SeqOccIn

INRAE
la science pour la vie, l'humain, la terre

SeqOccIn 2021

Axis 1 – Diploid Assembly

Arnaud Di Franco
Clément Birbes

GeT

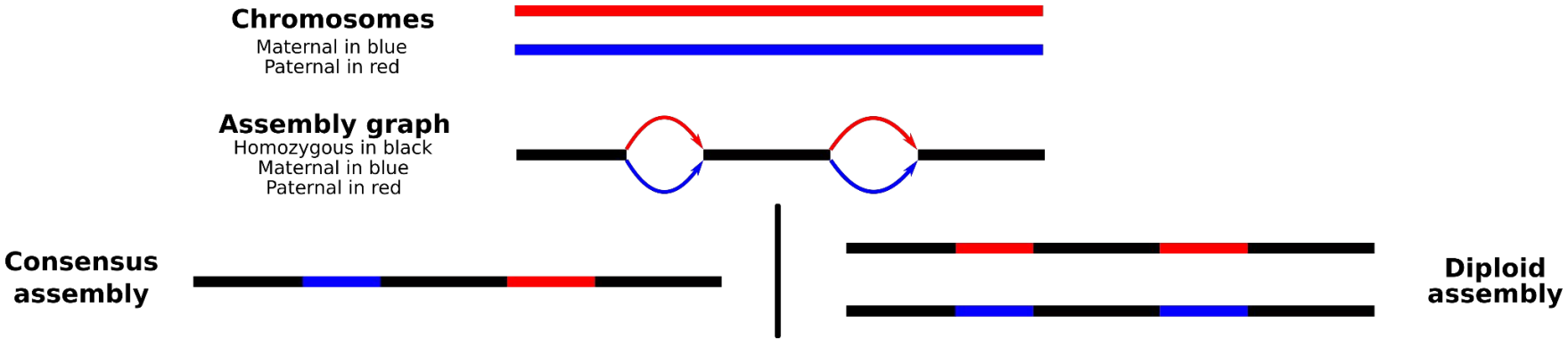
http://get.genotoul.fr
@get_genotoul

http://bioinfo.genotoul.fr

Bioinfo

Diploid Assembly

A diploid assembly is an assembly in which the maternal and paternal haplotypes are separated to create 2 sets of chromosomes.

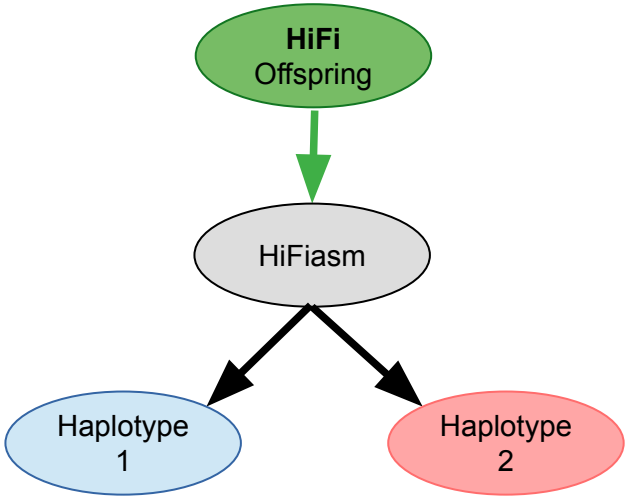


Different methods exists to create a diploid assembly.

- **1 Method using distant connection to separate haplotypes**
- **1 Method resolving haplotype based on parental K-mer**
- **1 Method using long reads only**

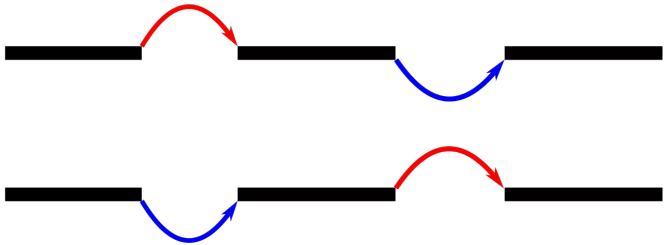
Hifiasm Diploid Assembly

Produce diploid assembly using HiFi reads (long reads with low error rate) and separate haplotype with heterozygous regions informations.



Assembly graph
Homozygous in black
Maternal in blue
Paternal in red

Pseudo-haploid assembly



This method produces chimeric assemblies due to the lack of information in some intra-chromosome areas but also between chromosomes

Long Reads phasing

A third way to separate haplotypes is through phasing, using SNPs

+

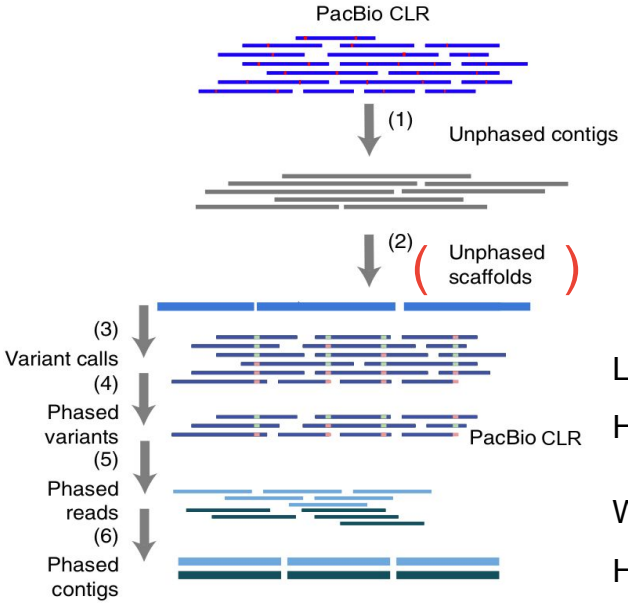
- No parental information needed
- No supplementary sequencing

-

- Affected by numerous factors
  - SNP density
  - Reads length distribution
  - Reads quality
  - ...

Long reads are particularly useful as their size allow for large phased blocks

Protocol based on DipAsm proof of concept *(Garg et al. 2020)* and adapted for long reads only



LongShot (HapCut2 ML)

HapCut2 (&& WhatsHap phase)

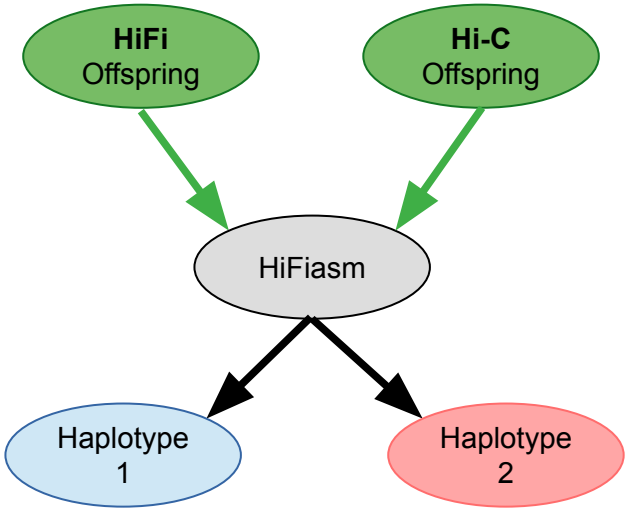
WhatsHap haplotag

Hifiasm

This method produces chimeric assemblies due to the lack of information in some intra-chromosome areas but also between chromosomes

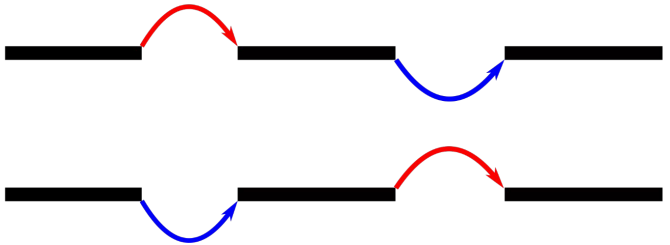
Hifiasm-HiC Diploid Assembly

Produce diploid assembly using HiFi reads (long reads with low error rate) and Hi-C connection information to separate haplotypes



Assembly graph
Homozygous in black
Maternal in blue
Paternal in red

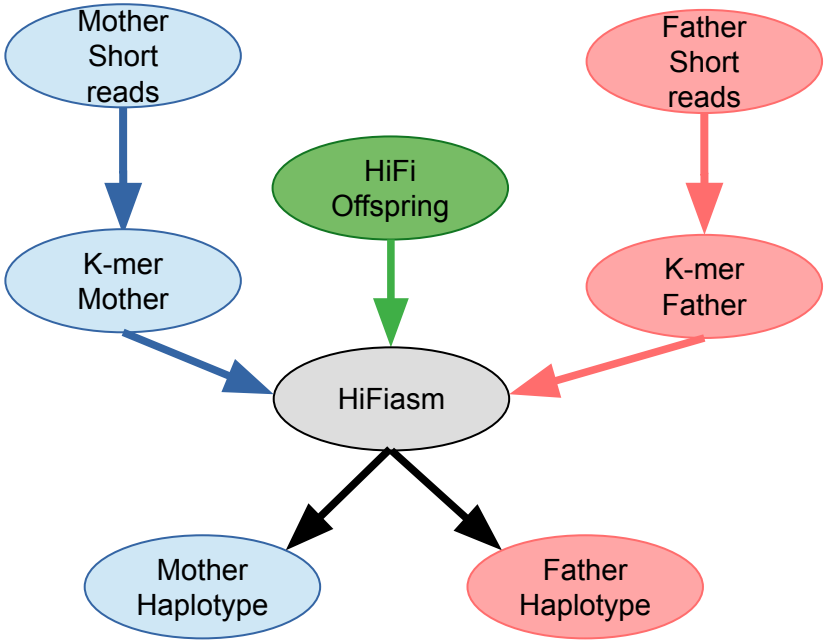
Pseudo-haploid assembly



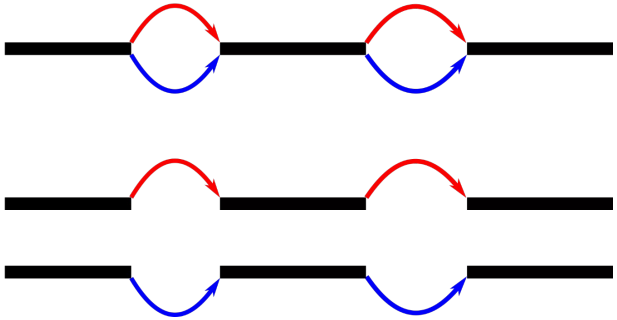
This method produces chimeric assemblies due to the lack of Hi-C contact in some intra-chromosome areas but also between chromosomes

Hifiasm-Kmer Diploid Assembly

Produce diploid assembly using HiFi reads (long reads with low error rate) and parental short reads to separate haplotypes



Assembly graph
Homozygous in black
Maternal in blue
Paternal in red



Haploid assembly

This method produces better diploid assemblies with fewer haplotyping errors through the use of parental kmers

Diploid Assemblies Statistics

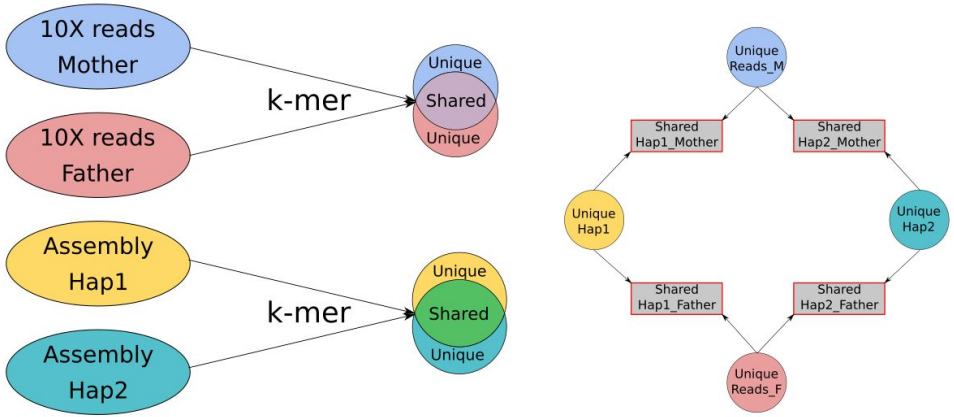
Données	CCS			CCS_Haplotag		CCS_Haplotag_Split		CCS/HiC		CCS_Yak	
Quantité	31X			31X		31X		31X+28X		31X+2x80X	
Assembleur	Hifiasm			Hifiasm		Hifiasm		Hifiasm_HiC		Hifiasm_Yak	
	Consensus	Hap1	Hap2	Hap1	Hap2	Hap1	Hap2	Hap1	Hap2	Hap1	Hap2
Number of contigs	1 444	2 173	2 241	2 582	2 701	3 773	3 009	2 658	2 136	2 871	2 300
Total size (Gb)	3.244	3.175	3.087	3.214	3.208	3.455	3 .61	3.078	3.184	3.2	3.1
Longest contigs (Mb)	158.4	158.4	159.4	132.5	159.1	159.6	158.3	159.8	159.2	158.6	159.2
N50 contigs length (Mb)	84.1	73.6	80.6	43.1	46.2	63.2	68.5	80.1	71.6	71.6	69.1
BUSCO	C:95.9%	C:95.7%	C:95.9%	C:95.4%	C:95.5%	C:95.8%	C:95.7%	C:95.7%	C:95.8%	C:95.8%	C:95.3%

Diploid Assemblies

Are produced haplotypes well-separated ?

Protocol based on k-mer to assess the haplotyping quality:

- Extract a list of all k-mers from parental reads (Reads_M and Reads_F) and assembled haplotypes (Hap1 and Hap2)
- Extract unique k-mers from Reads_M, Reads_F, Hap1 and Hap2
- Compare Shared Unique k-mers between each pair Reads-Hap



	Simple	DipAsm	Trio_Yak	Trio_HiC
Uniq 37161 (Father)	211 306 159			
Uniq 37162(Mother)	190 003 623			
Uniq Hap1	68 127 607	65 177 035	75 122 750	67 672 898
Uniq Hap2	66 389 188	63 610 620	68 786 188	68 174 024
Uniq Shared 37161 Hap1	15 644 523	16 224 181	159 521	12 513 350
Uniq Shared 37161 Hap2	17 557 070	16 076 371	33 337 798	21 004 304
Uniq Shared 37162 Hap1	18 411 760	16 179 561	34 054 852	21 078 367
Uniq Shared 37162 Hap2	15 409 073	15 693 869	116 586	13 021 761
Hamming error rate Hap1	0.376557	0.026206	0.027533	0.035184
Hamming error rate Hap2	0.383939	0.349655	0.028670	0.023762