De novo sequencing and assembly of plant genomes using nanopore long reads

Jean-Marc Aury

jmaury@genoscope.cns.fr

@J_M_Aury

ONT workshop, 01/15/2019

Genoscope overview

- **French National Sequencing Center located near Paris and lead by Patrick Wincker, created in 1997 and part of the CEA since 2007.**

- **Provide high-throughput sequencing data to the Academic community, and carry out in-house genomic projects**

- **Focus on biodiversity : *de novo* sequencing and metagenomic projects (TaraOceans)**

http://www.genoscope.cns.fr

Vitis vinifera
(grape)

Quercus robur
(oak)

Musa acuminata
(banana)

Brassica napus
(seed rape)

Sequencing capacities

2 Illumina HiSeq 2500

2 Illumina HiSeq 4000

1 Illumina NovaSeq

2 MiSeq

6 Oxford Nanopore MkI

1 PromethION

BIONANO GENOMICS

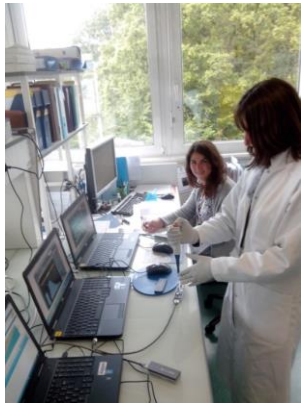
1 Saphyr System

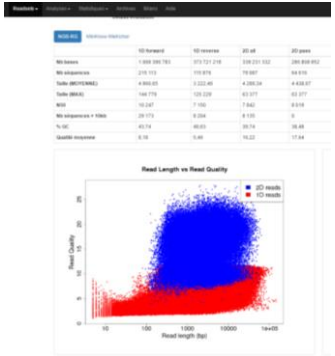
MinIOn sequencing at Genoscope

- **>1,000 MinION and >80 PromethION flowcells ; >100 different organisms; ~3.5 Tb of ONT reads ; DNA and RNA samples**

- ***de novo* assembly (22 yeast strains ~12Mb, 4 fungi genomes ~30Mb, several bacterial genomes, 15 plant genomes of 400-1200Mb) and gene prediction**

- **Software development : error correction tool**





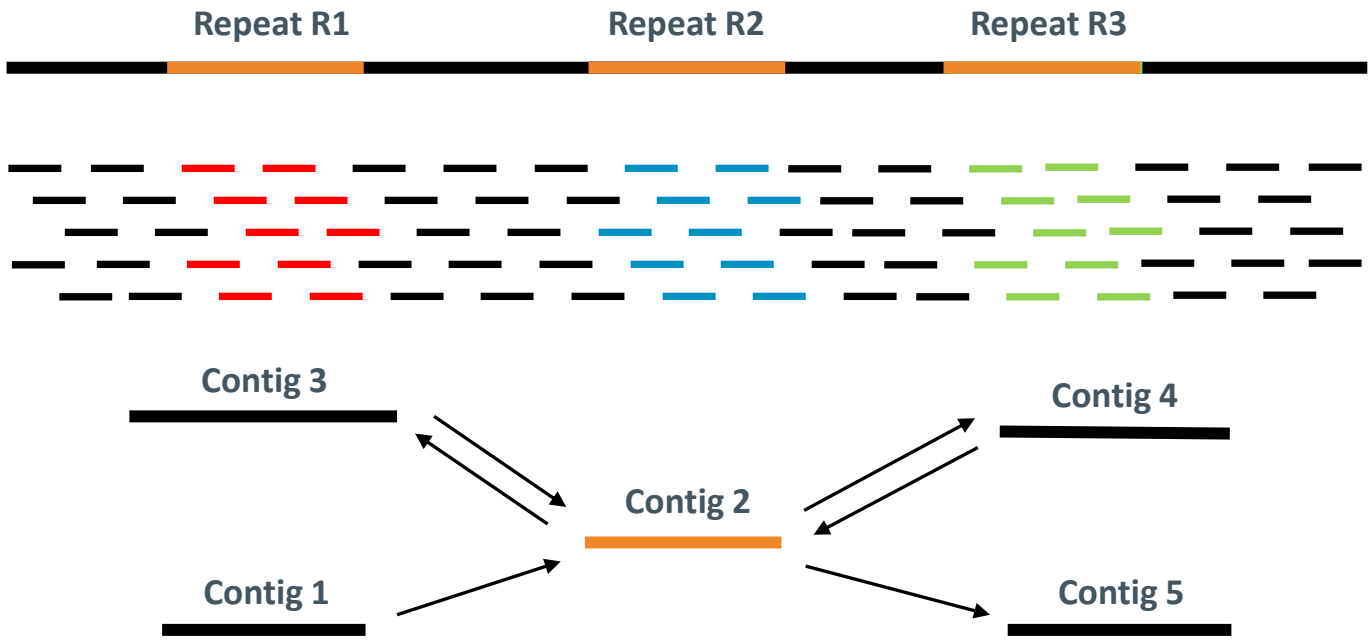
BMC Genomics

Genome assembly using Nanopore-guided long and error-free DNA reads

Mohammed-Amin Madoui†, Stefan Engelen†, Corinne Cruaud, Caroline Belser, Laurie Bertrand, Adriana Alberti, Arnaud Lemainque, Patrick Wincker and Jean-Marc Aury

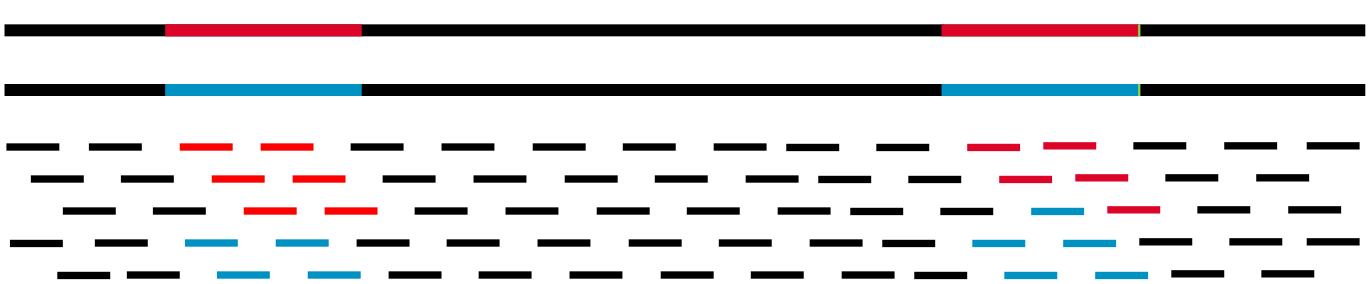
http://www.genoscope.cns.fr/nas

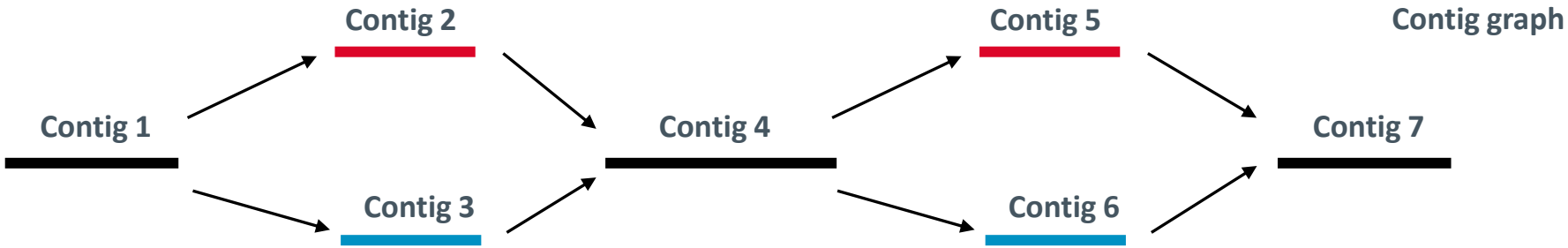
Genome assembly difficulties



=> Repetitive regions lead to fragmented assemblies and under-estimate repeat content

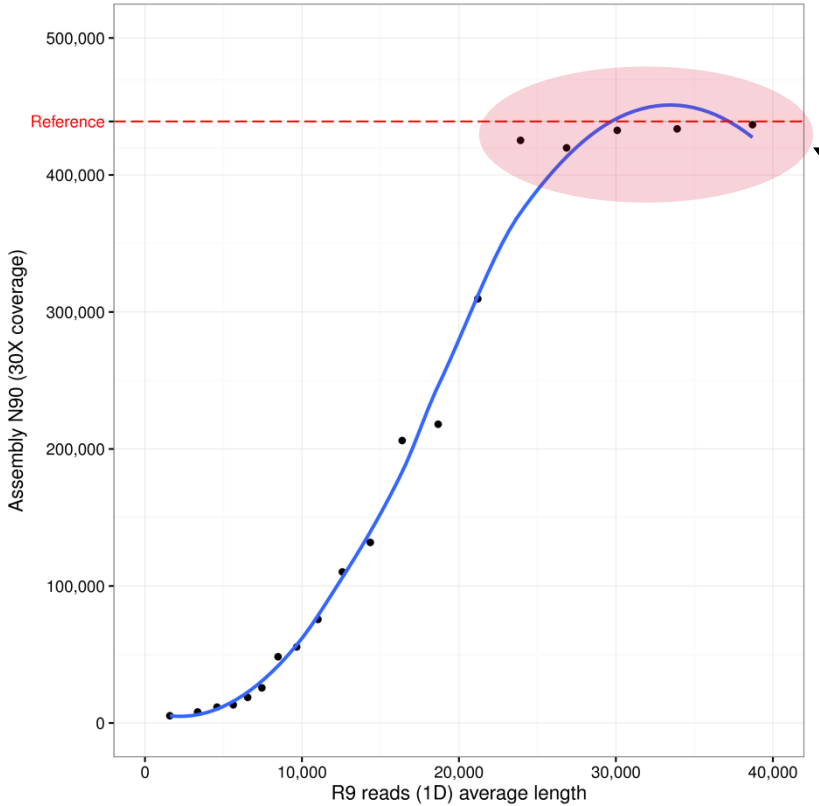
Genome assembly difficulties





=> Heterozygous regions lead to fragmented assemblies and cause allelic duplication (over-estimate the size of the haploid genome)

Read Length Matters



1 contig per chromosome assemblies

***de novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer**

Benjamin Istace, Anne Friedrich, Léo d'Agata, Sébastien Faye, Emilie Payen, Odette Beluche, Claudia Caradec, Sabrina Davidas, Corinne Cruaud, Gianni Liti Arnaud Lemainque, Stefan Engelen, Patrick Wincker, Joseph Schacherer ✉, Jean-Marc Aury ✉

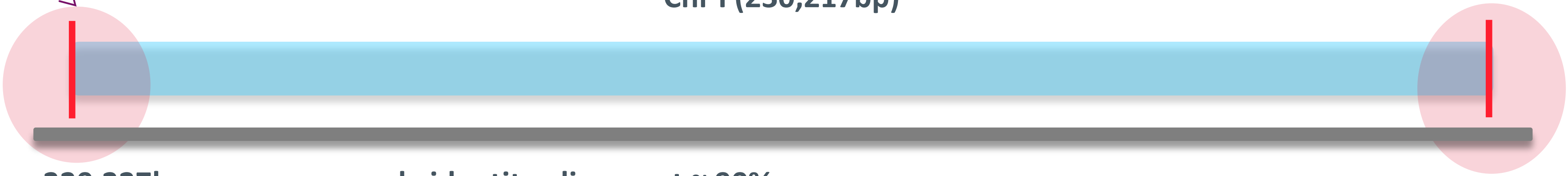
=> Yeast genome assembly is resolved when using 30X of 25Kb reads in average

Nanopore : a fast evolving technology

Chromosomes can be captured entirely, the example read span yeast chromosome 1 from telomere to telomere

```
                                                       ccaca.cca.cacccacacacccacacaccacaccacacaccacaccac
cattagcttcgttccagt .. 150 nt .. accccccacaccacccaccacacccacacccaccacccac.cacccacac.ccaca.cac.caccacaccac
                                                             ^   ^       ^         ^     ^   ^
```

Chr I (230,217bp)



220,227bp nanopore read ; identity alignment ~ 90%

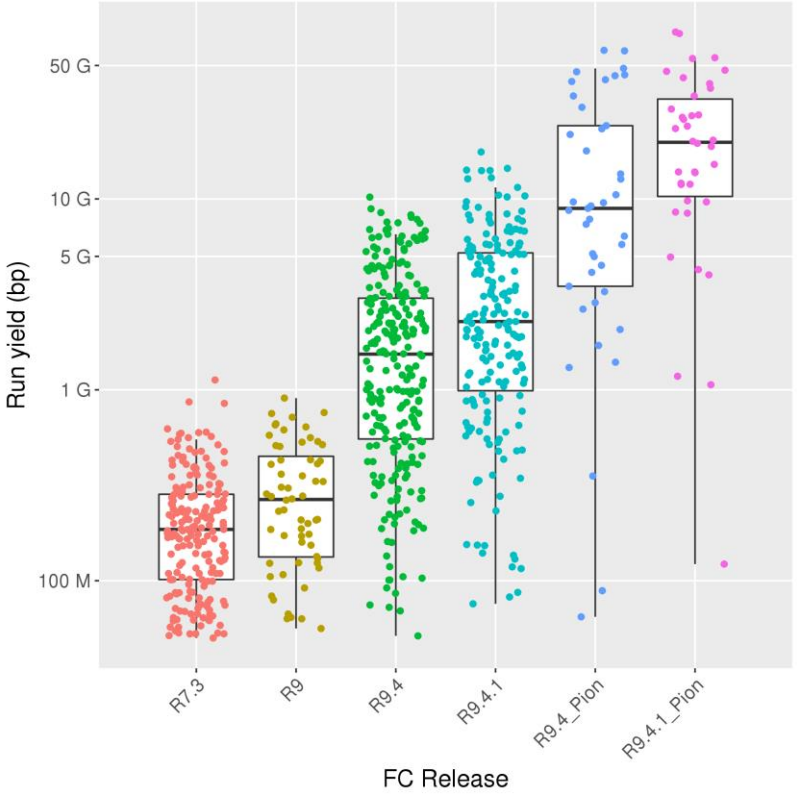
```
ggtgtgggtgtggtgtggtgtgtgggtgtggtgtgggtgtggtgtgtg
ggtgtaggtgtggtgtggtgtgtgggtgtggtgtg.gtgtggtgtgggtgtgggtgtattgtgggtgtgg .. 200 nt .. gtgtgggtgtgggtgtgtgtggt
    ^                              ^          ^
```

The nanopore read is smaller than the chromosome due to deletions

Nanopore : a fast evolving technology

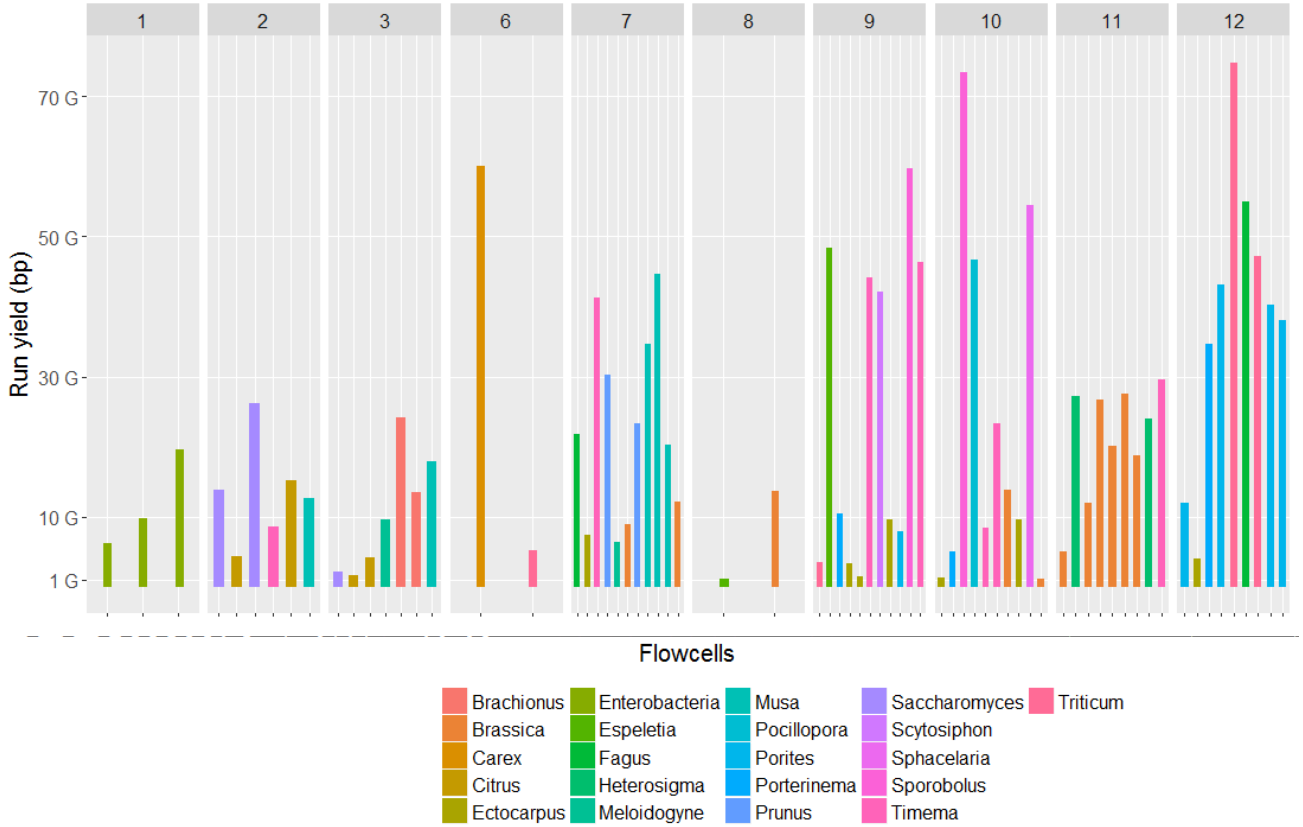
Yield improvement : ~100Mb to several Gb for the MinION and ~10Gb per PromethION flowcell

Throughput is still heterogeneous depending on the DNA sample



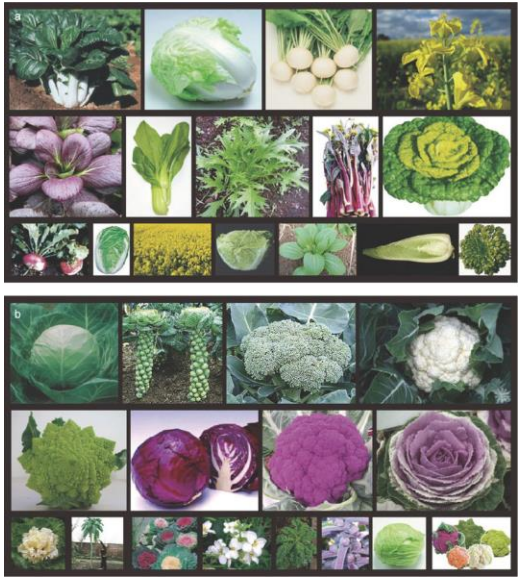
Nanopore : a fast evolving technology

A year of PromethION sequencing : throughput improvement in the last four months



Sequencing of plant genomes using the MinION

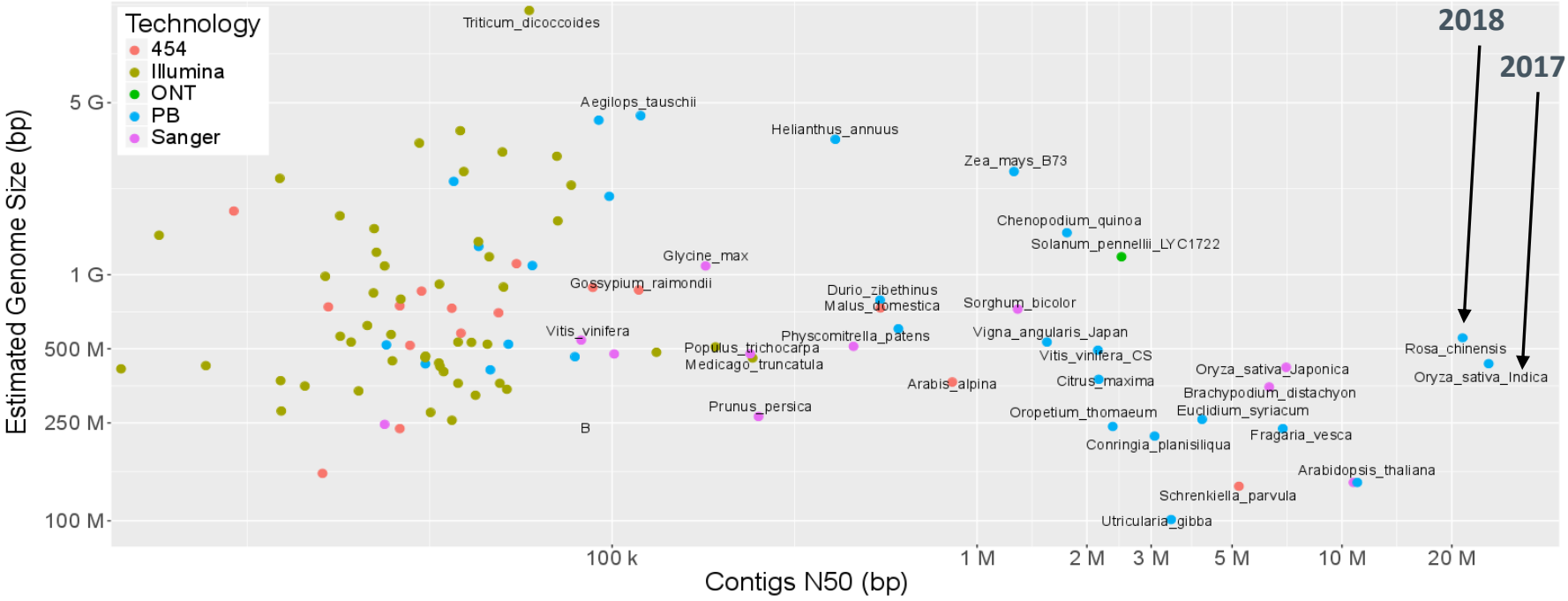
- Large scale genomic projects focused on *Brassica* and *Musa* genomes
- *Brassica* includes important vegetables for human nutrition and are important models for understanding polyploid plants
- The variability between two morphotypes of the same *Brassica* species is high
- *Musa* spp are essential crops in (sub-)tropical countries, and are interesting models for studying reticulate evolution
- In this context, we are currently sequencing 3 *Brassica* and 7 banana genomes.



Genome Triplication Drove the Diversification of Brassica Plants, Cheng et al. 2014

Continuity of current plant genome assemblies

A lot of plant genomes have already been sequenced, but only 6 plant species have an assembly with a contig N50 > 5Mb



http://www.genoscope.cns.fr/genomes

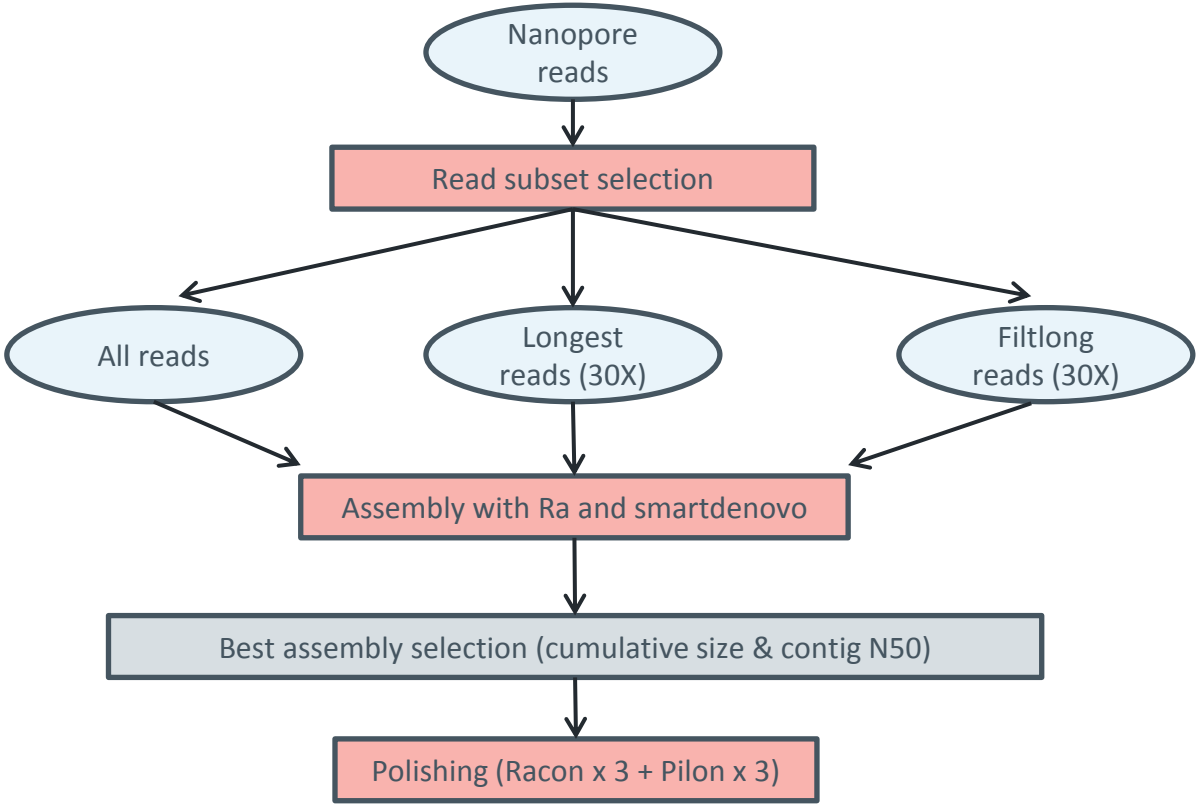
Genome assembly of plant genomes using long and short reads

So far, 2 Brassica and 5 Musa have been sequenced

	Brassica rapa *ssp* Z1	*Brassica oleracea* *ssp* HDEM	*Musa schizocarpa*	*Musa textilis*	*Musa acuminata* ssp zebrina	*Musa acuminata* ssp malaccensis	*Musa acuminata* ssp burmannica
Estimated Genome size	529 Mb	630 Mb	587 Mb	700 Mb	530 Mb	530 Mb	530 Mb
# flowcells	11	10	18	23	46	21	5
Cumul. Size	32 Gb	21 Gb	27 Gb	36 Gb	81 Gb	35 Gb	32 Gb
N50	15 kb	31 kb	24 kb	28 Kb	18 Kb	16 Kb	25 Kb
Coverage	58 X	32 X	51 X	51 X	150 X	66 X	60 X
N50 longest 30X	26 kb	33 kb	32 kb	36 Kb	32 Kb	27 Kb	30 Kb

with the goal of reaching at least 30X coverage and an N50 at 30Kb

Genome assembly process



Genome assembly results

	Brassica rapa *ssp* Z1	*Brassica oleracea* *ssp* HDEM	*Musa schizocarpa*	*Musa textilis*	*Musa acuminata* ssp zebrina	*Musa acuminata* ssp malaccensis	*Musa acuminata* ssp burmannica
Assembler	Ra	Ra	Ra	Smartdenovo	Smartdenovo	Smartdenovo	Ra
Dataset	All reads	All reads	30X fitlong	30X fitlong	30X longest	30X longest	30X longest
# contigs	544	244	437	608	718	427	704
Cumul. Size	375 Mb	546 Mb	527 Mb	601 Mb	510 Mb	477 Mb	481 Mb
N50	3.8 Mb	7.3 Mb	2.1 Mb	3.2 Mb	2.0 Mb	2.7 Mb	1.9 Mb
Max size	21.6 Mb	25.4 Mb	12.8 Mb	21.5 Mb	13.1 Mb	16.0 Mb	11.2 Mb

High contiguity of the assemblies, but insufficient to decipher genome organization at the chromosome-level

BIONANO GENOMICS

Chromosome-scale assemblies

BIONANO GENOMICS

Bionano Direct Label and Stain (DLS) technology

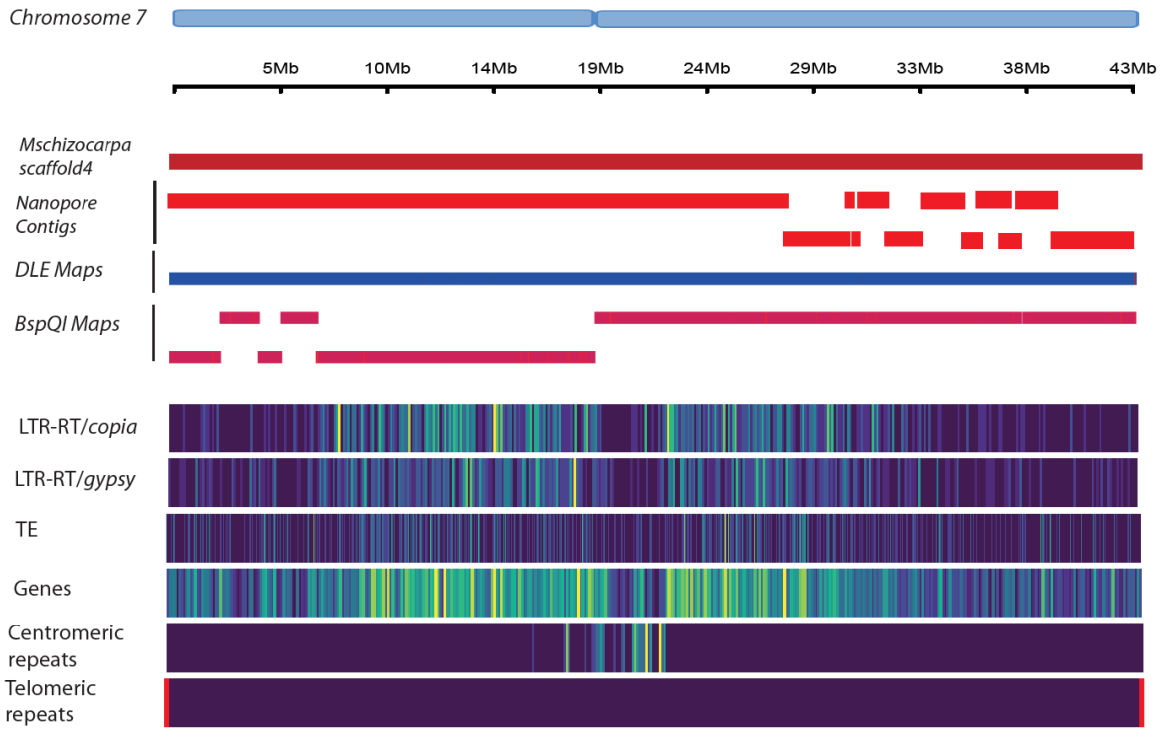
Organization of nanopore contigs using optical maps

	Brassica rapa *ssp* Z1	***Brassica oleracea*** *ssp* HDEM	***Musa schizocarpa***	***Musa acuminata*** *ssp* malaccensis
# scaffolds	335	140	227	144
Cumul. Size (N's)	402 Mb (8.2%)	555 Mb (1.8%)	525 Mb (1.5%)	473 Mb (0.8%)
N50	15.4 Mb	29.5 Mb	36.8 Mb	34.6 Mb
Contig N50 (nanopore assembly)	5.5 Mb (3.8 Mb)	9.5 Mb (7.3 Mb)	6.5 Mb (2.1 Mb)	8.6 Mb (2.7 Mb)
% chromosomes in ≤3 scaffolds	9 /10	8 / 9	11 / 11	11 / 11

Hybrid scaffolding generated chromosome scale assemblies and but also improved the contig N50

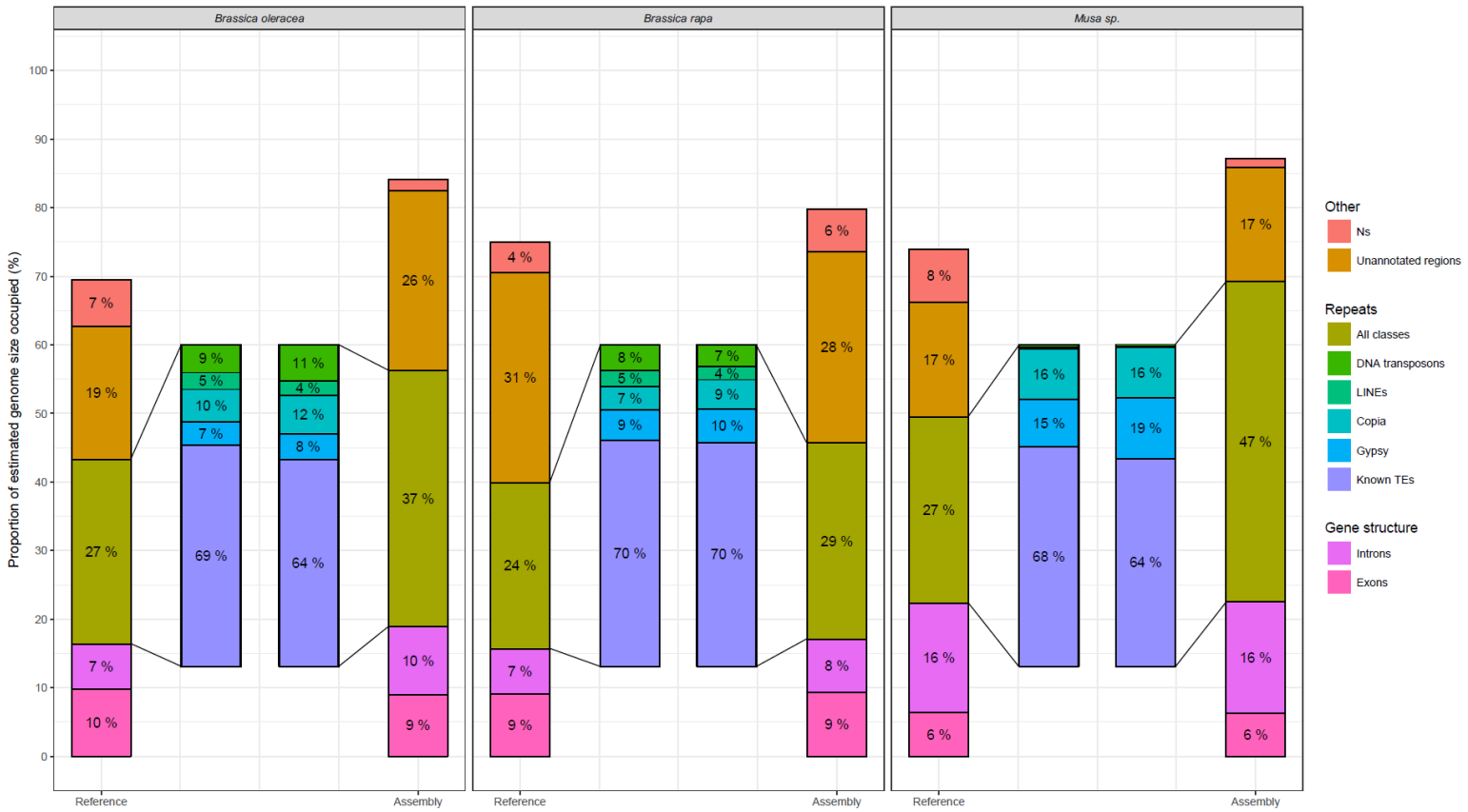
Chromosome-scale assemblies

Schematic view of chromosome 7 from banana genome assembly



Chromosome-scale assemblies

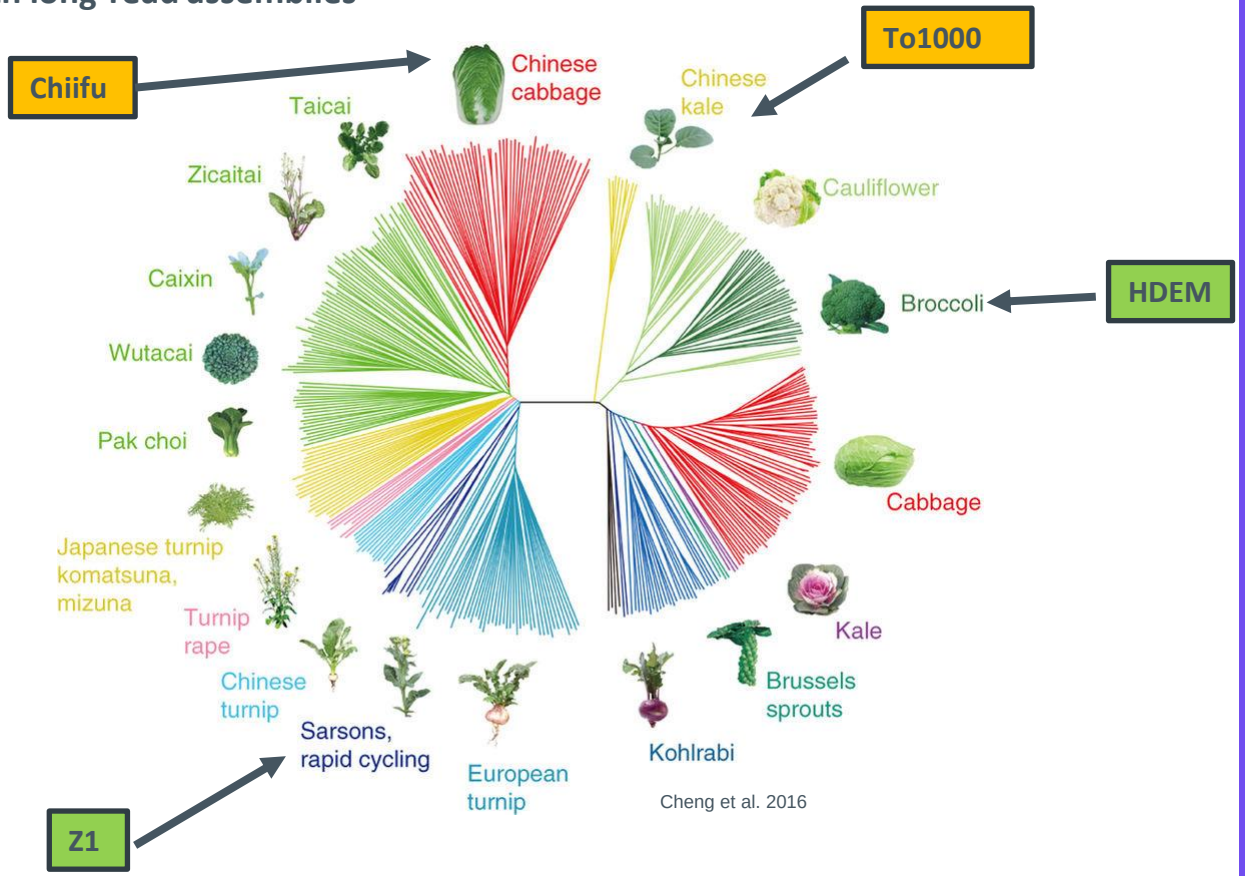
Comparison of existing references with long-read assemblies



Chromosome-scale assemblies

Comparison of existing references with long-read assemblies

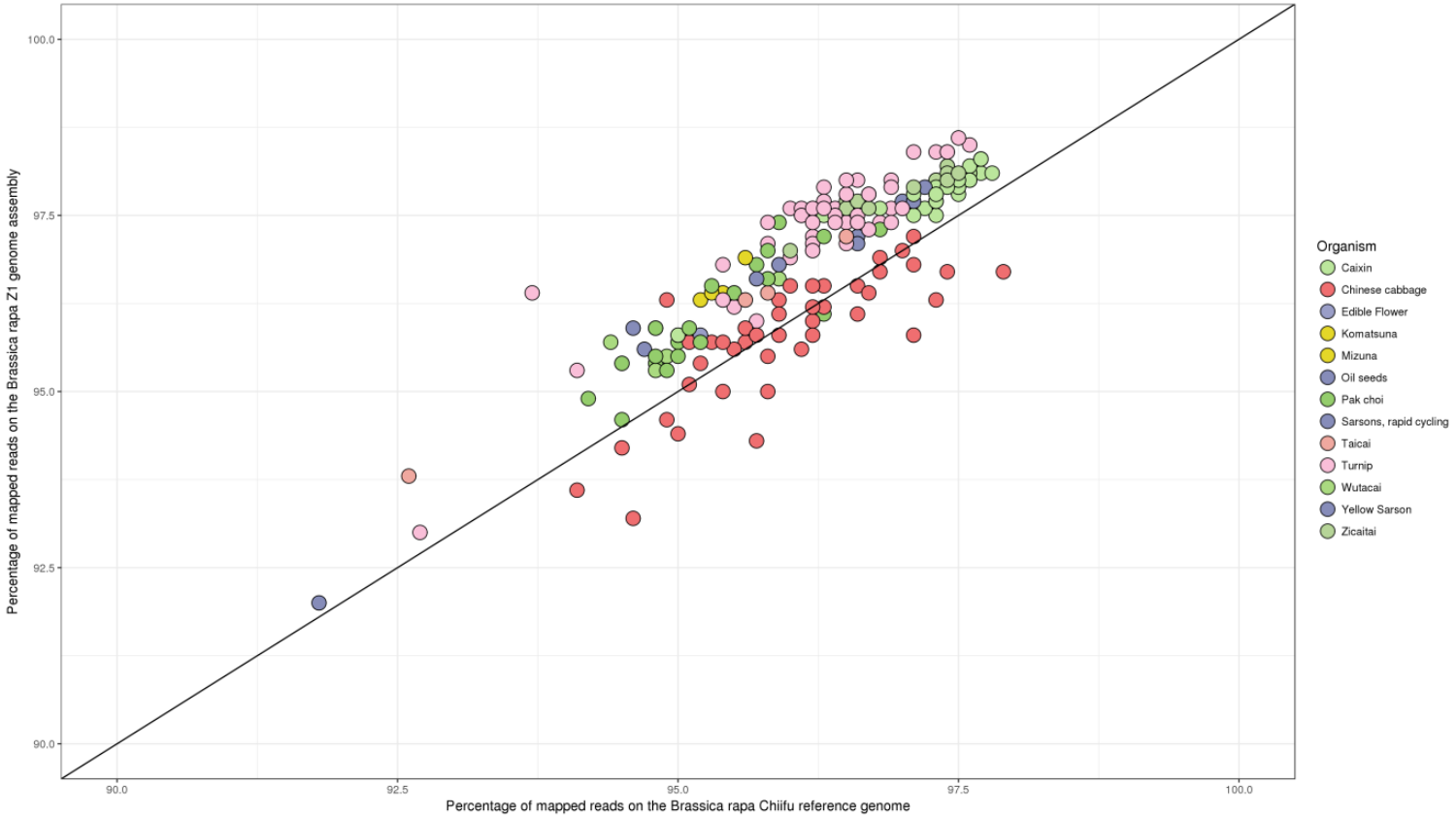
- resequencing data of 199 *B. rapa* and 119 *B. oleracea* accessions.
- representing various morphotypes,
- some closer to the reference genomes Chinese cabbage for *B. rapa* Chiifu and Chinese kale for *B. oleracea* To1000)
- and others closer to our Z1 and HDEM accessions (sarsons for *B. rapa* and broccoli for *B. oleracea*).



Cheng et al. 2016

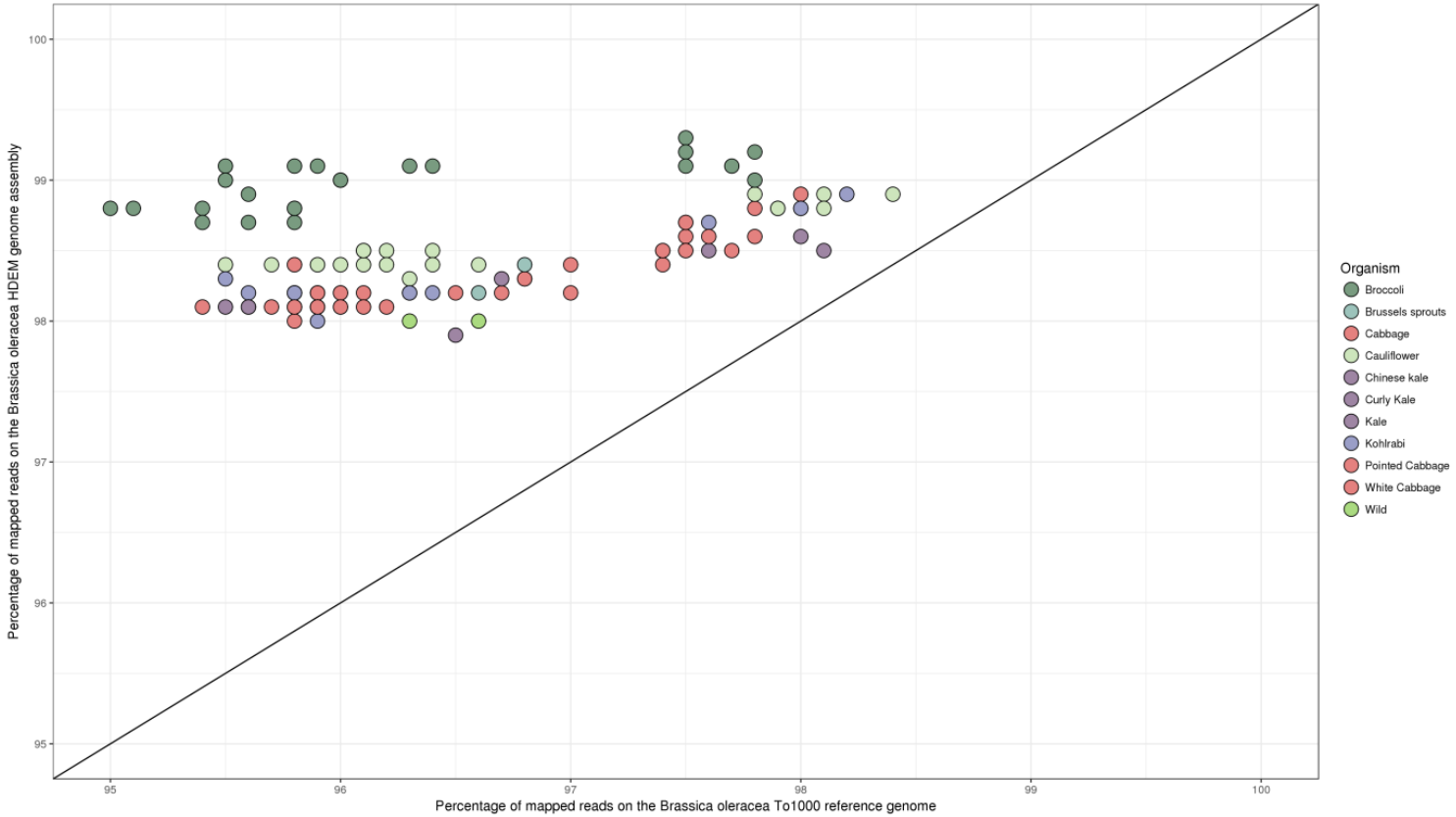
Chromosome-scale assemblies

Comparison of existing references with long-read assemblies



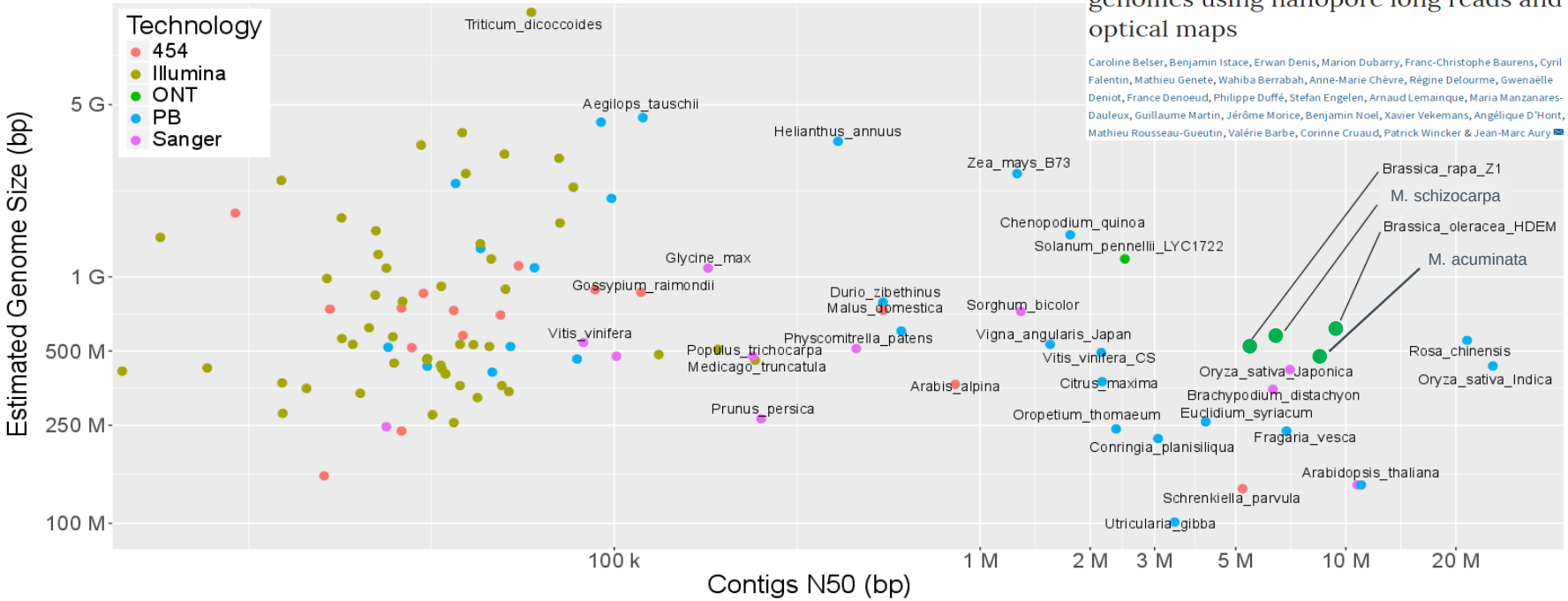
Chromosome-scale assemblies

Comparison of existing references with long-read assemblies



Continuity of current plant genome assemblies

Using Nanopore+Bionano we were able to add four more species with contig N50 > 5Mb



nature plants

Letter | Published: 02 November 2018

Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps

Caroline Belser, Benjamin Istace, Erwan Denis, Marion Dubarry, Franc-Christophe Baurens, Cyril Falentin, Mathieu Genete, Wahiba Berrabah, Anne-Marie Chèvre, Régine Delourme, Gwenaëlle Deniot, France Denoeud, Philippe Duffé, Stefan Engelen, Arnaud Lemainque, Maria Manzanares-Dauleux, Guillaume Martin, Jérôme Morice, Benjamin Noel, Xavier Vekemans, Angélique D'Hont, Mathieu Rousseau-Gueutin, Valérie Barbe, Corinne Cruaud, Patrick Wincker & Jean-Marc Aury

http://www.genoscope.cns.fr/genomes

Sequencing of the banana genome using the PromethION

	Musa schizocarpa	*Musa schizocarpa*
Estimated Genome size	PromethION	MinION
# flowcells	1	18
Cumul. Size	17.6 Gb	27 Gb
N50	26 Kb	24 kb
Coverage	34 X	51 X
# scaffolds	199	227
Cumulative size	519.5 Mb	525.6 Mb
N50	36.8 Mb	36.9 Mb
Contig N50	10.0 Mb	6.5 Mb
Sequencing Costs	~ $6,000	~ $16,000

What's next

- **Recent hybrid synthetized in lab : *Brassica napus* (1.2Gb)**
  - 4 PromethION flowcells : 28 Gb, 27 Gb, 20 Gb and 19 Gb
  - => 85X of long reads ; N50 = 44 Kb ; ~5X of ultra-long reads > 100 Kb

- **Hexaploid wheat genome : *Triticum aestivum* (17Gb)**
  - 2 PromethION flowcells : 75 Gb and 47 Gb
  - => representing 7X (and ~1X from reads >50Kb)

- **100 genomes of *Arabidopsis lyrata*, focus on the dynamic and impact of TE mobilization**

Conclusion

- Already 40 sequenced eukaryotic genomes (200Mb-1500Mb ; plants, brown algae, insects, …) and currently working on optical maps and genome assemblies
- Download assemblies from EBI/NCBI or http://www.genoscope.cns.fr/plants

- Heterozygous genomes/regions are still complicate to manage for actual assemblers

- PromethION throughput allows sequencing of large genomes
- Error rate is acceptable for de novo sequencing projects, but still an issue with homopolymers

- The potential of the device to sequence long reads is impressive
- DNA extraction is a key point (quantity and quality) to obtain “ultra-long” reads and generate optical maps

Acknowledgments

R&DBioSeq Team
www.genoscope.cns.fr/rdbioseq

jmaury@genoscope.cns.fr

@J_M_Aury

- Genoscope labs
  - Bioinformatic : Benjamin Istace, Stefan Engelen, Caroline Belser and Marion Dubarry
  - Sequencing lab: Corinne Cruaud, Erwan Denis, Karine Labadie, Arnaud Lemainque
- Angélique D'Hont & Anne-Marie Chèvre
- Oxford Nanopore Tech Support team
- Funding agencies : CEA, Genoscope and France Génomique