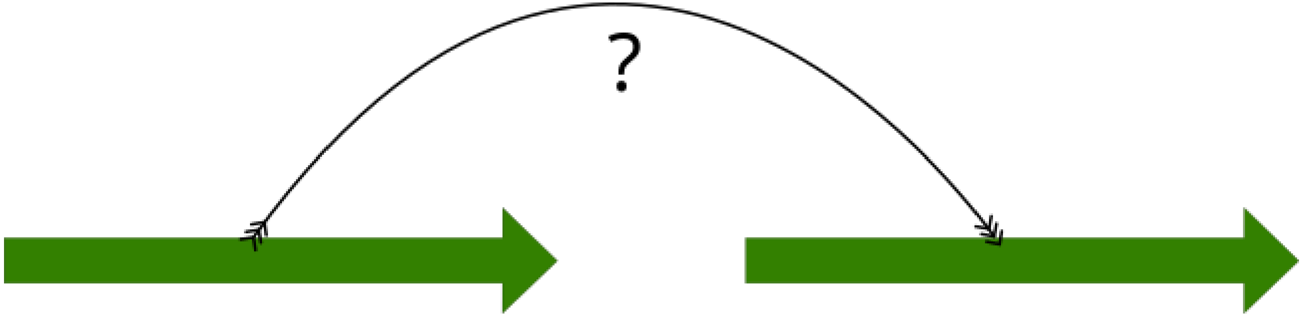
How to build a scaffold

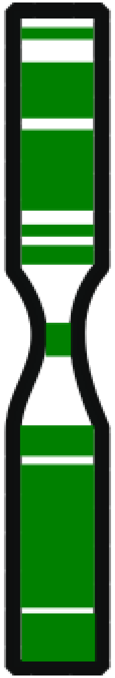


J.Kreplak, A. Cormier, C. Klopp

What are you going to learn ?

- *What a scaffold is*
- *What scaffolding is*
- *How you can scaffold contigs using optical map or Hi-C*
- *How to scaffold with an optical map*
- *How to scaffold with Hi-C*

What would we like our assembly to look like ?

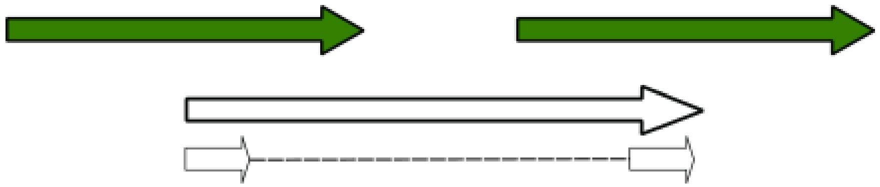


Pseudo-molecule / Chromosome :

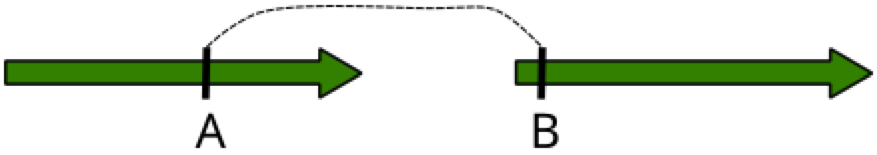
A sequence representing a complete and individual physical segment of DNA

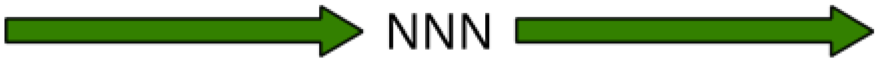
How do we link two contigs and for what ?

With a sequence that span between my contigs :

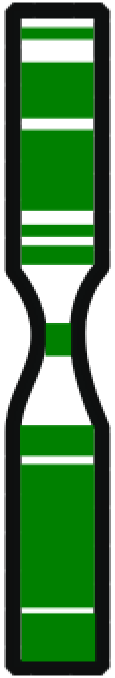


With Specific data linking two markers :





Scaffolds



Pseudo-molecule / Chromosome :
A sequence representing a complete and individual physical segment of DNA

Technic overview

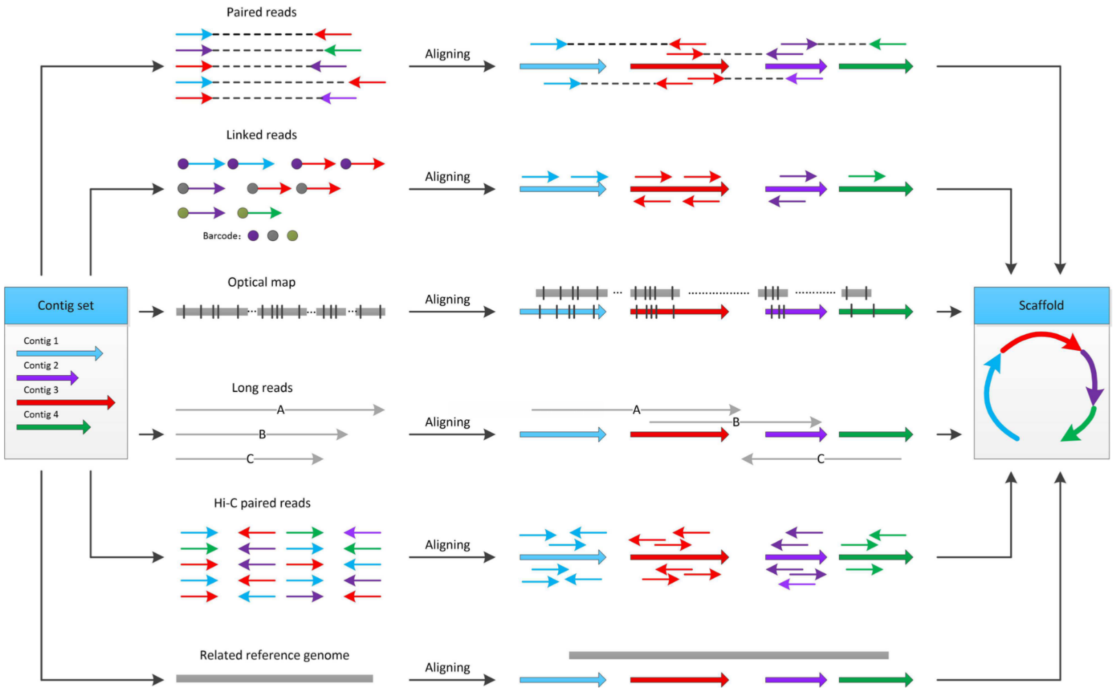


Figure 1. Processes of six scaffolding method types. First, these reads are aligned against the contigs, or these contigs are aligned against optical maps or the reference genomes. Second, based on the alignment information, the order and orientation among contigs are deduced. Finally, scaffolds are output by these methods.

(Luo et al. 2021)

Choice criteria

- Range :
  Can you connect contigs separated by 10k, 100k , 1Mb… ?
- Density/Coverage :
  Do you cover all you contigs ? With an appropriate number of reads/markers ?
- Accuracy:
  Can you estimate properly the distance between two contigs ?
- Initial quality of the assembly :
  Is your contigs size long enough ?Do you have chimeric/low quality contigs ?

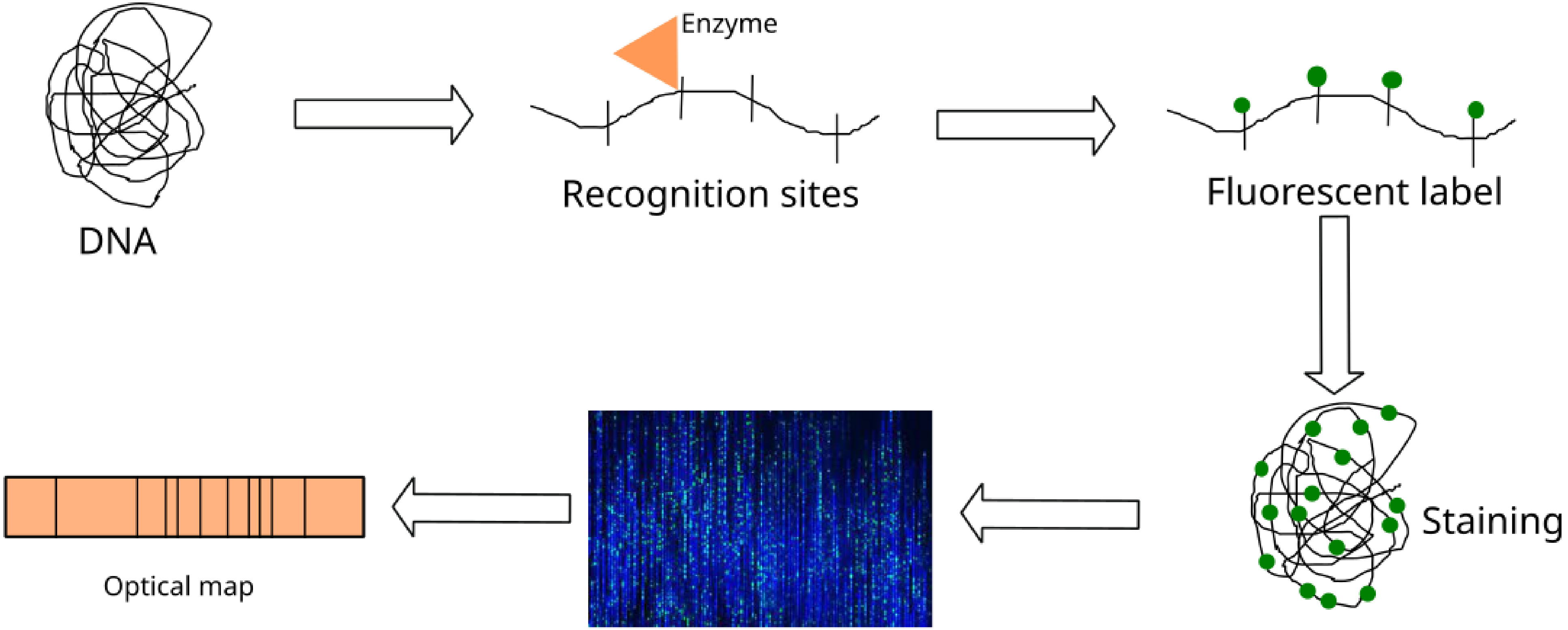
A change of paradigm

- **Long-reads assemblies** have **high-continuity** and **large-sized contigs** which reduce the utility of short range methods like mated-pairs
- Focus is now on **long-range scaffolding** with **accurate estimation** of the distance to go directly from contigs to pseudomolecules:

  - **Optical map** (Bionano Saphyr)

  - **Hi-C**

Optical mapping (1)



DNA

Enzyme

Recognition sites

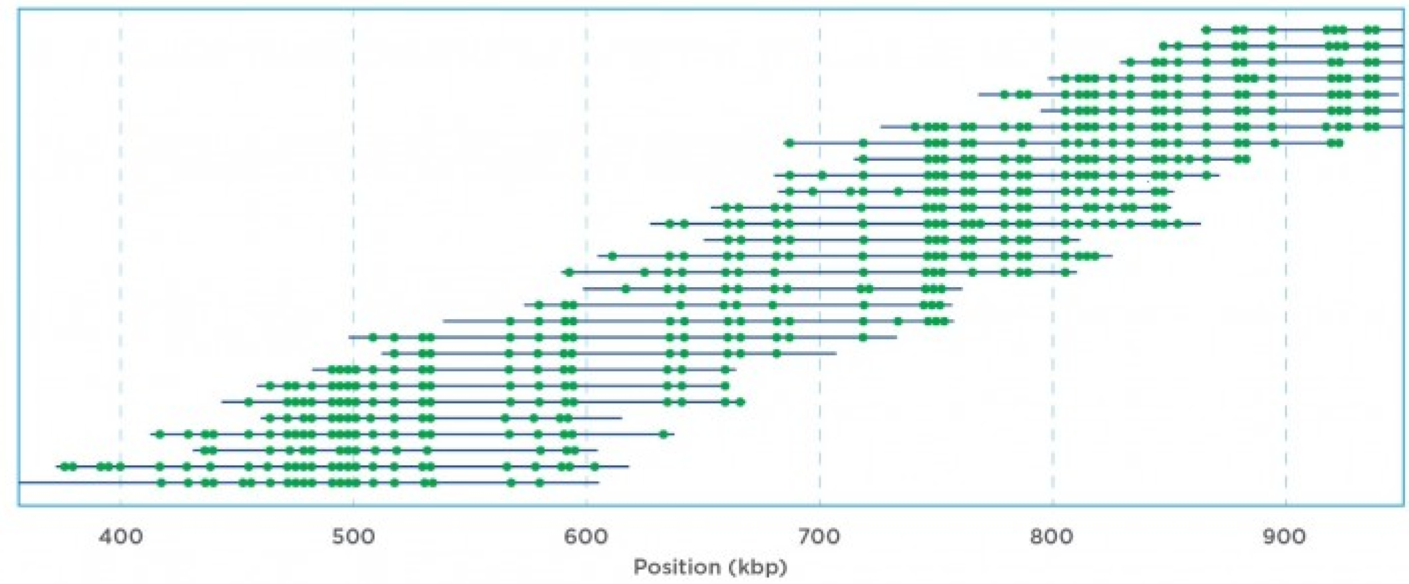
Fluorescent label

Staining

Optical map

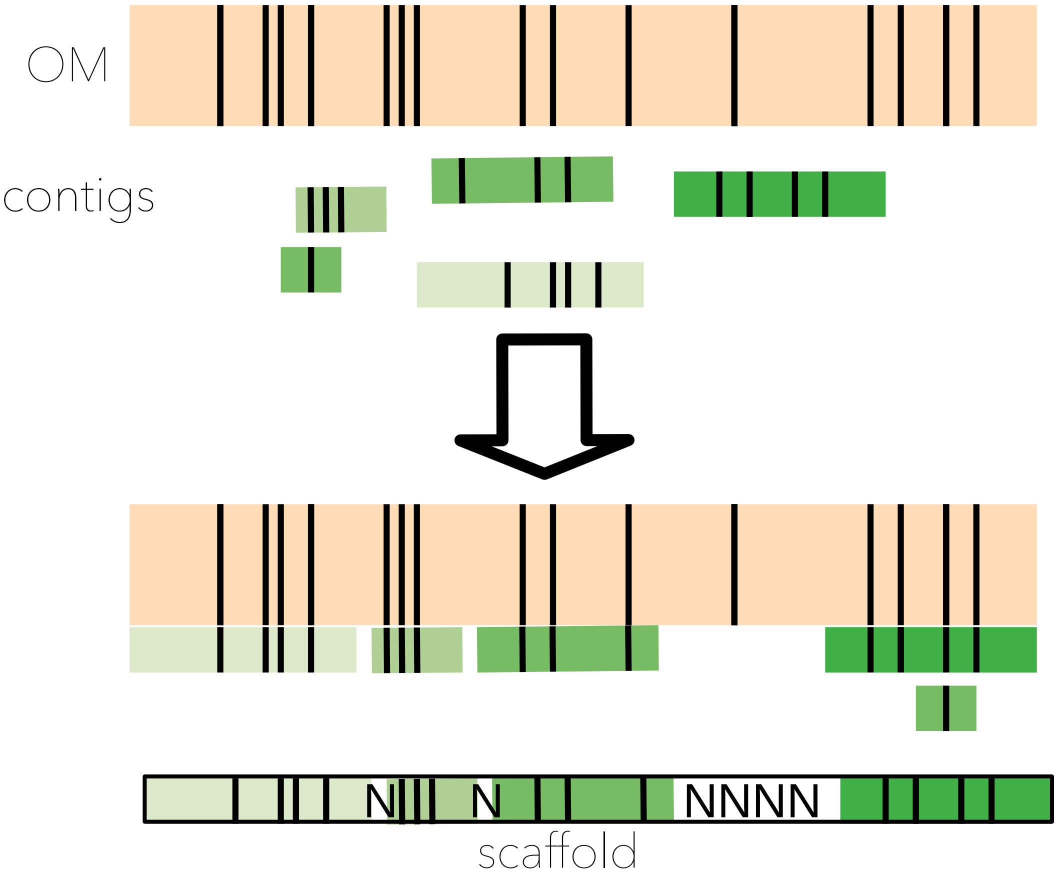
(Bionano Saphyr results)

Optical mapping (2)



https://www.france-genomique.org/technological-expertises/whole-genome/optical-mapping/?lang=en

Optical mapping (3)



- There is no sequencing involved in optical maps
- To process it we will do an in silico digestion of the contigs with the same restriction enzyme
- compare recognition sites (labels) on our contigs to those on our map
- If you have the same frequencies/distance between them it's a match !
- Size of optical maps and density of label are really important

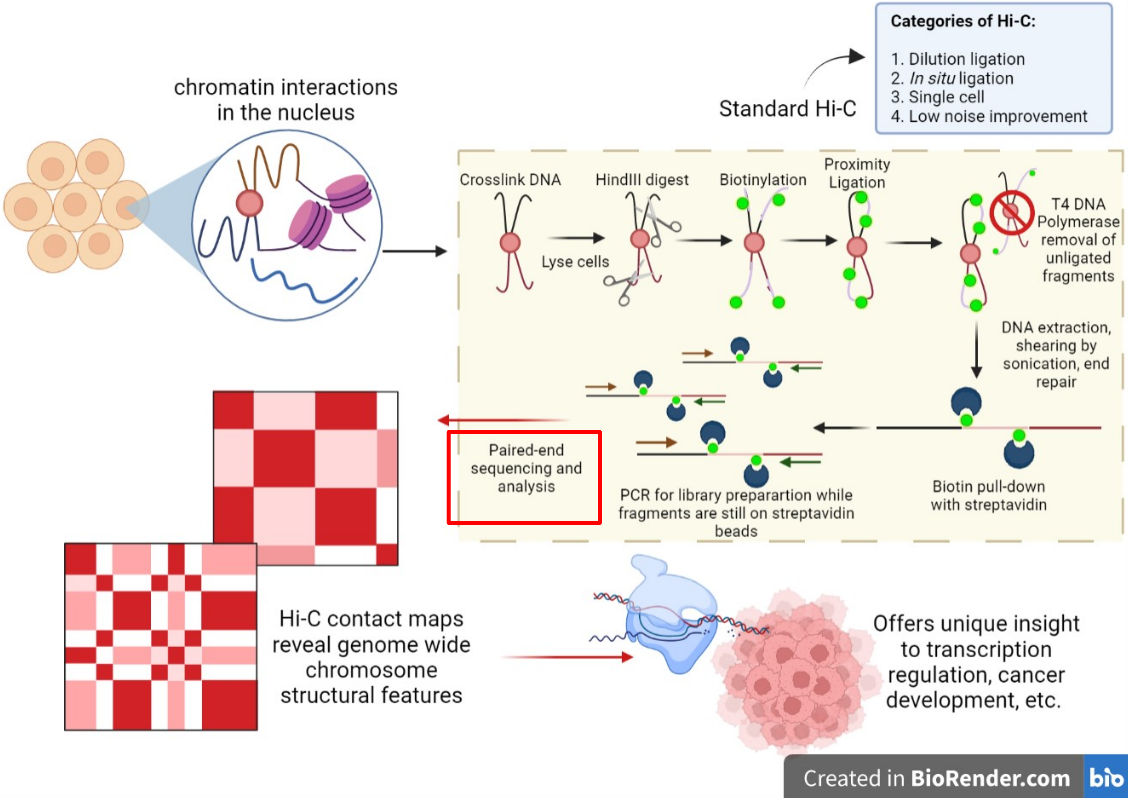
Bionano Saphyr

- Last generation of optical maps
- Bioinformatics steps often done by the provider
- Only a few available tools :

    - Bionano Solve to scaffold
    - BiSCot to improve assembly

- For large genome won't be able to assemble properly centromeres and do pseudo-molecules
- Cheaper than Hi-C
- No sequencing

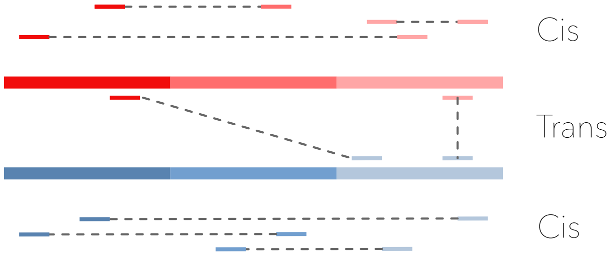
Hi-C


Categories of Hi-C:
1. Dilution ligation
2. *In situ* ligation
3. Single cell
4. Low noise improvement
Standard Hi-C
chromatin interactions in the nucleus
Crosslink DNA
Lyse cells
HindIII digest
Biotinylation
Proximity Ligation
T4 DNA Polymerase removal of unligated fragments
DNA extraction, shearing by sonication, end repair
Biotin pull-down with streptavidin
PCR for library preparartion while fragments are still on streptavidin beads
Paired-end sequencing and analysis
Hi-C contact maps reveal genome wide chromosome structural features
Offers unique insight to transcription regulation, cancer development, etc.
Created in BioRender.com bio

https://en.wikipedia.org/wiki/Hi-C_(genomic_analysis_technique)

- A method to capture chromatin conformation by sequencing
- First used for long-range interactions (Lieberman-Aiden et al. 2009)
- Hypothesis that those interactions could be use to scaffolds sequences into pseudo-molecules (Burton et al. 2013)
- Protocols with different or multiple enzymes were developed to boost the quality of scaffolding

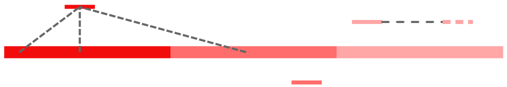
How to obtain a contact map?



Mapping

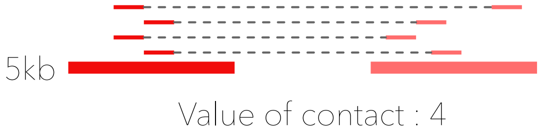
Mapper need to be tuned for variable insert size.

- bwa mem -5 -SP
- minimap2



Filtering

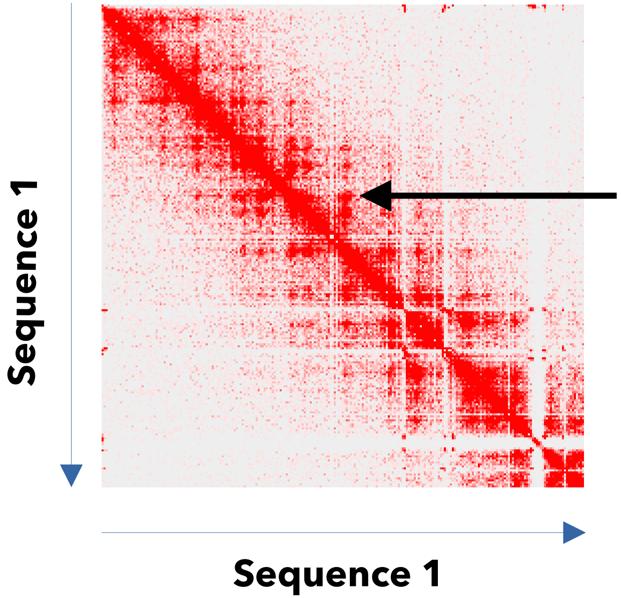
Remove multi-mapping,
low-quality mapping,
Invalids pair,
Singletons,
Invalid ligation products



Binning

Choose windows of different size (5kb – 25 kb – 100 kb - 500 kb – 2,5 Mb) to regroup information

How to read a contact map for scaffolding ?



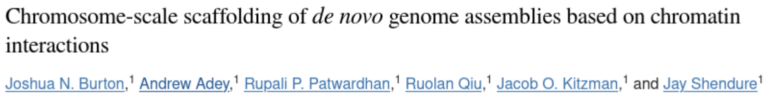
Each point is a bin between two windows

Think really big heatmap /dotplot

- For scaffolding, you need to use the contact map to find spatially near contigs
- You need to use the signal to order contigs correctly
- Can that be done by a computer ?
- Maximazing diagonal signal

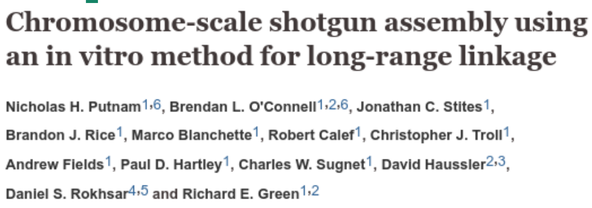
Bioinformatics will help you !

2013



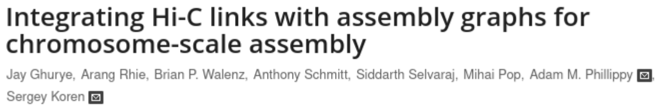
Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions

Joshua N. Burton,[1] Andrew Adey,[1] Rupali P. Patwardhan,[1] Ruolan Qiu,[1] Jacob O. Kitzman,[1] and Jay Shendure[1]



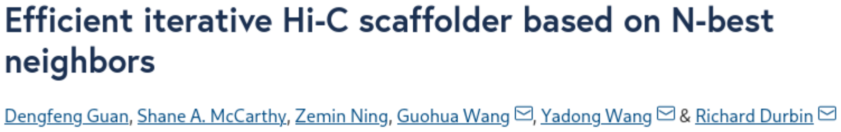
Chromosome-scale shotgun assembly using an in vitro method for long-range linkage

Nicholas H. Putnam[1,6], Brendan L. O'Connell[1,2,6], Jonathan C. Stites[1], Brandon J. Rice[1], Marco Blanchette[1], Robert Calef[1], Christopher J. Troll[1], Andrew Fields[1], Paul D. Hartley[1], Charles W. Sugnet[1], David Haussler[2,3], Daniel S. Rokhsar[4,5] and Richard E. Green[1,2]



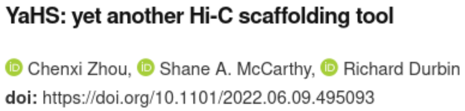
Integrating Hi-C links with assembly graphs for chromosome-scale assembly

Jay Ghurye, Arang Rhie, Brian P. Walenz, Anthony Schmitt, Siddarth Selvaraj, Mihai Pop, Adam M. Phillippy ✉, Sergey Koren ✉



Efficient iterative Hi-C scaffolder based on N-best neighbors

Dengfeng Guan, Shane A. McCarthy, Zemin Ning, Guohua Wang ✉, Yadong Wang ✉ & Richard Durbin ✉



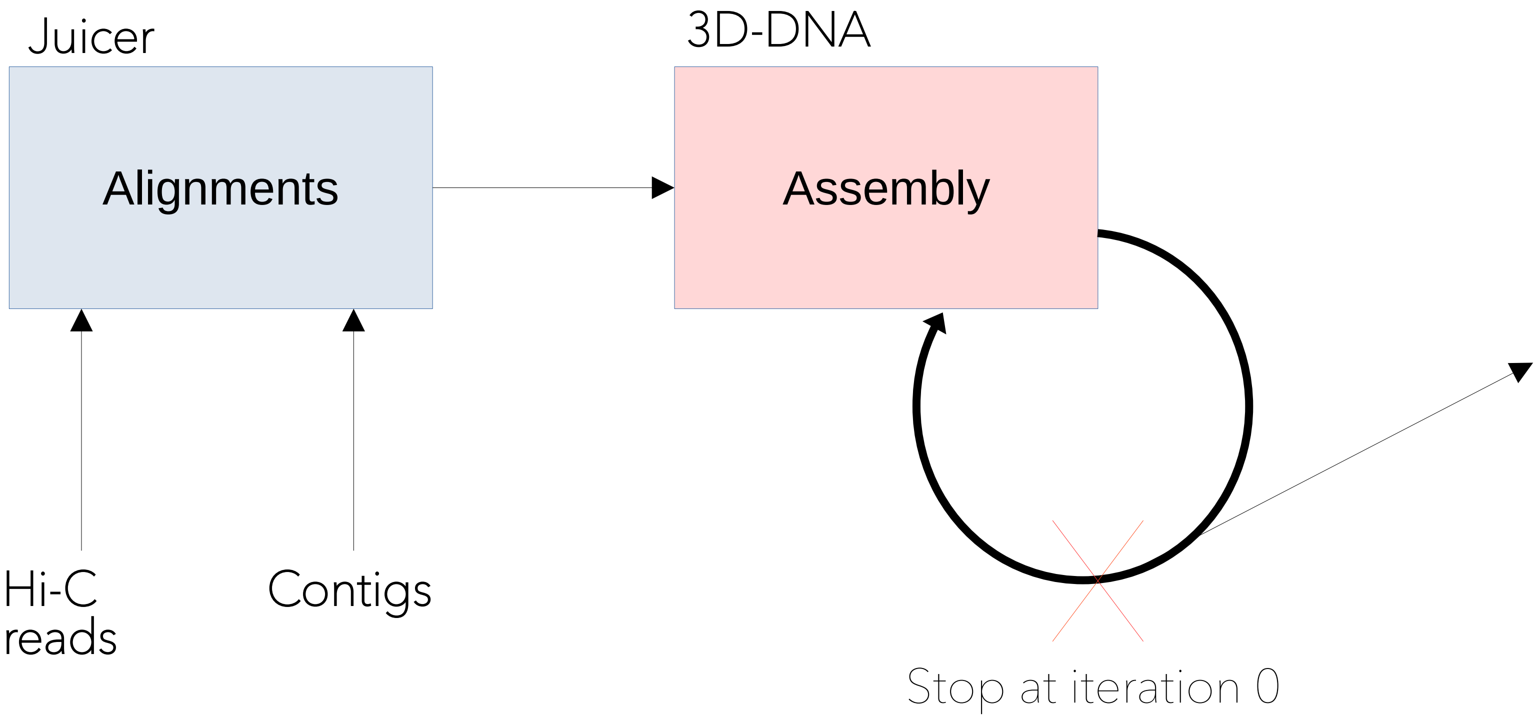
YaHS: yet another Hi-C scaffolding tool

Chenxi Zhou, Shane A. McCarthy, Richard Durbin

doi: https://doi.org/10.1101/2022.06.09.495093

- Hi-C methods are still developped by the community
- The graal is to be able to cluster properly a contact map to obtain correct pseudomolecules for every genomes
- Some providers (Dovetail, Phase…) have also their own pipelines …

A pipeline for Hi-C analysis



Juicer

Alignments

3D-DNA

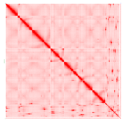
Assembly

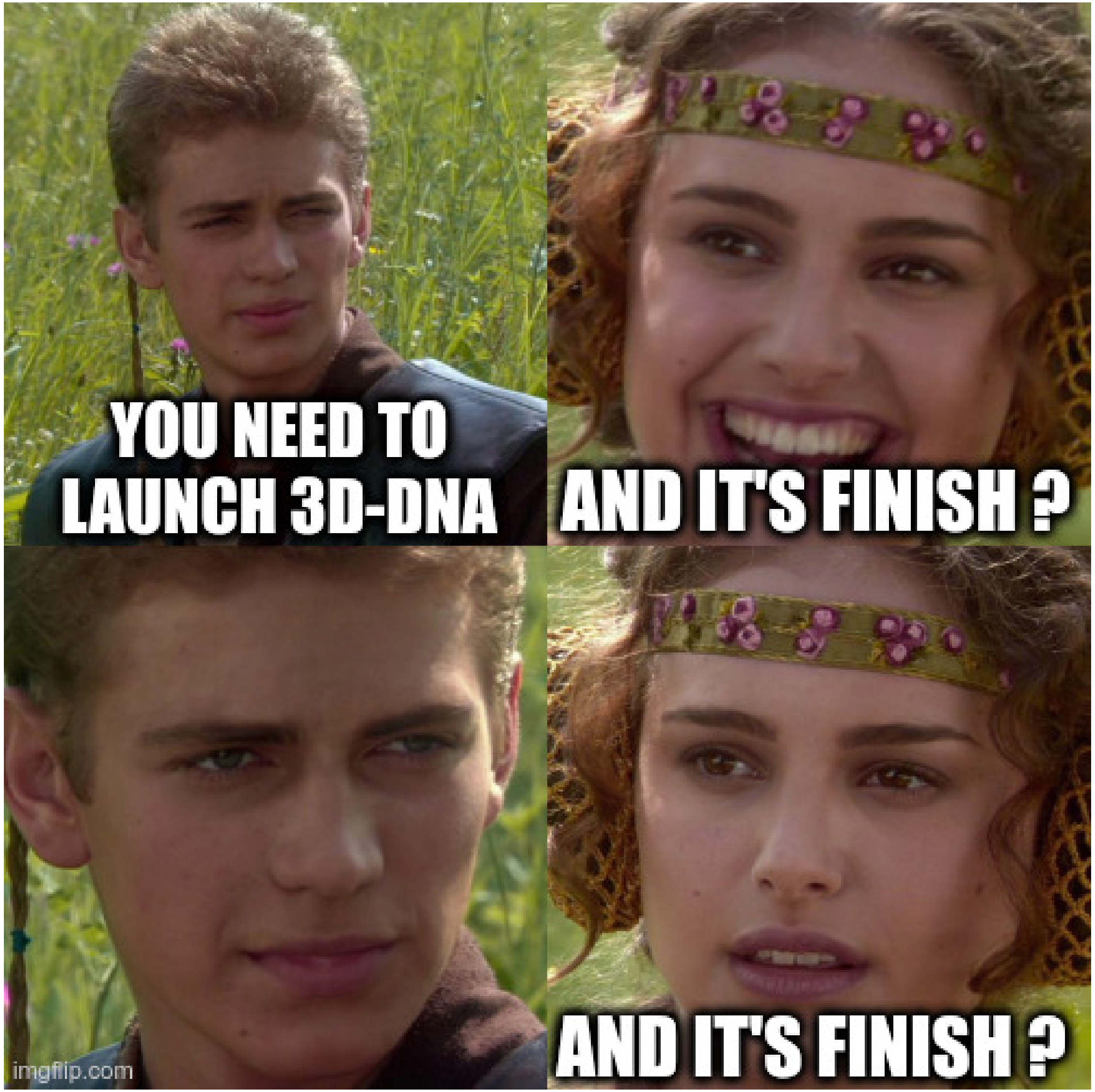
Hi-C reads

Contigs

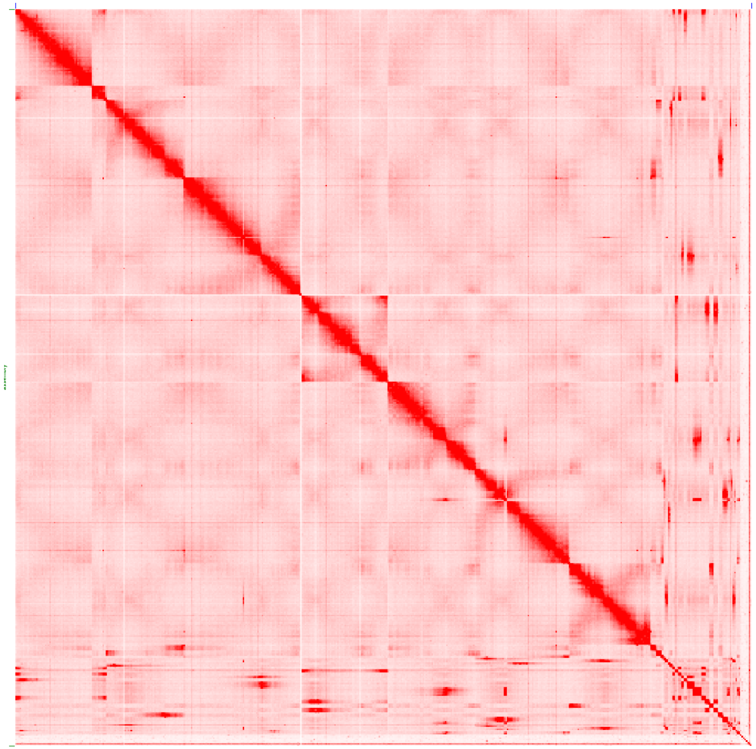
Stop at iteration 0

- A clustered contact map (.hic)
- Contigs position on contact map (.assembly)
- View the assembly using Juicebox

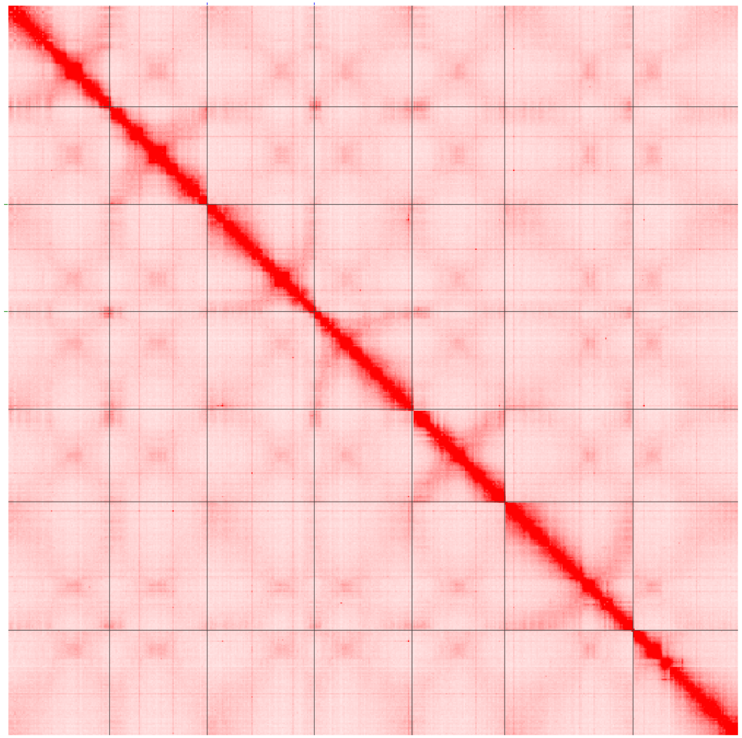



YOU NEED TO LAUNCH 3D-DNA
AND IT'S FINISH ?
AND IT'S FINISH ?
imgflip.com

Last step !



3D-DNA results



After manual curation

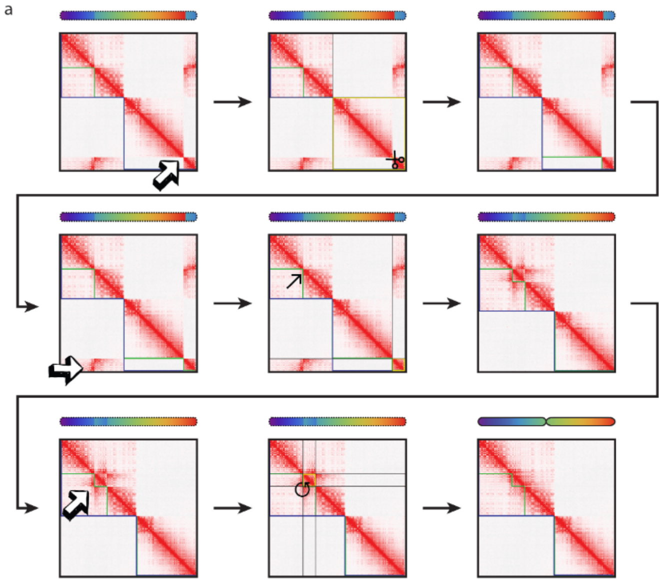
Visualisation using JuiceBox
(mystery plant with 7 chromosomes)

A method to rule them all ?

- Hi-C is the only technics able to scaffold directly a long-reads contigs assembly into linear chromosome without any others data.
- For now, you can't stop after the automated assembly
- You'll need to check the contact map and correct it
- Expensive, you need to sequence with a coverage of

  30-40x
- Other datas like genetic map can help to detect or correct difficult zone to assemble
- Hi-C could be also used to distinct between different organisms in a metagenomic experiment

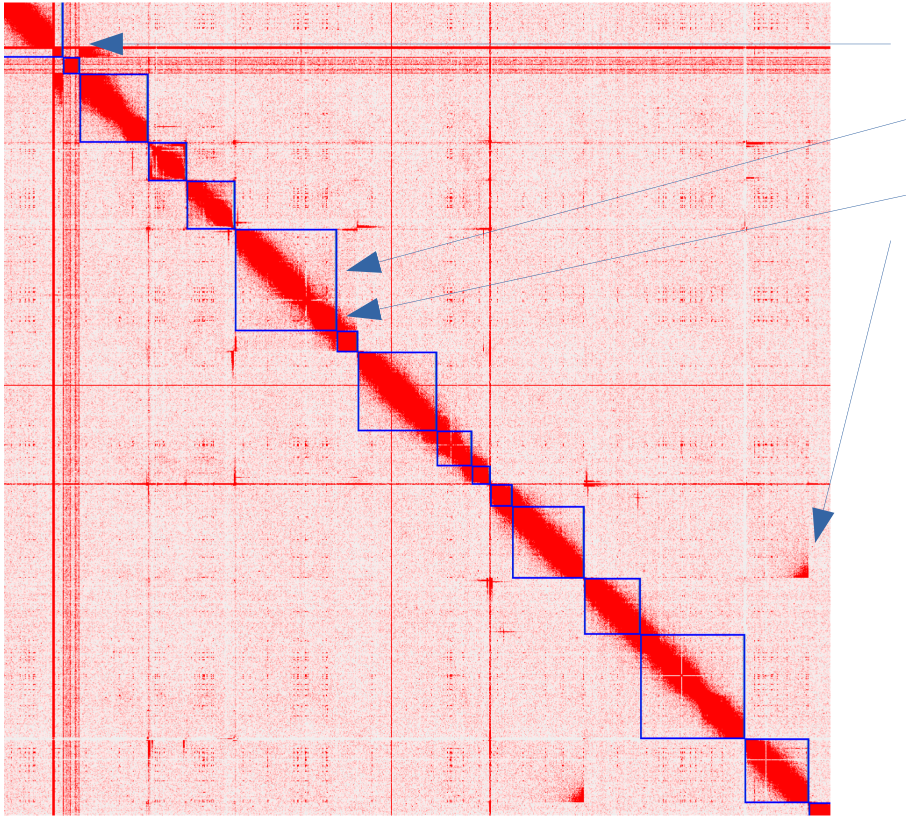
JuiceBox TP



Shift + left click = selection of the element
Shift + left click + move = selection of several elements
Pointing arrow + left click = move element at this position
Cisors + left click = split contig at this position
Angle bloc + left click = add or remove chromosome boundaries

https://www.biorxiv.org/content/10.1101/254797v1.full.pdf

To be done



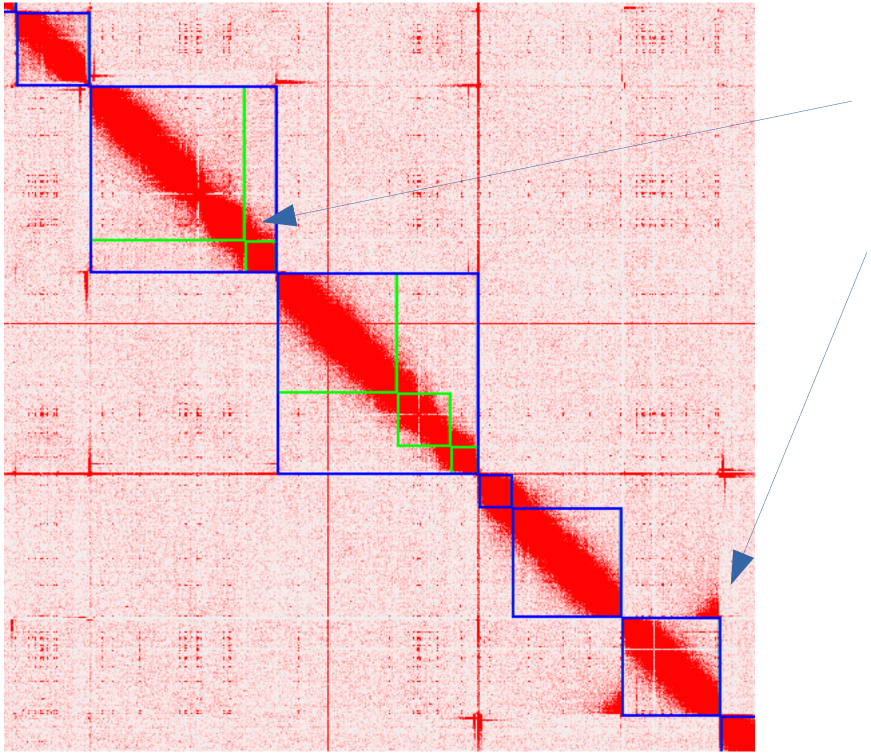
Split contigs

Add chromosomes boundaries

Group contigs in chromosomes

Move contig to build a chromosome

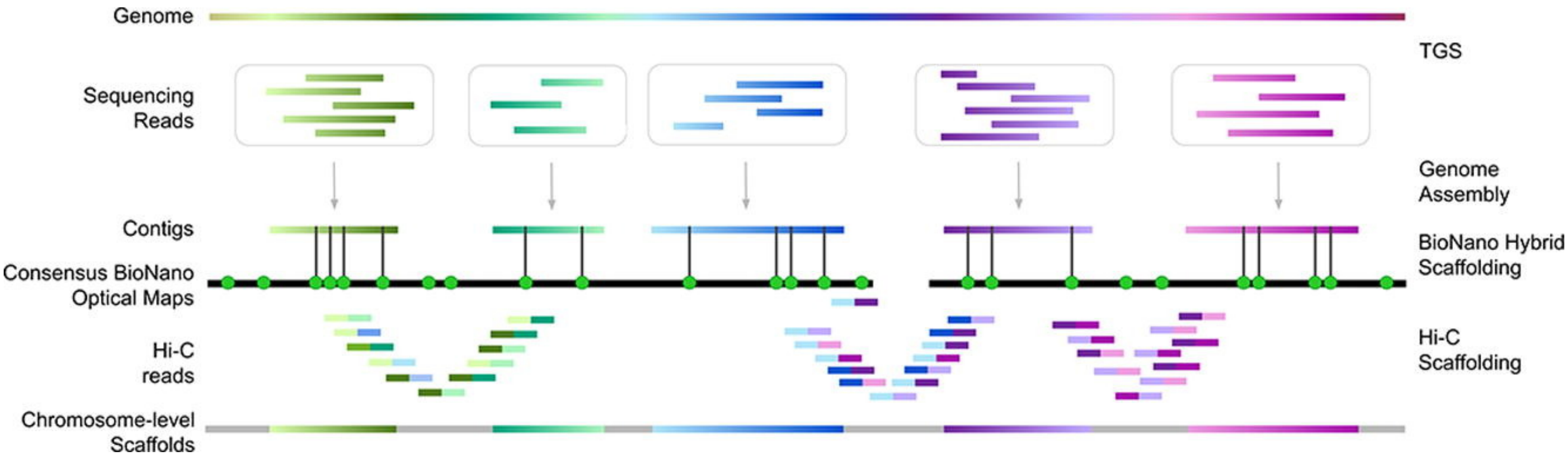
To be done (2)



Add chromosomes boundaries

Rotate contig

Ideal world of scaffolding ?


Genome
TGS
Sequencing Reads
Genome Assembly
Contigs
BioNano Hybrid Scaffolding
Consensus BioNano Optical Maps
Hi-C reads
Hi-C Scaffolding
Chromosome-level Scaffolds