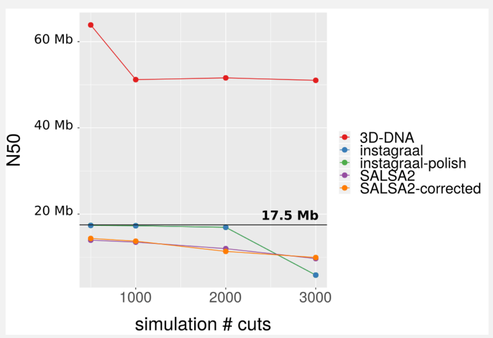
Three published Hi-C scaffolders were included in the benchmark:
3D-DNA: build a megascaffold then split it according to interaction variations
SALSA2: add Hi-C links to iteratively join contigs
instaGRAAL: reassemble to obtain a more likely structure

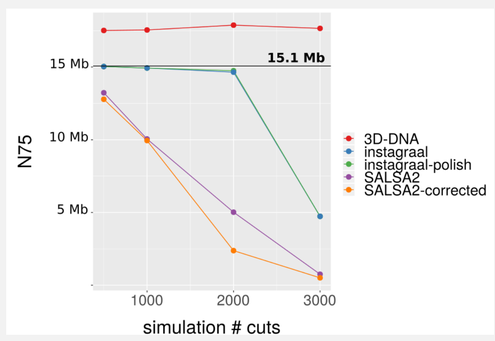
Selected data: *C. elegans* and *D. melanogaster*
+ **good reference genomes**
+ **Hi-C data available**
Reference genomes were cut randomly with increasing numbers of cuts.



Results for *C. elegans*

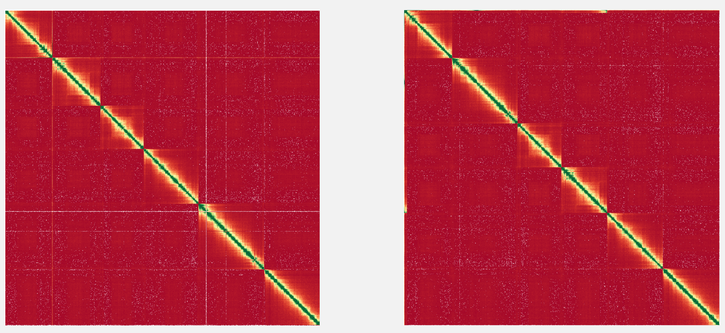
Genome: 100 Mb, 6 chromosomes
Hi-C data: 57 million reads





Assessment of scaffolding over increasing numbers of cuts
5 to 10 replicates were run for each simulation. These graphs display the median N50 and N75.

Reference **3D-DNA**



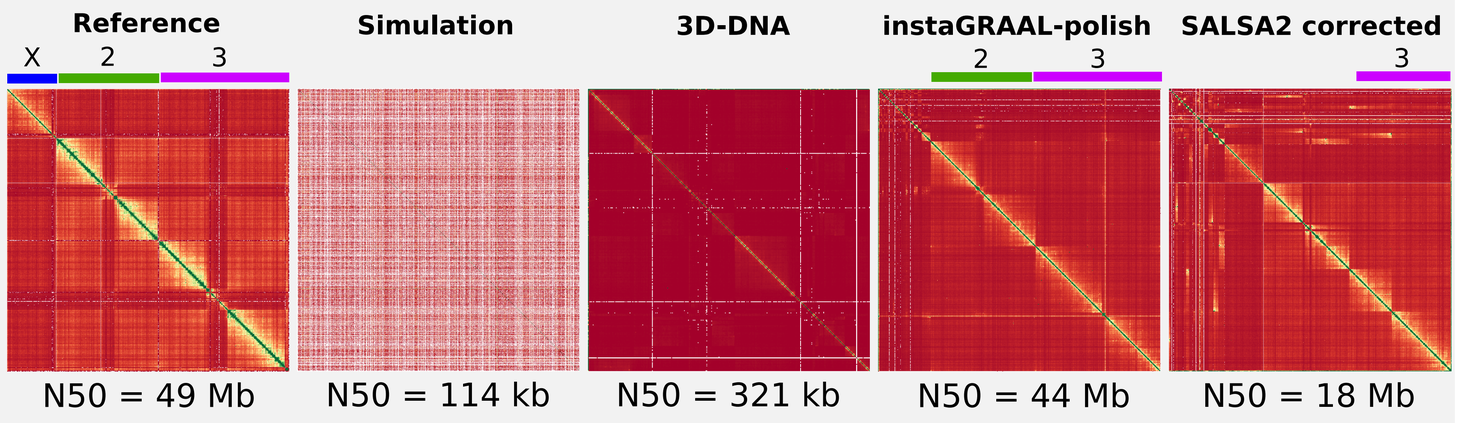
Comparison of the reference and 3D-DNA contact maps
Simulated data: 500 cuts

All tools managed to reconstruct chromosome-level scaffolds :
▸ instaGRAAL shows consistent N50 and N75 up to 2000 cuts
▸ SALSA2 has the lowest N50 and N75
▸ 3D-DNA merged scaffolds, but the quality of the contact map shows that the assembly can be manually corrected

Results for *D. melanogaster*

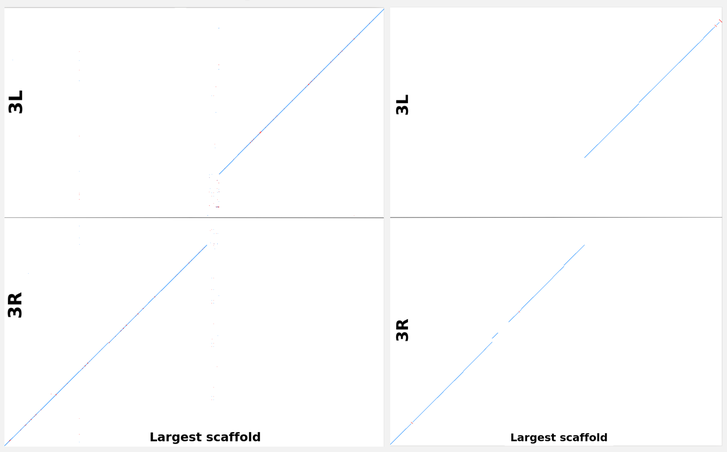
Genome: 144 Mb, 4 chromosomes
Hi-C data: 155 million reads

Simulated data: 2000 cuts

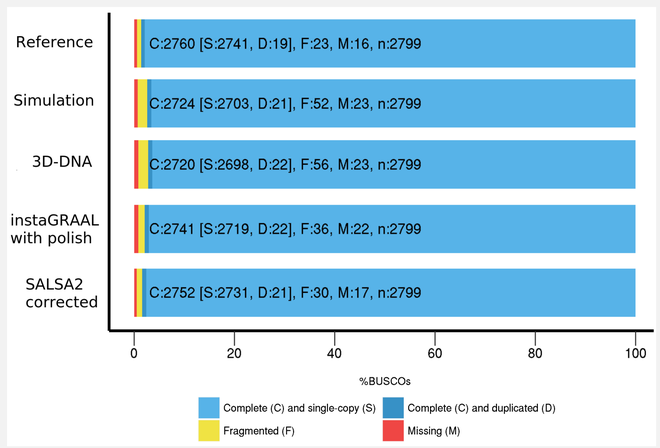


Contact maps to evaluate the quality of scaffoldings

instaGRAAL-polish **SALSA2 corrected**



Mapping of the largest scaffold against chr3



BUSCO assessment

▸ instaGRAAL and SALSA2 achieved chromosome-level scaffolds: chr3 for SALSA2, chr2 and chr3 for instaGRAAL; however, the scaffolded chr3 were shorter in the centromere area
▸ every scaffolders scored a high BUSCO completeness

Conclusion and perspectives

The three Hi-C scaffolders performed well on *C. elegans*. The structure of *D. melanogaster* proved more challenging, and only instaGRAAL and SALSA2 generated better scaffolds. instaGRAAL displayed the most consistent results compared to reference genomes.
Hi-C scaffolders should be further tested for the following cases:
▸ **Gb-sized genomes**
▸ **different Hi-C libraries** (cell phase, protocol, more or less reads)
▸ with **other published scaffolders**: HiCAssembler, ALL-HiC

For more information and command lines, take a look at the github page:
https://github.com/nadegeguiglielmoni/Benchmarking_Hi-C_scaffolders

Bibliography

[1] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. 295(5558):1306–1311, 2002.
[2] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950):289–93, 2009.
[3] Jean-François Flot, Hervé Marie-Nelly, and Romain Koszul. Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. FEBS Letters, 589(20):2966–2974, 2015.
[4] Olga Dudchenko, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, Ido Machol, Eric S. Lander, Aviva Presser Aiden, and Erez Lieberman Aiden. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science, 356(6333):92–95, 2017.
[5] Hervé Marie-Nelly, Martial Marbouty, Axel Cournac, Jean-François Flot, Gianni Liti, Dante Poggi Parodi, Sylvie Syan, Nancy Guillen, Antoine Margeot, Christophe Zimmer, and Romain Koszul. High-quality genome (re)assembly using chromosomal contact data. Nature Communications, 5:1–10, 2014.
[6] github.com/koszullab/instaGRAAL
[7] Jay Ghurye, Mihai Pop, Sergey Koren, Derek Bickhart, and Chen Shan Chin. Scaffolding of long read assemblies using long range contact information. BMC Genomics, 18(1):1–11, 2017.
[8] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer,Eric S. Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7):1665-1680, 2014.

Acknowledgements

We thank the SFBI, the LRZ supercomputing centre for their resources, and Antoine Régnier for his help and his computational resources.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764840