Ifremer

GigaStore2

-

Assemblage *de novo* résolu au niveau des haplotype du génome de *Crassostrea gigas* (huitre creuse)

Alexandre CORMIER
IRSI-SeBiMER
Jérémie VIDAL-DUPIOL
RBE-IHPE

PEPI IBIS 2021 – Toulouse

16 novembre 2021

SeBiMER at IFREMER

Centre Ifremer Bretagne at Plouzané

SeBiMER history

2010-2018 : Sub unit of RIC team

- 2 permanent contracts
- Technical support

February 2019 : creation of “Service de Bioinformatique de l’Ifremer”

- Mutualisation of bioinformatics activities at Ifremer and its 9 UMRs
- Transversal bioinformatics activities
- For teams and research projects
- Team expansion

SeBiMER team

Patrick Durand, IR
(CDI since 09/2017)

Laura LEROI, IE
(CDI since 04/2019)

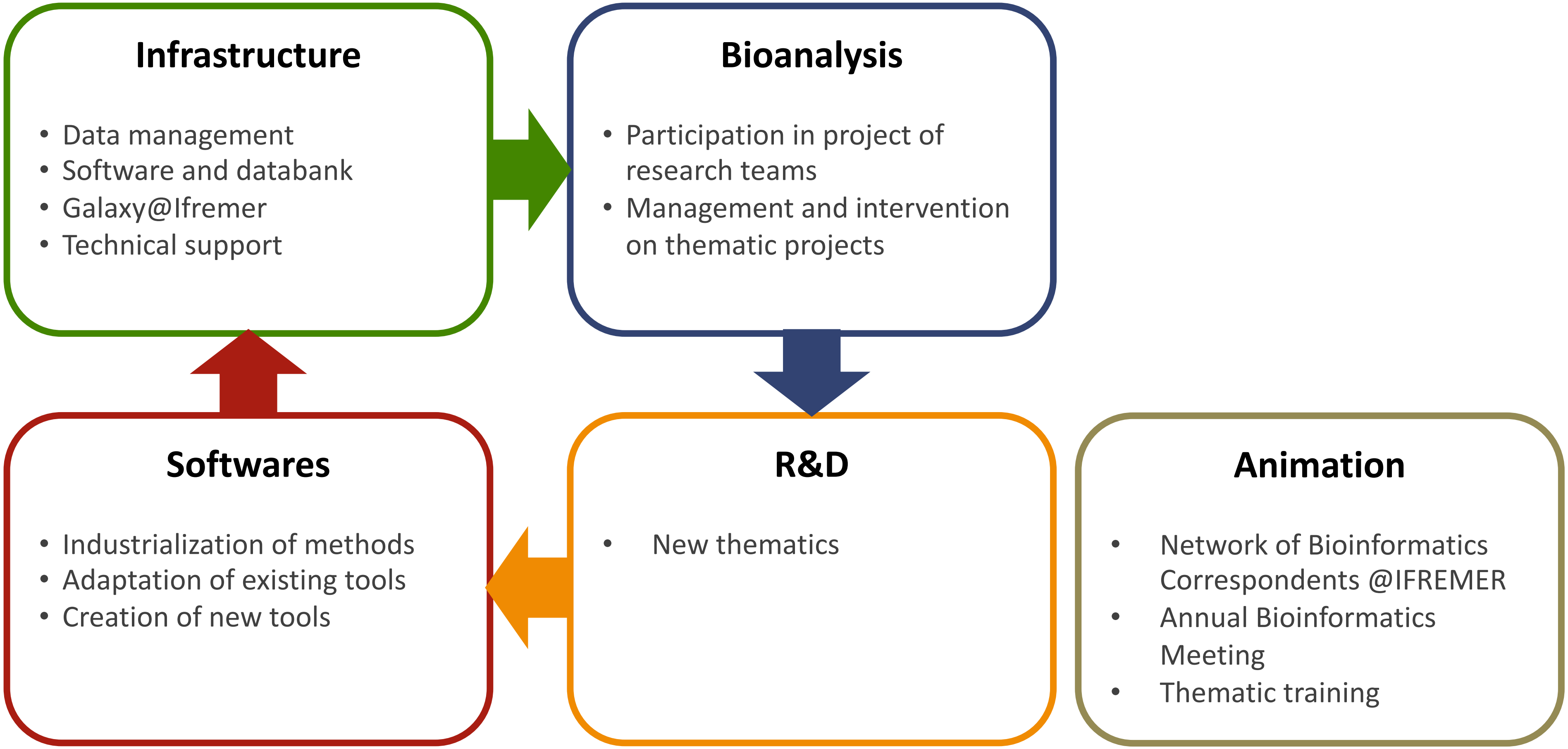
IE
(Start in January 2022)

Cyril NOEL, IR
(CDI since 07/2019)

Alexandre CORMIER, IR
(CDI since 01/2020)

Alizée BARDON
(Apprenticeship 2021-23)

SeBiMER's missions



Infrastructure

- Data management
- Software and databank
- Galaxy@Ifremer
- Technical support

Bioanalysis

- Participation in project of research teams
- Management and intervention on thematic projects

Softwares

- Industrialization of methods
- Adaptation of existing tools
- Creation of new tools

R&D

- New thematics

Animation

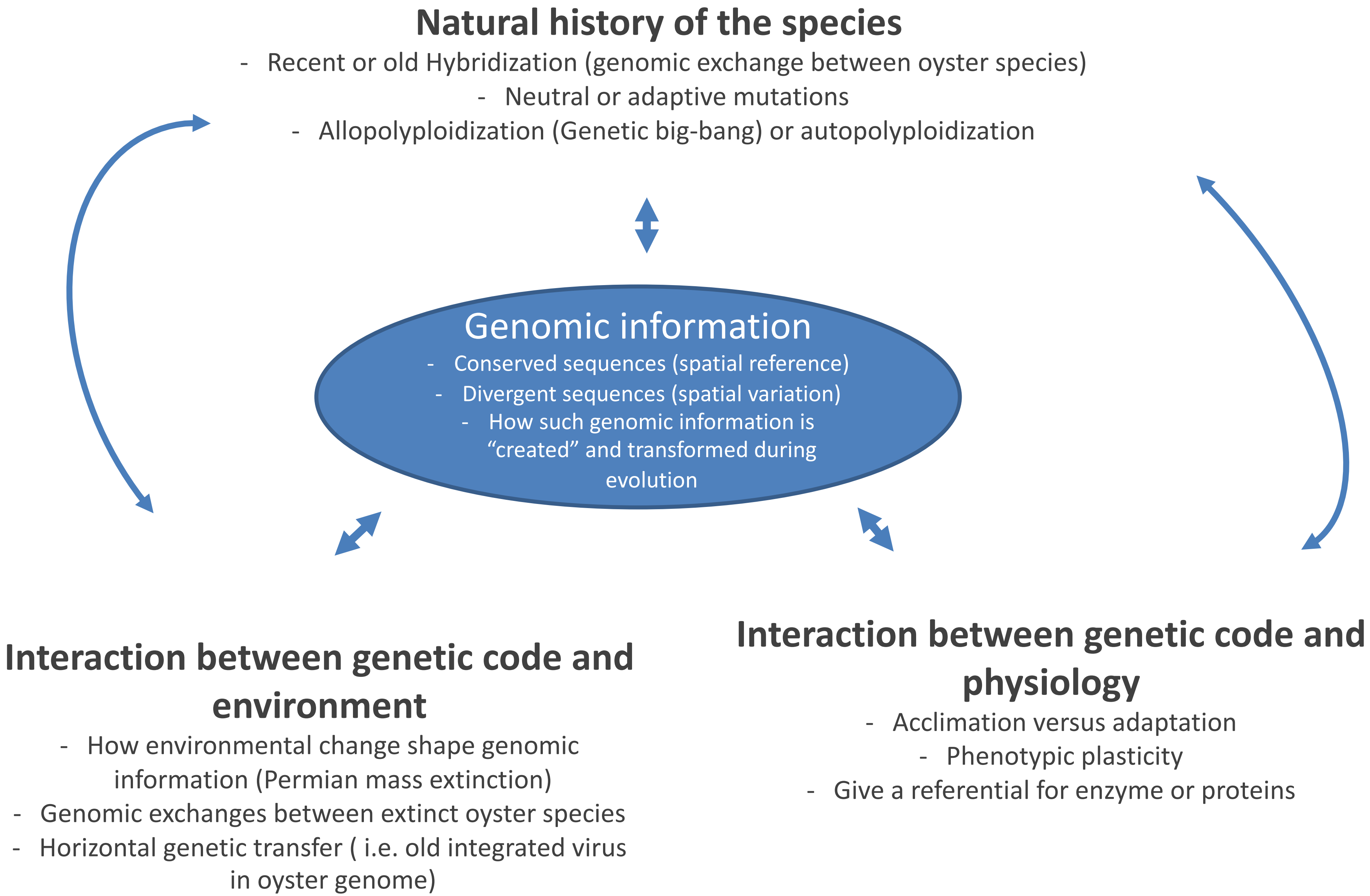
- Network of Bioinformatics Correspondents @IFREMER
- Annual Bioinformatics Meeting
- Thematic training

GigaStore2

-

Assemblage *de novo* résolu au niveau des haplotype du génome de *Crassostrea gigas* (huitre creuse)

Biological questions



Natural history of the species

- Recent or old Hybridization (genomic exchange between oyster species)
- Neutral or adaptive mutations
- Allopolyploidization (Genetic big-bang) or autopolyploidization

Genomic information

- Conserved sequences (spatial reference)
- Divergent sequences (spatial variation)
- How such genomic information is "created" and transformed during evolution

Interaction between genetic code and environment

- How environmental change shape genomic information (Permian mass extinction)
- Genomic exchanges between extinct oyster species
- Horizontal genetic transfer (i.e. old integrated virus in oyster genome)

Interaction between genetic code and physiology

- Acclimation versus adaptation
- Phenotypic plasticity
- Give a referential for enzyme or proteins

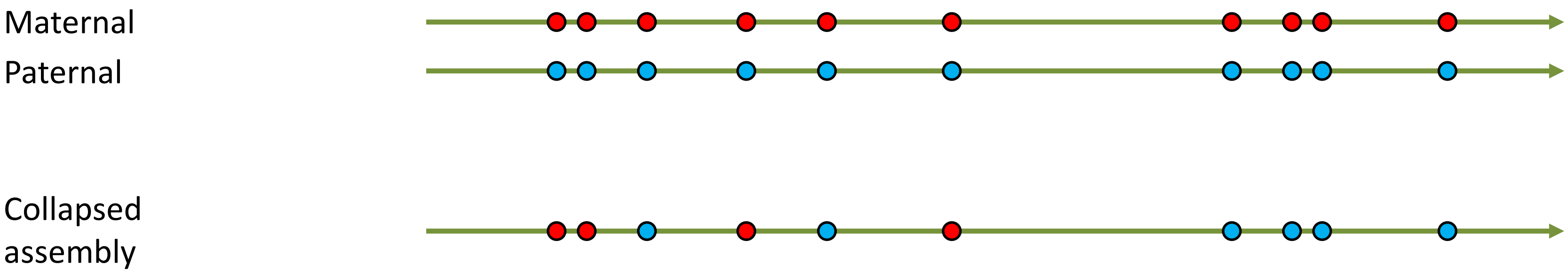
GigaStore1 to GigaStore2

	BGI-Shenzhen	Northwestern Polytechnical University	Institute of Oceanology, Chinese Academy of Sciences	The Roslin Institute
Level	Scaffold	Contig	Chromosome	Chromosome
Sequencing	Illumina	Oxford Nanopore + Illumina	PacBio Sequel I + Illumina	PacBio Sequel I + Illumina
Contigs	5 530	3 676	10	234
Total length	561 804 531	587 503 506	586 856 703	647 883 482
N50 (bp)	290 825	581 941	60 957 391	58 462 999

“Only” unphased genome assemblies with 2 at chromosome level

GigaStore2 → provide a fully phased chromosome scale assembly of *C. gigas*

Current limitation of diploid genome assemblies



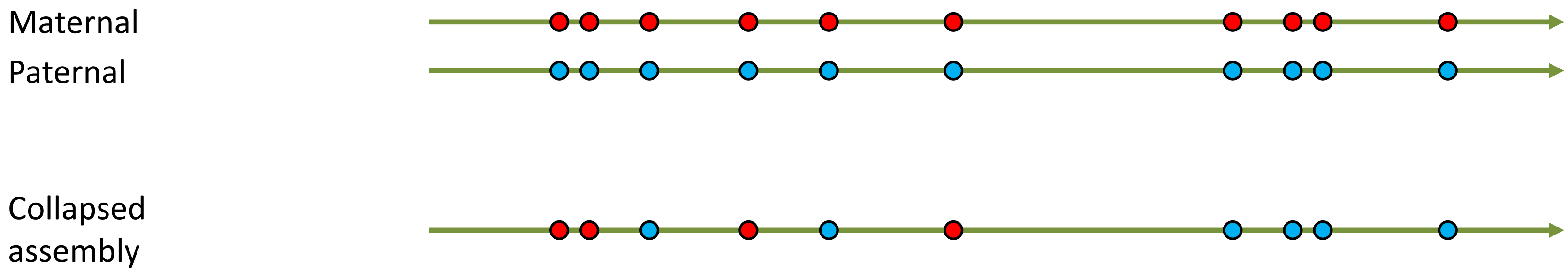
Maternal

Paternal

Collapsed assembly

Heng Li’s blog: lh3.github.io/2021/04/17/concepts-in-phased-assemblies

Current limitation of diploid genome assemblies



Maternal

Paternal

Collapsed assembly

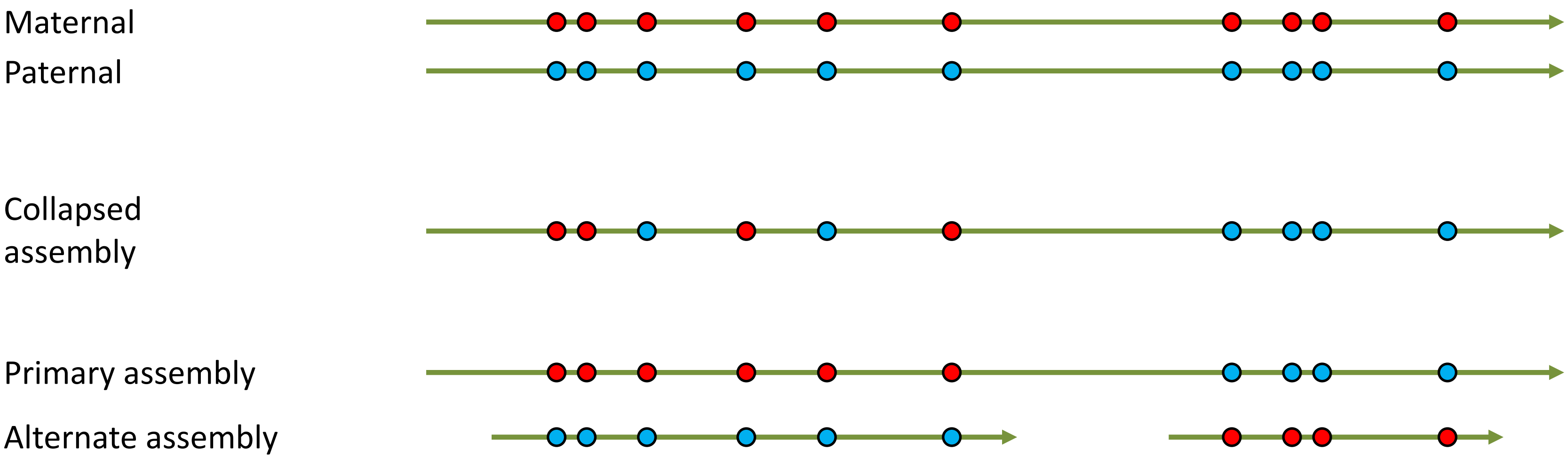
Incomplete picture of genetic variation, problematic for post-analysis

- Loss of half of heterozygous variations
- Errors in heterozygote regions
- May lead to inflated assembly for **species with high heterozygosity**

Sequence of each copy is essential to correctly understand allele-specific DNA methylation and gene expression, to analyse evolution, forensics, genetic diseases…

Tewhey, R., Bansal, V., Torkamani, A. *et al.* The importance of phase information for human genomics. *Nat Rev Genet* **12,** 215–223 (2011)
Vinson JP, Jaffe DB, O'Neill K, et al. Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res*. 2005;15(8):1127-1135

Current limitation of diploid genome assemblies



Maternal

Paternal

Collapsed assembly

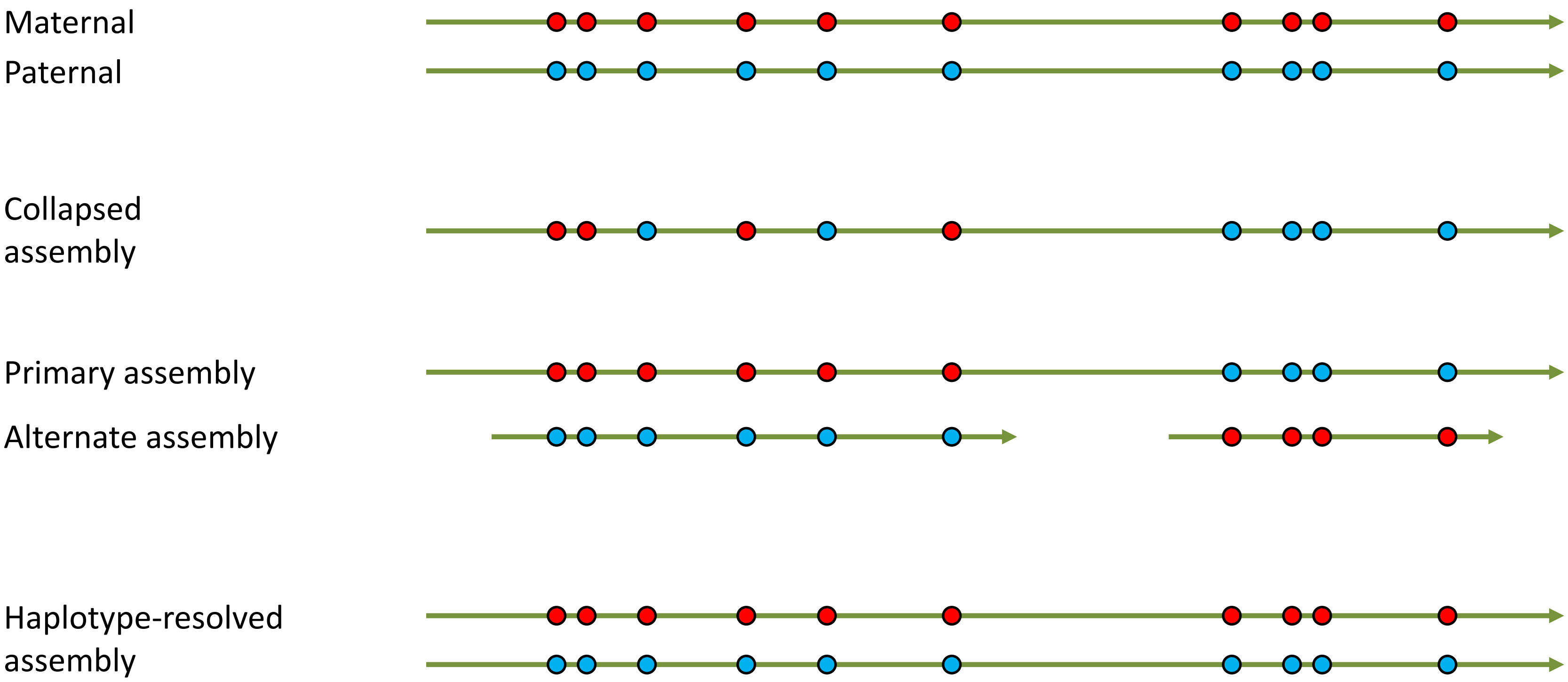
Primary assembly

Alternate assembly

Assembly of haplotigs in heterozygous regions

Alternate assembly, not useful by itself as it is fragmented and incomplete

Current limitation of diploid genome assemblies



Maternal

Paternal

Collapsed assembly

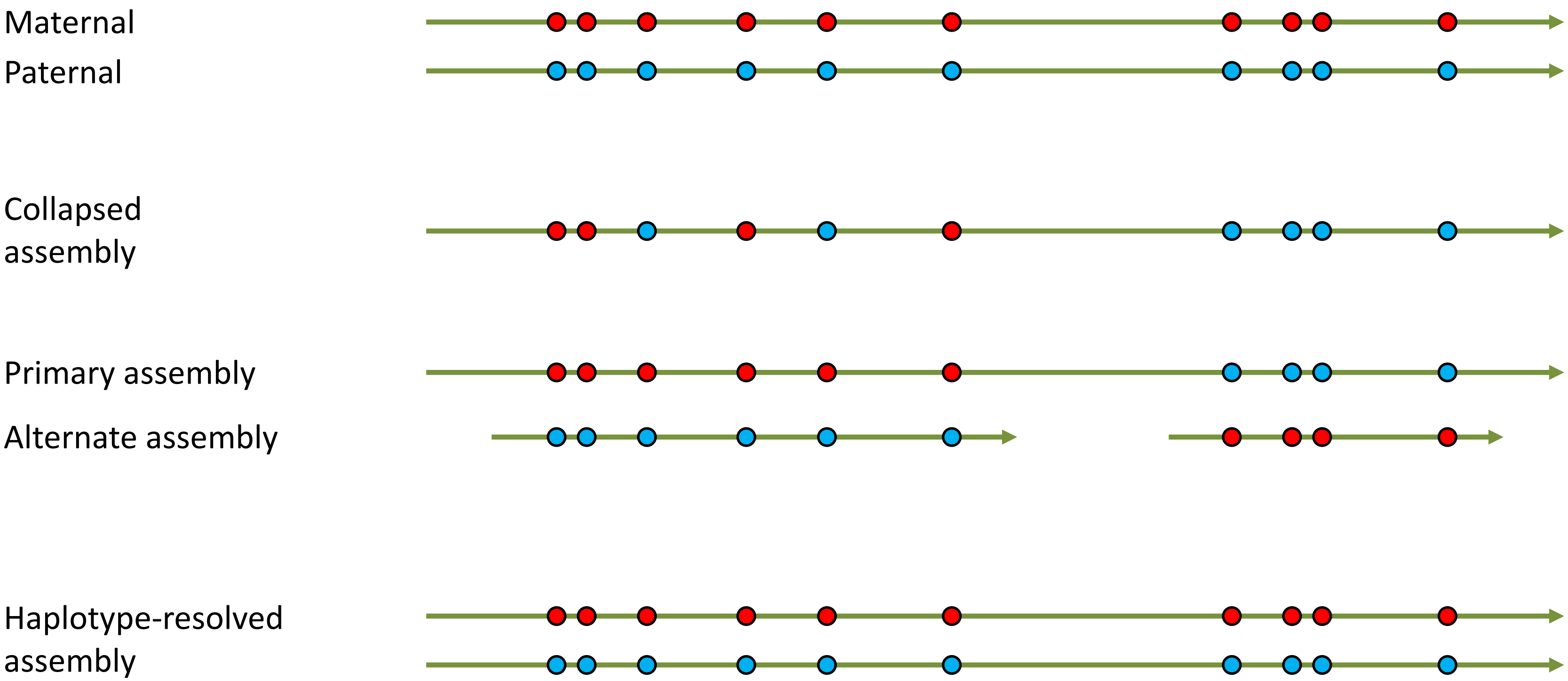
Primary assembly

Alternate assembly

Haplotype-resolved assembly

Each haplotype assembled separately → **true and complete picture of genetic variation**

Current limitation of diploid genome assemblies



Maternal

Paternal

Collapsed assembly

Primary assembly

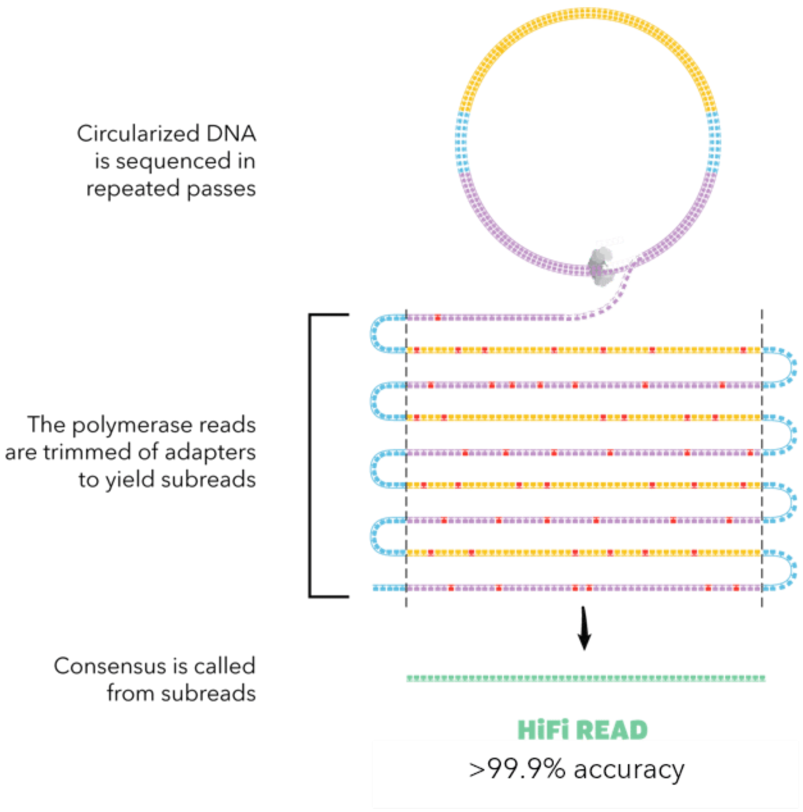
Alternate assembly

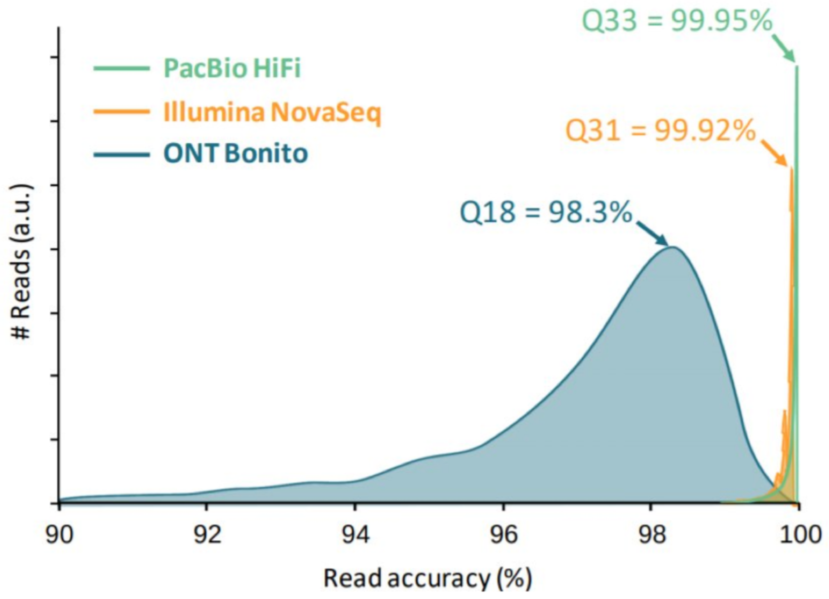
Haplotype-resolved assembly

Each haplotype assembled separately → **true and complete picture of genetic variation**

PacBio Hifi reads allows easily to generate fully phased diploid assembly

PacBio HiFi sequencing technology

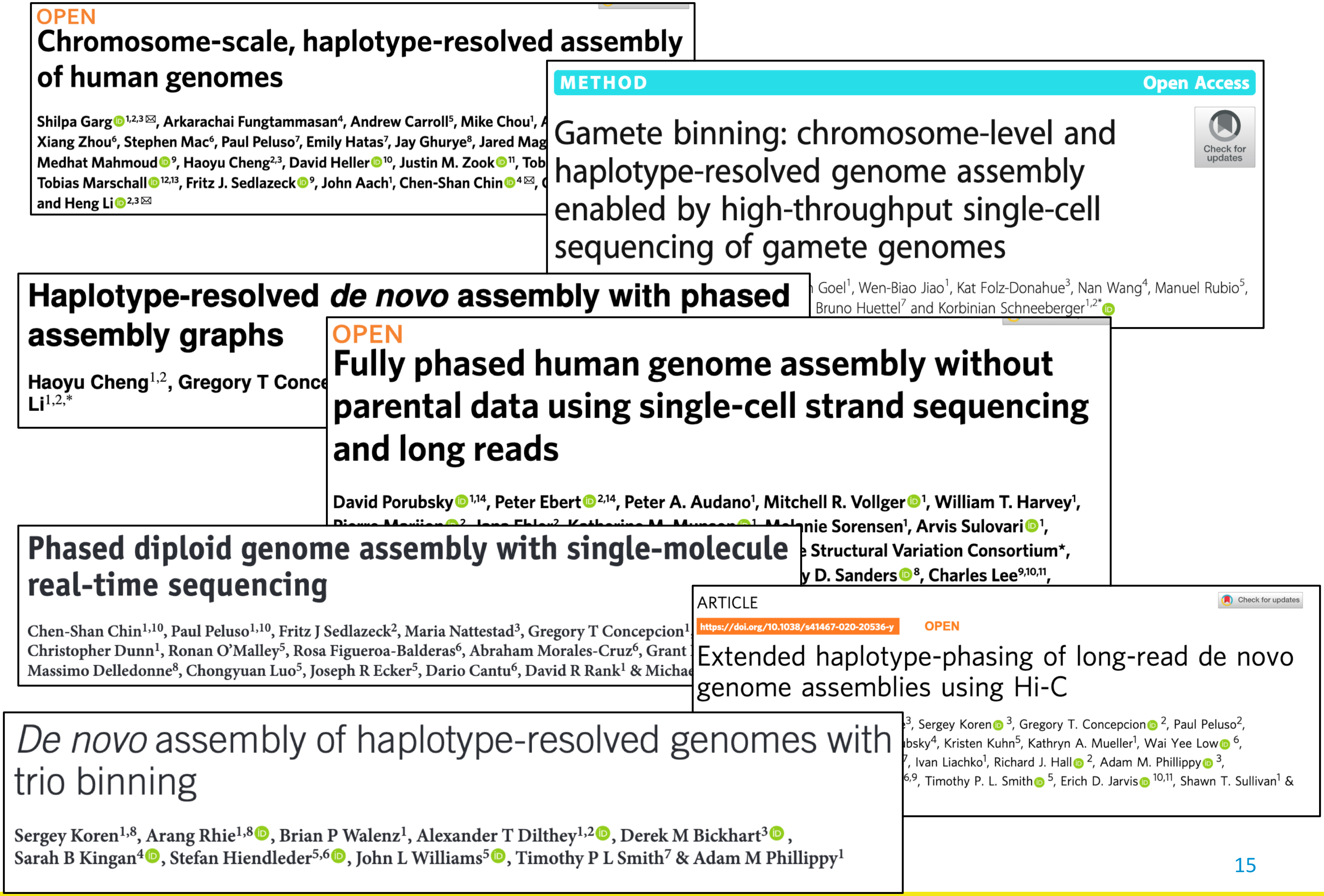

Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus is called from subreads
HiFi READ
>99.9% accuracy



HiFi read qualification:

- Minimal number of pass: **3**
- Minimal Qscore: **20**

Strategies to obtain haplotype phased genome assembly



OPEN

Chromosome-scale, haplotype-resolved assembly of human genomes

Shilpa Garg[1,2,3], Arkarachai Fungtammasan[4], Andrew Carroll[5], Mike Chou[1],
Xiang Zhou[6], Stephen Mac[6], Paul Peluso[7], Emily Hatas[7], Jay Ghurye[8], Jared Mag
Medhat Mahmoud[9], Haoyu Cheng[2,3], David Heller[10], Justin M. Zook[11], Tob
Tobias Marschall[12,13], Fritz J. Sedlazeck[9], John Aach[1], Chen-Shan Chin[4],
and Heng Li[2,3]

METHOD — Open Access

Check for updates

Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes

Goel[1], Wen-Biao Jiao[1], Kat Folz-Donahue[3], Nan Wang[4], Manuel Rubio[5],
Bruno Huettel[7] and Korbinian Schneeberger[1,2*]

Haplotype-resolved *de novo* assembly with phased assembly graphs

Haoyu Cheng[1,2], Gregory T Conce
Li[1,2,*]

OPEN

Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads

David Porubsky[1,14], Peter Ebert[2,14], Peter A. Audano[1], Mitchell R. Vollger[1], William T. Harvey[1],
Melanie Sorensen[1], Arvis Sulovari[1],
Structural Variation Consortium*,
y D. Sanders[8], Charles Lee[9,10,11],

Phased diploid genome assembly with single-molecule real-time sequencing

Chen-Shan Chin[1,10], Paul Peluso[1,10], Fritz J Sedlazeck[2], Maria Nattestad[3], Gregory T Concepcion[1],
Christopher Dunn[1], Ronan O'Malley[5], Rosa Figueroa-Balderas[6], Abraham Morales-Cruz[6], Grant
Massimo Delledonne[8], Chongyuan Luo[5], Joseph R Ecker[5], Dario Cantu[6], David R Rank[1] & Michae

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-020-20536-y — OPEN

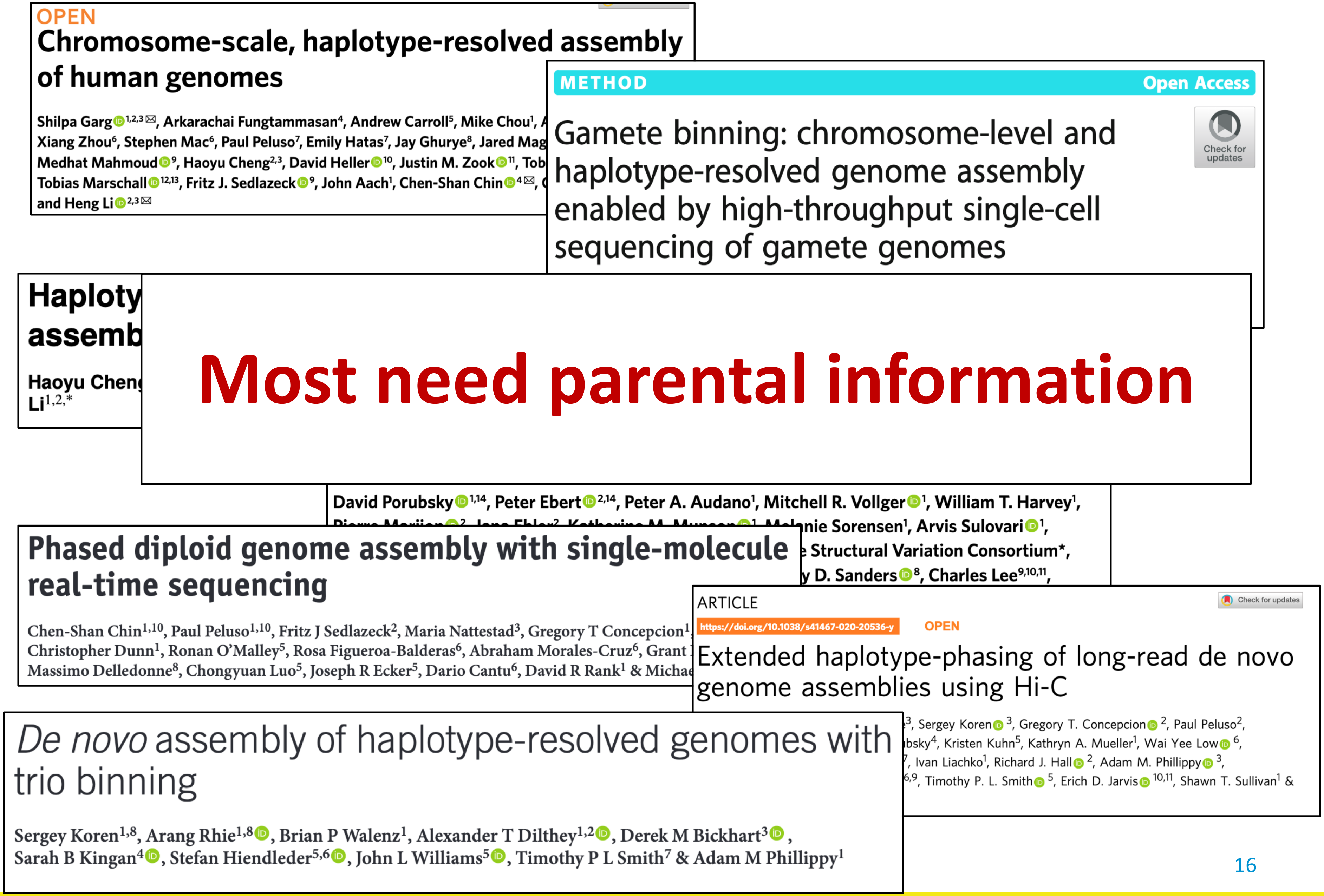
Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C

[3], Sergey Koren[3], Gregory T. Concepcion[2], Paul Peluso[2],
bsky[4], Kristen Kuhn[5], Kathryn A. Mueller[1], Wai Yee Low[6],
[7], Ivan Liachko[1], Richard J. Hall[2], Adam M. Phillippy[3],
[6,9], Timothy P. L. Smith[5], Erich D. Jarvis[10,11], Shawn T. Sullivan[1] &

De novo assembly of haplotype-resolved genomes with trio binning

Sergey Koren[1,8], Arang Rhie[1,8], Brian P Walenz[1], Alexander T Dilthey[1,2], Derek M Bickhart[3],
Sarah B Kingan[4], Stefan Hiendleder[5,6], John L Williams[5], Timothy P L Smith[7] & Adam M Phillippy[1]

Strategies to obtain haplotype phased genome assembly



OPEN

Chromosome-scale, haplotype-resolved assembly of human genomes

Shilpa Garg[1,2,3]✉, Arkarachai Fungtammasan[4], Andrew Carroll[5], Mike Chou[1], A
Xiang Zhou[6], Stephen Mac[6], Paul Peluso[7], Emily Hatas[7], Jay Ghurye[8], Jared Mag
Medhat Mahmoud[9], Haoyu Cheng[2,3], David Heller[10], Justin M. Zook[11], Tob
Tobias Marschall[12,13], Fritz J. Sedlazeck[9], John Aach[1], Chen-Shan Chin[4]✉, (
and Heng Li[2,3]✉

METHOD Open Access

Check for updates

Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes

Haploty
assemb

Haoyu Chen
Li[1,2,*]

Most need parental information

David Porubsky[1,14], Peter Ebert[2,14], Peter A. Audano[1], Mitchell R. Vollger[1], William T. Harvey[1],
Pierre Marijon[2], Jana Ebler[2], Katherine M. Munson[1], Melanie Sorensen[1], Arvis Sulovari[1],
e Structural Variation Consortium*,
y D. Sanders[8], Charles Lee[9,10,11],

Phased diploid genome assembly with single-molecule real-time sequencing

Chen-Shan Chin[1,10], Paul Peluso[1,10], Fritz J Sedlazeck[2], Maria Nattestad[3], Gregory T Concepcion[1],
Christopher Dunn[1], Ronan O'Malley[5], Rosa Figueroa-Balderas[6], Abraham Morales-Cruz[6], Grant
Massimo Delledonne[8], Chongyuan Luo[5], Joseph R Ecker[5], Dario Cantu[6], David R Rank[1] & Michae

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-020-20536-y OPEN

Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C

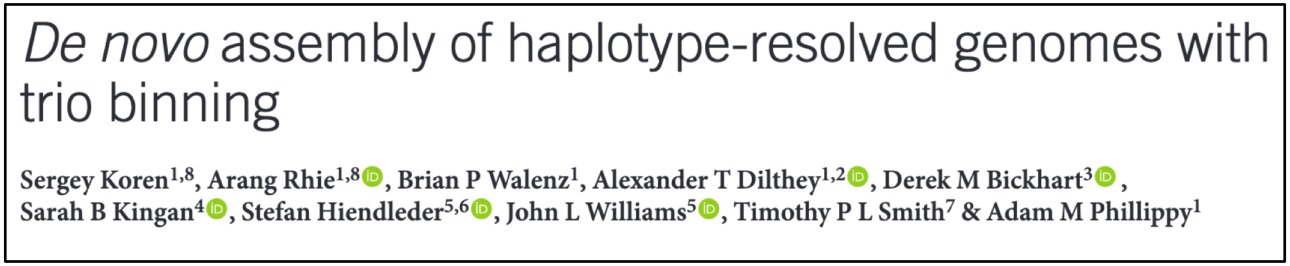
[3], Sergey Koren[3], Gregory T. Concepcion[2], Paul Peluso[2],
bsky[4], Kristen Kuhn[5], Kathryn A. Mueller[1], Wai Yee Low[6],
[7], Ivan Liachko[1], Richard J. Hall[2], Adam M. Phillippy[3],
[6,9], Timothy P. L. Smith[5], Erich D. Jarvis[10,11], Shawn T. Sullivan[1] &

De novo assembly of haplotype-resolved genomes with trio binning

Sergey Koren[1,8], Arang Rhie[1,8], Brian P Walenz[1], Alexander T Dilthey[1,2], Derek M Bickhart[3],
Sarah B Kingan[4], Stefan Hiendleder[5,6], John L Williams[5], Timothy P L Smith[7] & Adam M Phillippy[1]

Strategies to obtain haplotype phased genome assembly

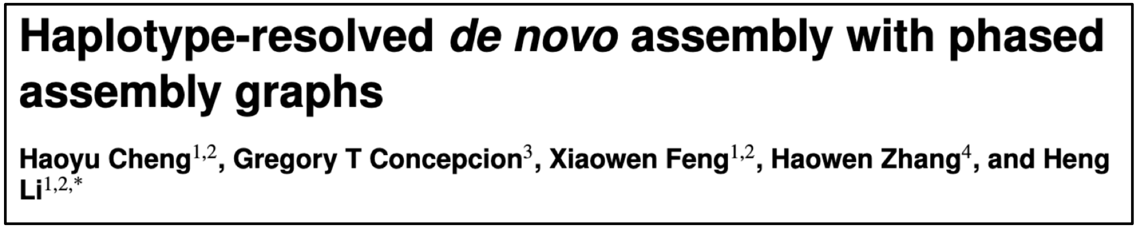
Trio Binning



De novo assembly of haplotype-resolved genomes with trio binning

Sergey Koren[1,8], Arang Rhie[1,8], Brian P Walenz[1], Alexander T Dilthey[1,2], Derek M Bickhart[3], Sarah B Kingan[4], Stefan Hiendleder[5,6], John L Williams[5], Timothy P L Smith[7] & Adam M Phillippy[1]

Canu/HiCanu



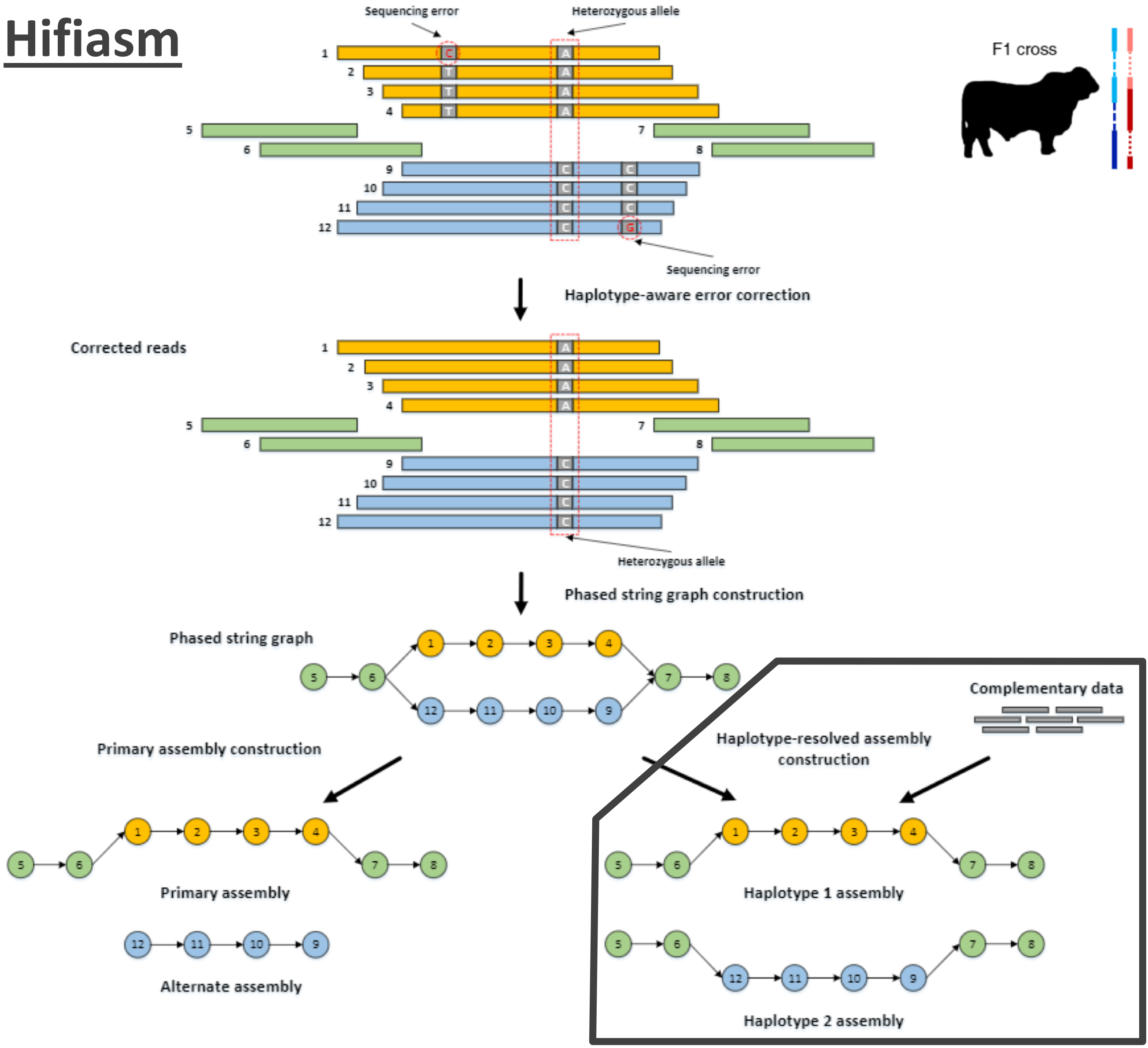
Haplotype-resolved *de novo* assembly with phased assembly graphs

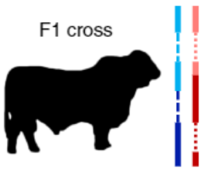
Haoyu Cheng[1,2], Gregory T Concepcion[3], Xiaowen Feng[1,2], Haowen Zhang[4], and Heng Li[1,2,*]

Hifiasm

Strategies to obtain haplotype phased genome assembly

Hifiasm



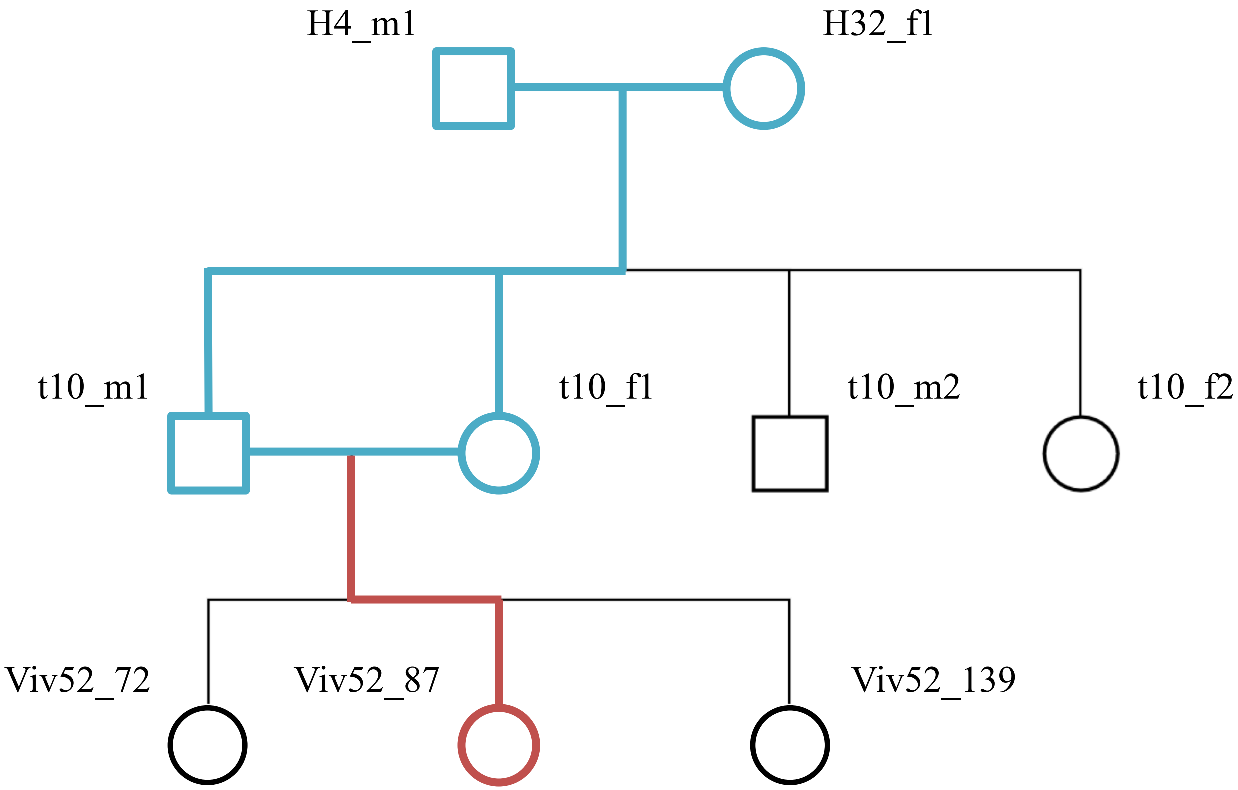


1 - Single assembly and graph phasing

2 – Full phasing using parental data DNA-seq

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, **18**:170-175.

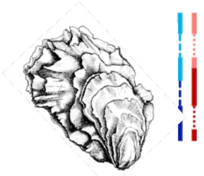
Sampled individuals



Checking:

- Parentage assignment on all the offspring's
- Flux cytometry – Autotriploid (3N) in the offspring's…

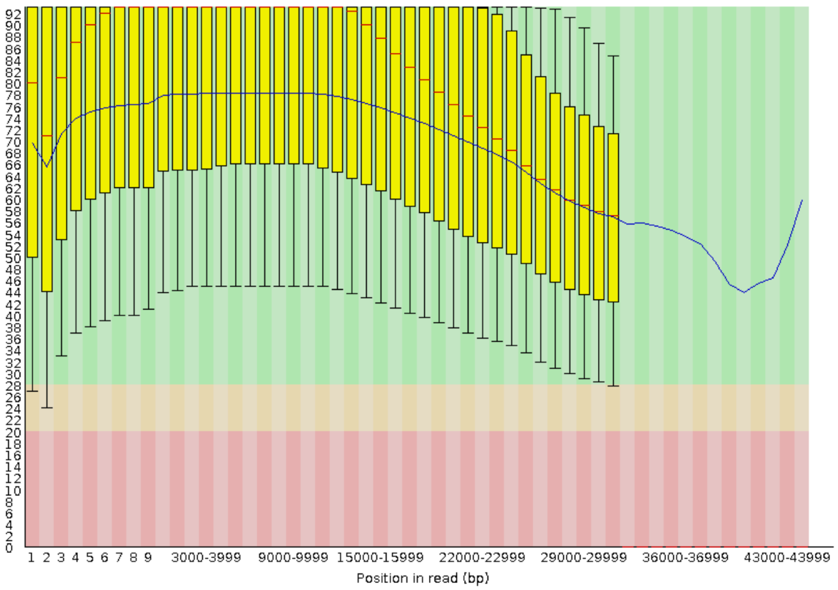
F1 generation – PacBio Hifi



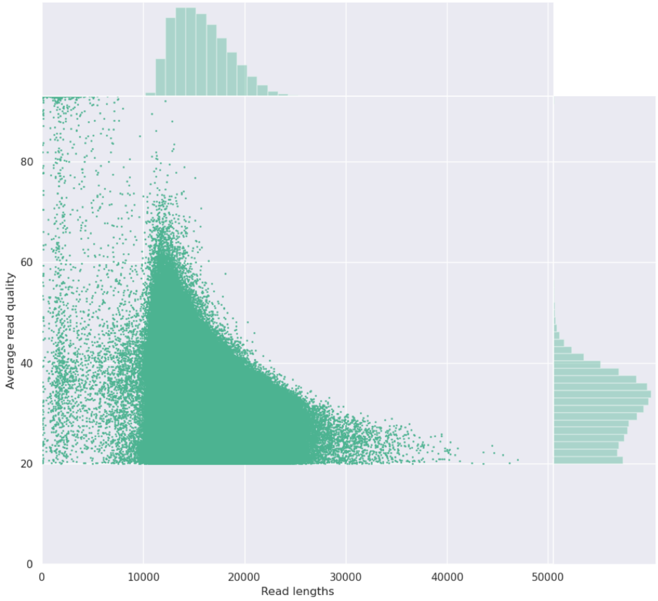
GS2VIV52-87m

	Reads	Gbp	N50 (bp)	Coverage
m64122_210120_191426	791 989	12 193 385 003	15 520	22
m64122_210214_033118	1 006 503	15 847 937 897	15 965	26
m64244_210612_174252	1 218 551	19 343 585 808	16 093	32

Quality score across all bases

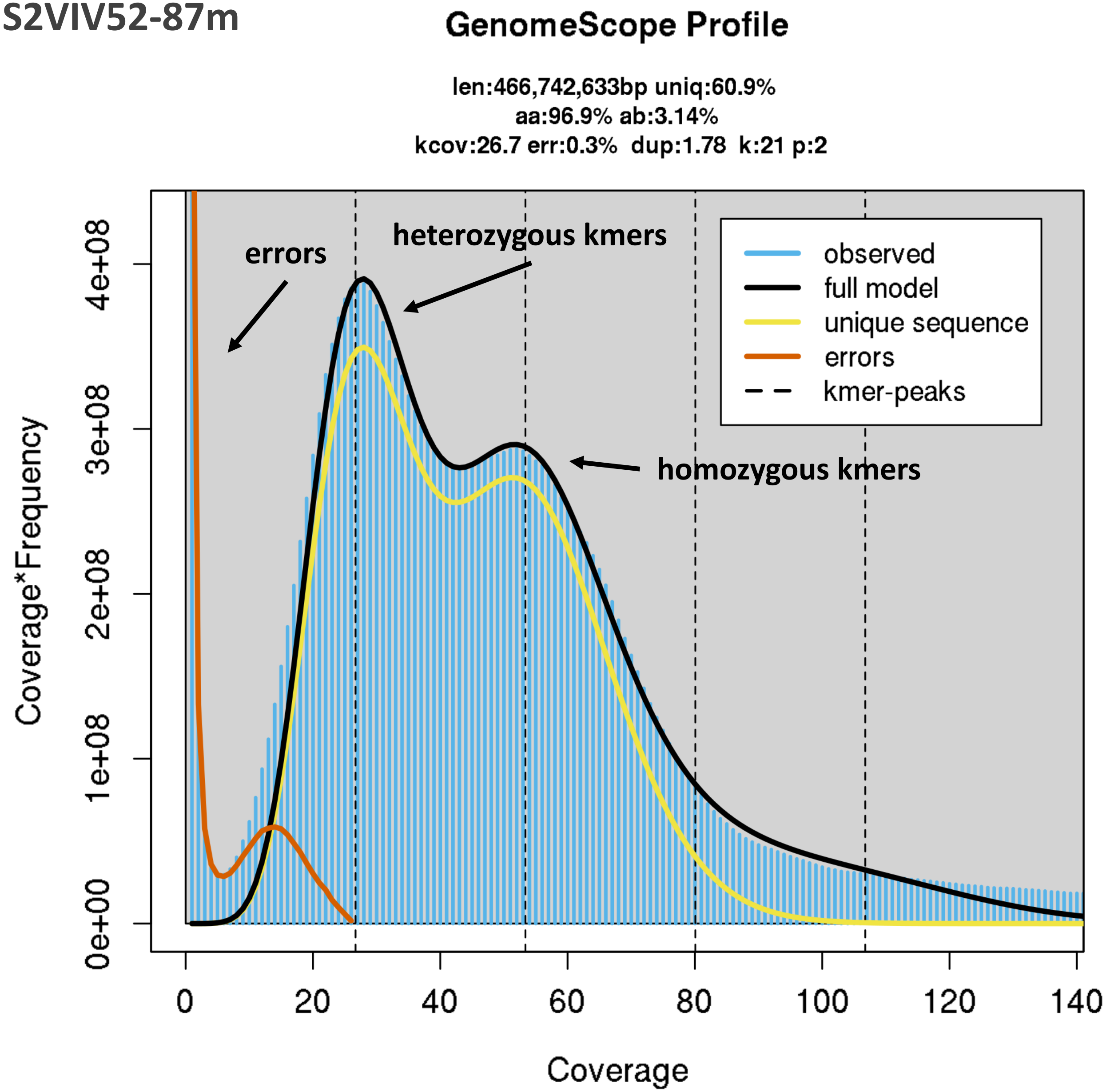


Read length vs Average read quality



F1 generation – PacBio Hifi

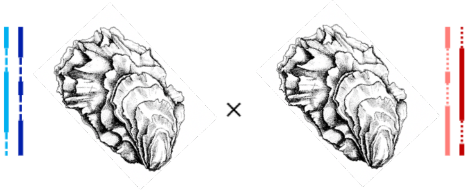
GS2VIV52-87m



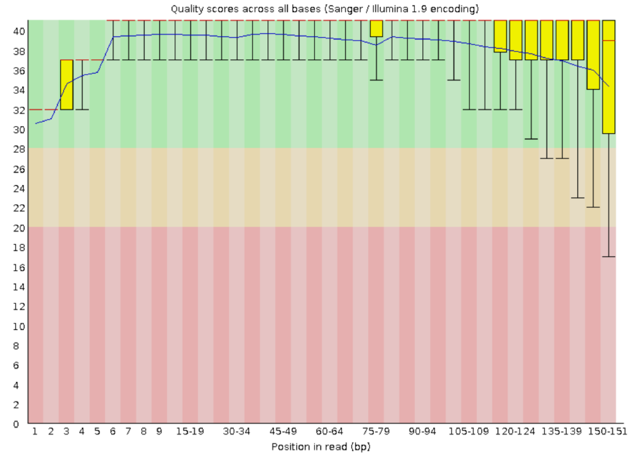
Homozygous (aa)	96.9%
Heterozygous (ab)	**3.1%**
Genome Haploid	467 Mbp
Genome Repeat	182 Mbp
Genome Unique	284 Mbp

Parental generation – Illumina

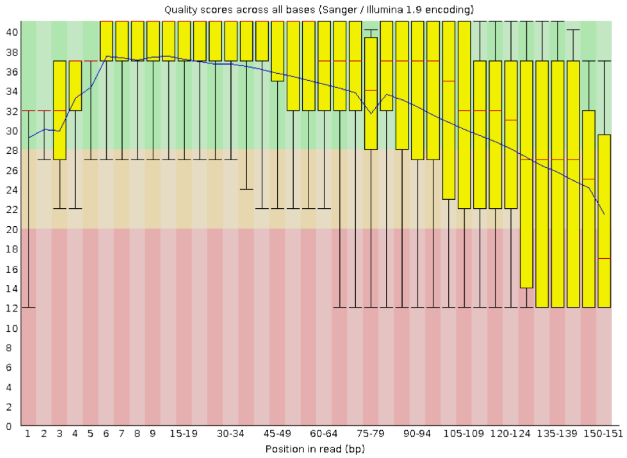
LIGAN
GENOMIC PLATFORM
LILLE INTEGRATED GENOMIC ADVANCED NETWORK



Per base sequence quality: **raw data R1**



Per base sequence quality: **raw data R2**



		Raw reads				**Clean reads (fastp)**			
	Samples	Reads	Gbp	Coverage	Coverage / ind	Reads	Gbp	Coverage	Coverage / ind
♀	t10-f1-a	178 530 702	26 468 838 432	48	112	145 219 932	19 435 943 632	35	82
	t10-f1-b	196 020 960	29 202 970 540	53		155 869 662	20 663 292 209	38	
	t10-f1-v	39 794 326	5 891 657 116	11		35 27 4982	4 973 583 965	9	
♂	t10-m1-a	135 797 024	19 968 062 329	36	88	108 255 872	14 311 258 750	26	64
	t10-m1-b	125 758 558	18 233 590 963	33		96 224 928	12 589 337 628	23	
	t10-m1-v	68 117 642	10 001 854 454	18		59 551 596	8 417 026 407	15	

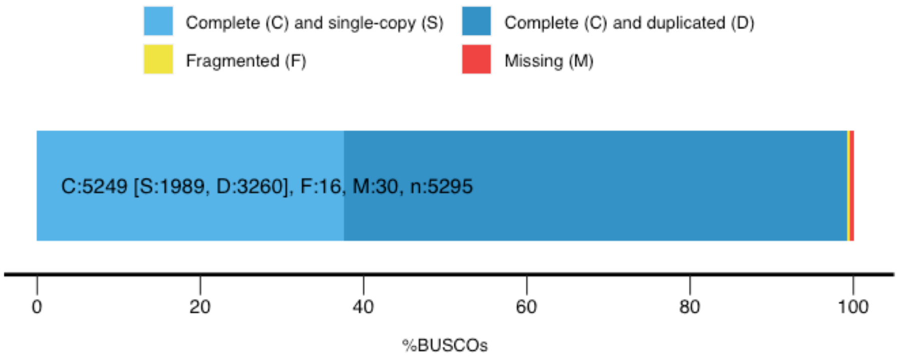
Huge loss of coverage + lower coverage for male haplotype

Not enough to correctly discriminate haplotypes

Primary genome assembly - Hifiasm

# contigs	836
Largest contig	33 768 982
Total length	**1 016 854 405**
N50	5 350 846

BUSCO (v5.2.2 - mollusca_odb10)

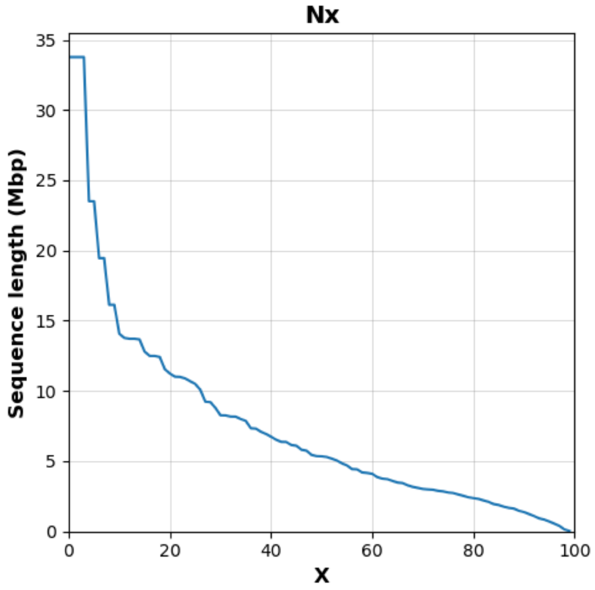


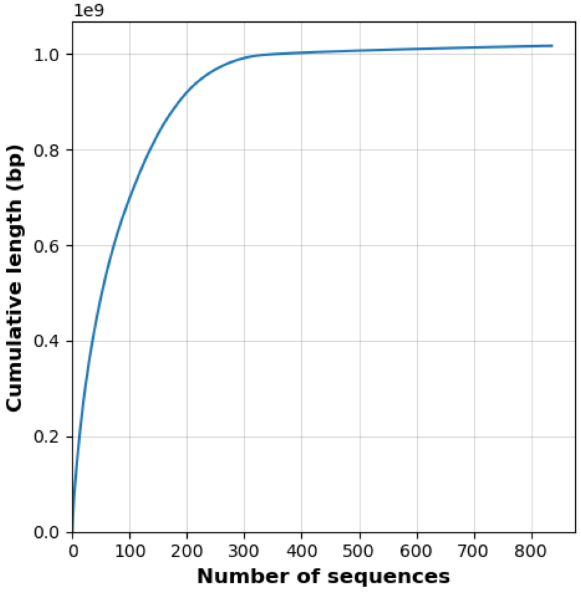
Genome size twice as expected
More than 50% of duplicated core genes

Hifi reads + high heterozygosity rate

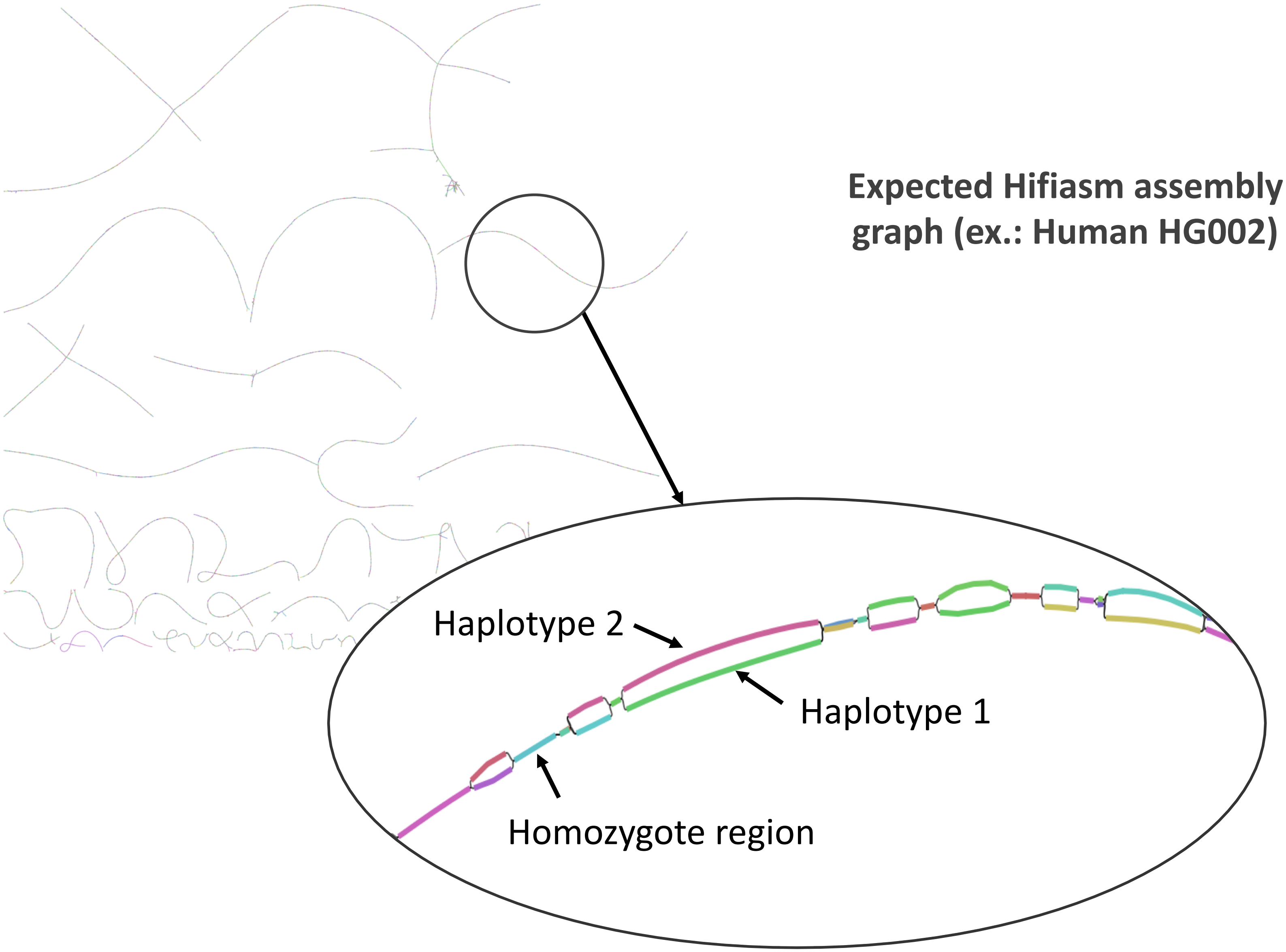
↓

Direct assembly of both haplotypes





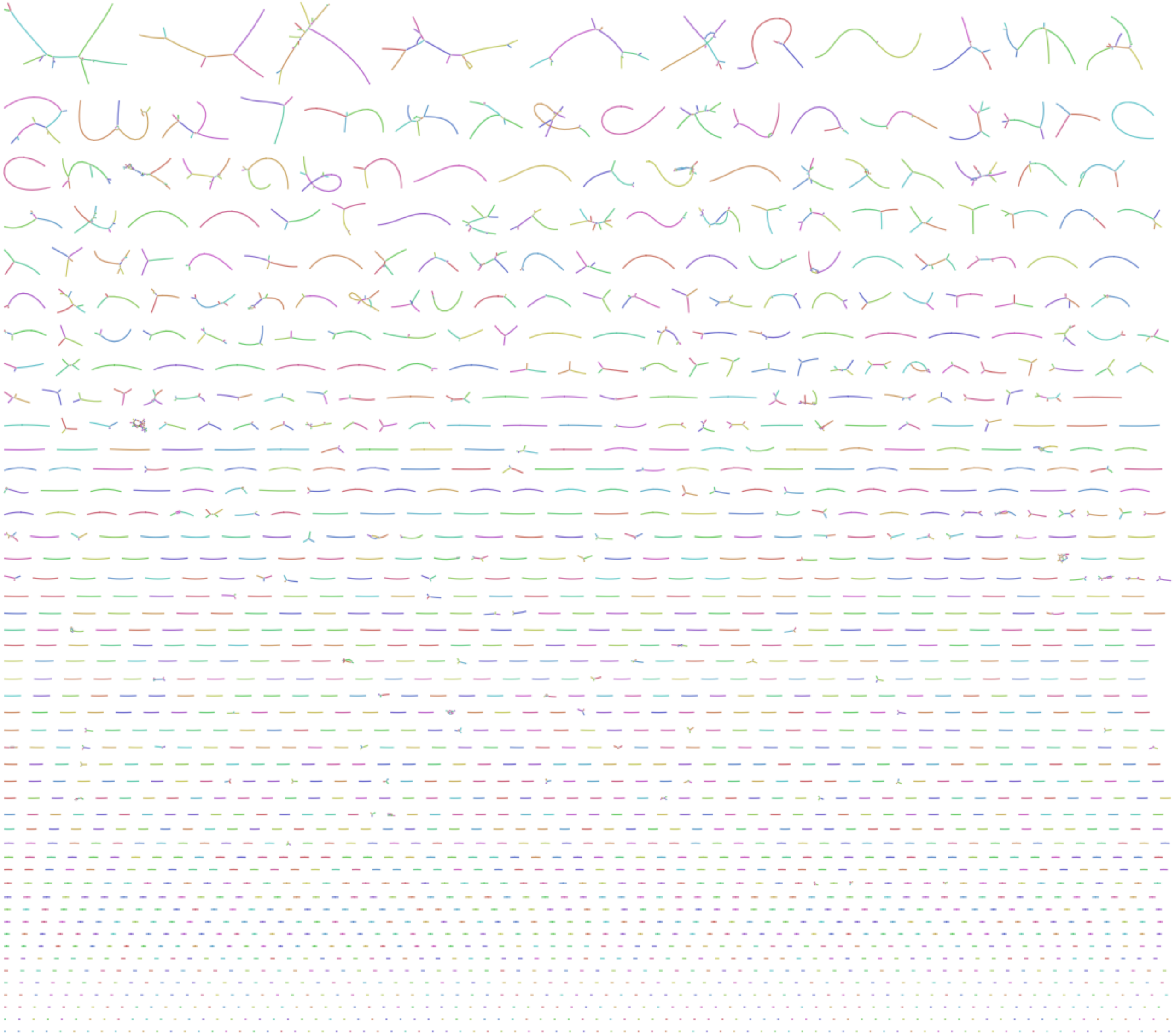
Primary genome assembly - Hifiasm



Expected Hifiasm assembly graph (ex.: Human HG002)

Primary genome assembly - Hifiasm

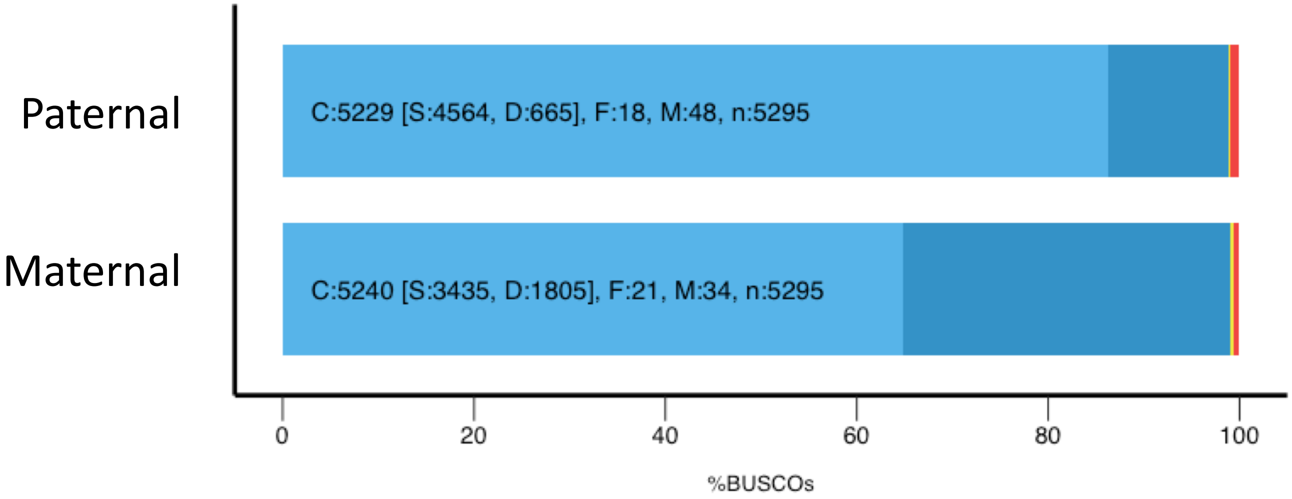
Hifiasm graph for *C. gigas*

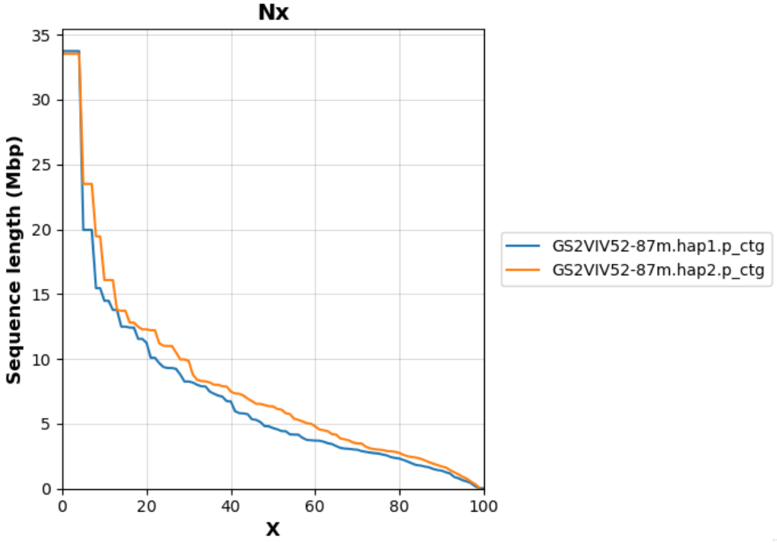


Phased genome assemblies - Hifiasm

	Paternal	Maternal
# contigs	561	464
Largest contig	33 738 176	33 540 007
Total length	**699 233 093**	**769 755 738**
N50	4 671 686	6 344 792

BUSCO (v5.2.2 - mollusca_odb10)





Still duplicated core genes

↓

Need an increase parental k-mer coverage

Conclusion

PacBio HiFi give excellent results (contiguity, completeness, mis-assemblies) even a low coverage for a lower computational cost

Still challenging for highly heterozygous species

* Cheng, H. *et al.* Robust haplotype-resolved assembly of diploid individuals without parental data. *Arxiv* (2021).

Conclusion

PacBio HiFi give excellent results (contiguity, completeness, mis-assemblies) even a low coverage for a lower computational cost

Still challenging for highly heterozygous species

Parental data are not longer necessary to obtain a fully phased genome (Hi-C module in Hifiasm) but results remains a little bit better with*

→ Switch to Hi-C phasing module for new genome projects

* Cheng, H. *et al.* Robust haplotype-resolved assembly of diploid individuals without parental data. *Arxiv* (2021).

Acknowledgements and collaborations

Zootechnics Hatchery

Lionel Degremont

Elise Maurouard

Wet laboratory

Florence Cornette

Serge Heurtebise

Abdellah Benabdelmouna

Sequencing

Marie Gislard

Celine Lopez-Roques

Julien Derop

Bioinformatics analysis

Me ;)

Romain Koszul

Christophe Klopp

Jérémie Vidal-Dupiol

Funding & coordinations

Jean-Baptiste Lamy

Me ;)

Guillaume Mitta

Sylvie Lapégue

Pierre Boudry

Pierre-Alexandre Gagnaire

Guillaume Rivière

Romain Koszul

Yannick Gueguen

Ifremer

IHPE
Interactions-Hôtes-Pathogènes-Environnements

Institut Pasteur

INRAE

isem
Institut des Sciences de l'Evolution-Montpellier

BOREA
Biologie des organismes et écosystèmes aquatiques

Genotoul GeT

LIGAN
GENOMIC PLATFORM
Lille Integrated Genomic Advanced Network

UNIVERSITÉ PERPIGNAN VIA DOMITIA

cnrs