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

Jean-Marc Aury

jmaury@genoscope.cns.fr

@J_M_Aury

MinIOn sequencing at Genoscope

- 6 MinION devices
- >1,000 flowcells ; >50 different organisms; ~1.5 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

(GIGA)n SCIENCE

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

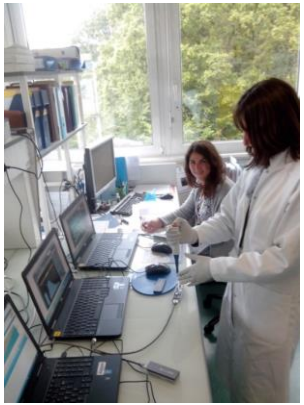
Benjamin Istace, Anne Friedrich, Léo d'Agata, Sébastien Faye, Emilie Payen, Odette Beluche, Claudia Caradec, Sabrina Davidas, Corinne Cruaud, Gianni Liti Arnaud Lemainque, Stefan Engelen, Patrick Wincker, Joseph Schacherer ✉, Jean-Marc Aury ✉

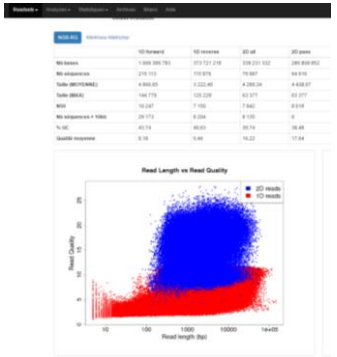
- Software development : error correction tool

BMC Genomics

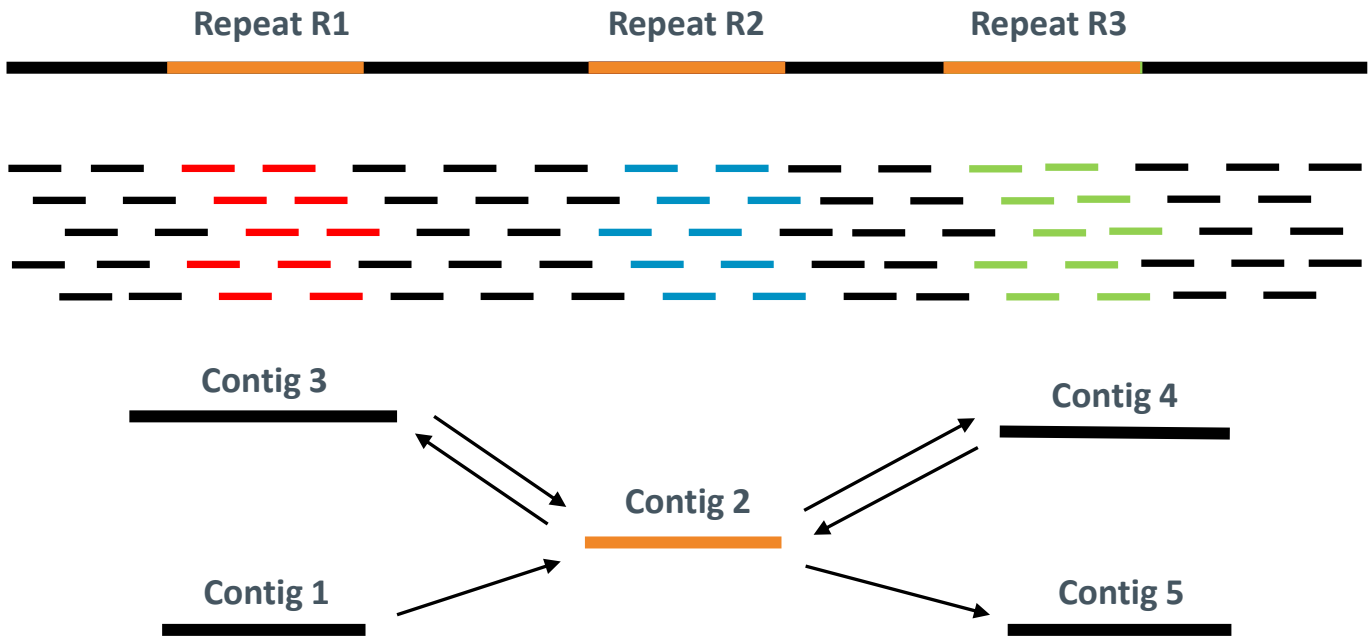
Genome assembly using Nanopore-guided long and error-free DNA reads

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



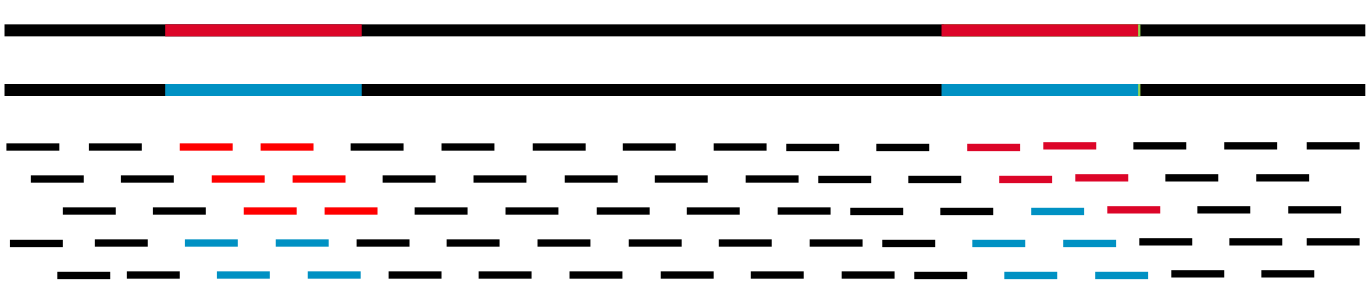


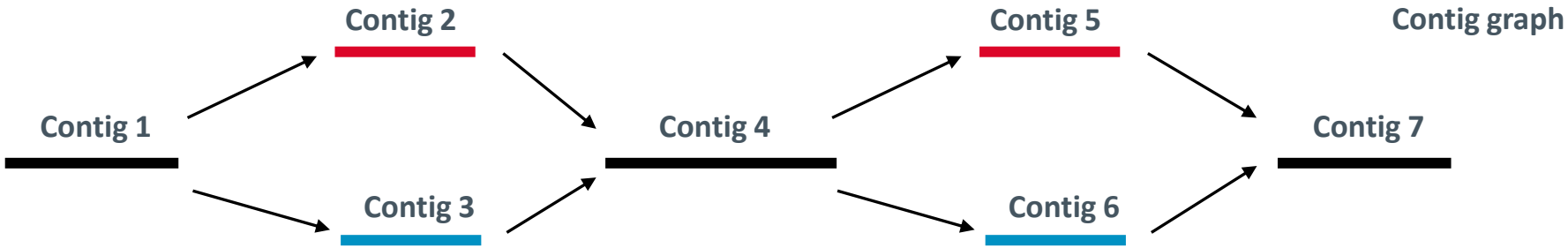
Genome assembly difficulties



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

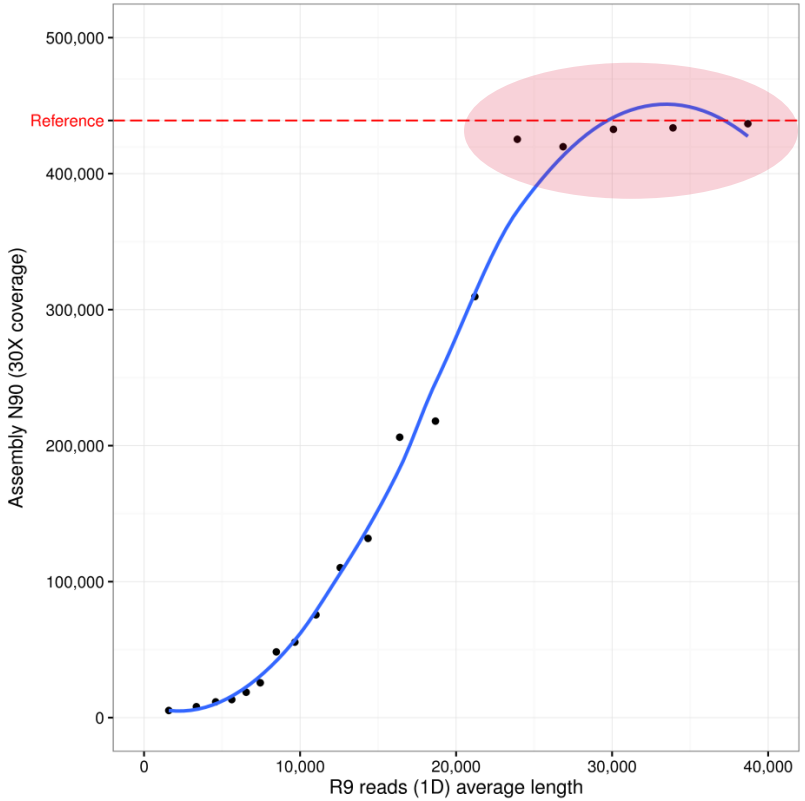
Genome assembly difficulties





=> Heterozygous regions lead to fragmented assemblies and over-estimate the size of the haploid genome

Read Length Matters



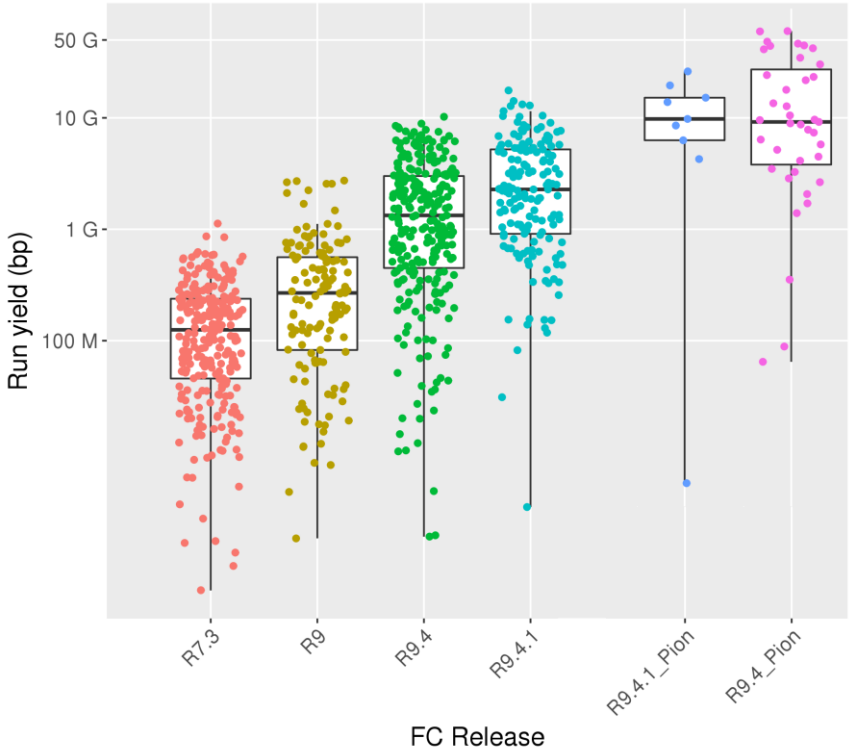
1 contig per chromosome assemblies

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

Nanopore : a fast evolving technology

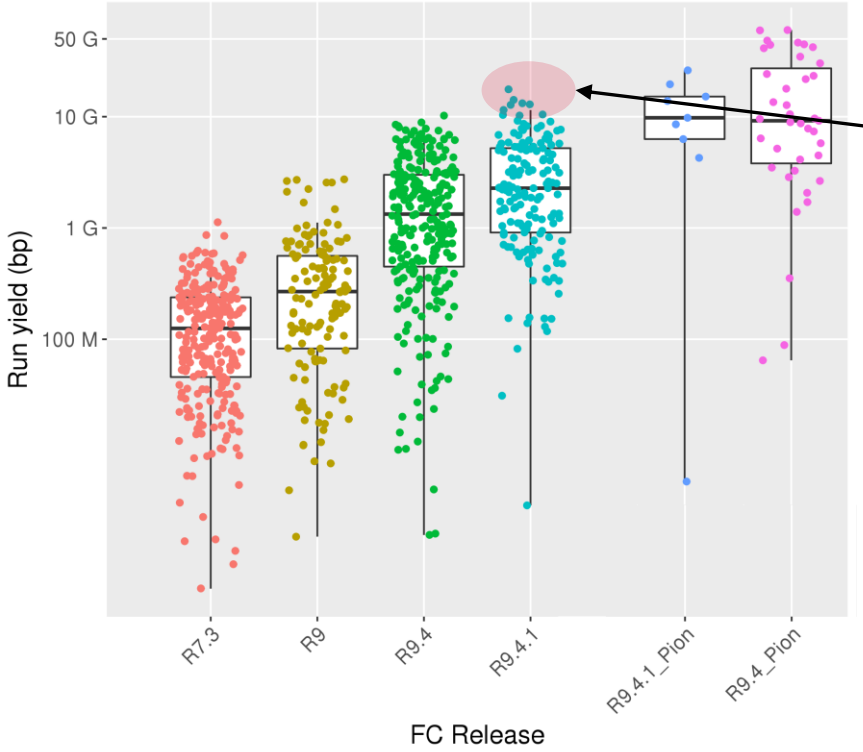
Yield improvement : ~100Mb to several Gb for the MinION and ~10Gb per PromethION flowcell

Throughput is still heterogeneous depending on the DNA sample



Nanopore : a fast evolving technology

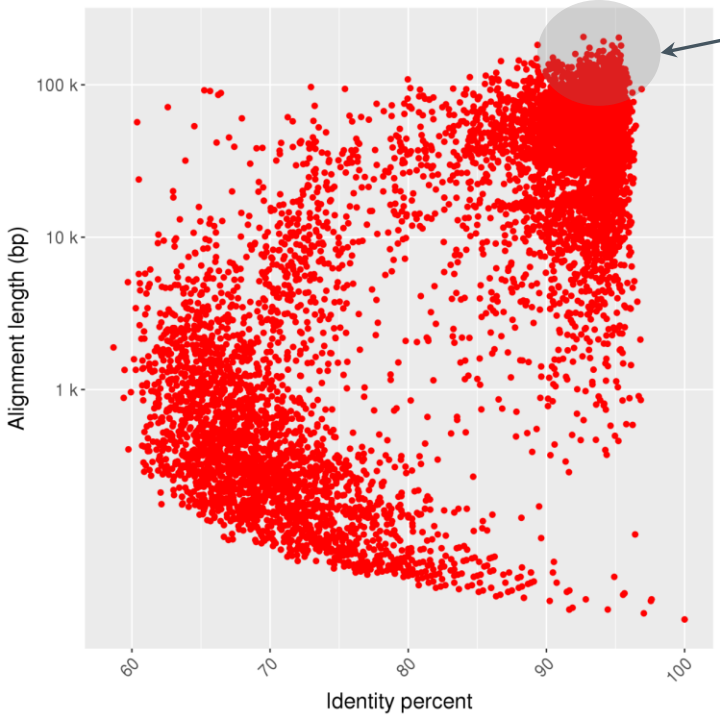
Two best runs were based on lambda samples provided by ONT, 3 runs ~10Gb with our own samples



1. **Lambda : 17.6Gb**
2. **Lambda : 11.5Gb**
3. **Yeast : 10.6Gb**
4. **Yeast : 10.2Gb**
5. **Fish : 10.2Gb**

Nanopore : a fast evolving technology

The device is able to sequence very long DNA fragments (>100Kb)



~400 high quality reads with alignment length > 100Kbp

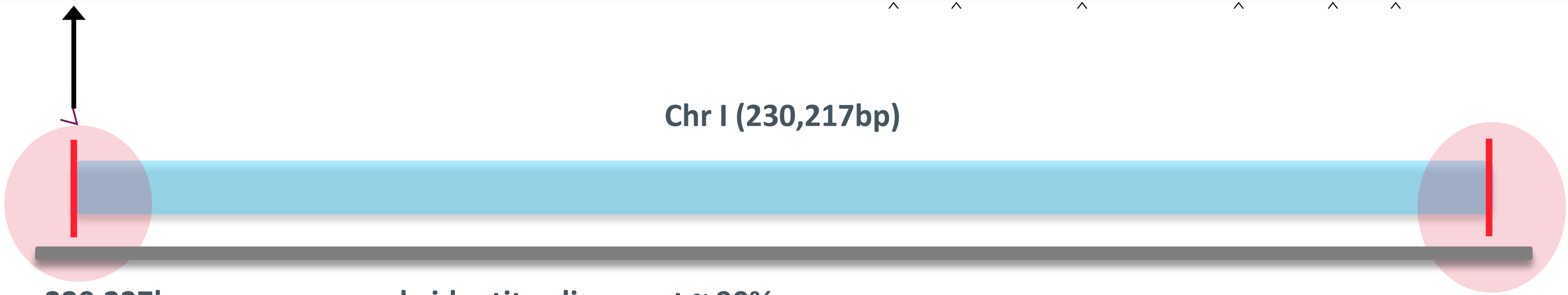
=> ~4X of yeast genome

# bases	2 036 675 349
# sequences	137 109
Max length (bp)	461 529
N50 (bp)	50 800
Nb seq. > 50kb	11 695
Nb seq. > 100kb	3 275

Nanopore : a fast evolving technology

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

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



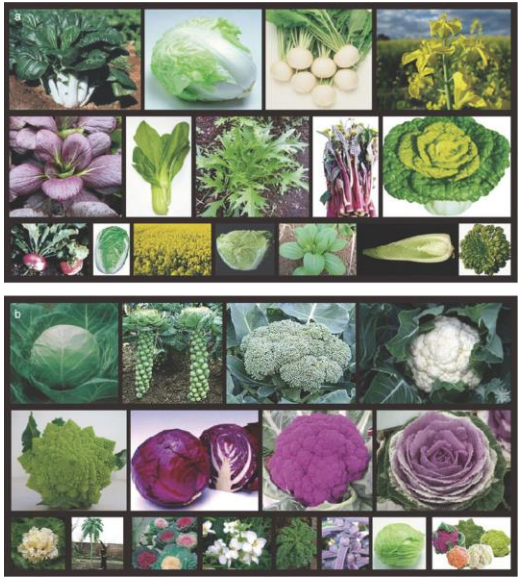
220,227bp nanopore read ; identity alignment ~ 90%

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

The nanopore read is smaller than the chromosome due to deletions

Sequencing of plant genomes using the MinION

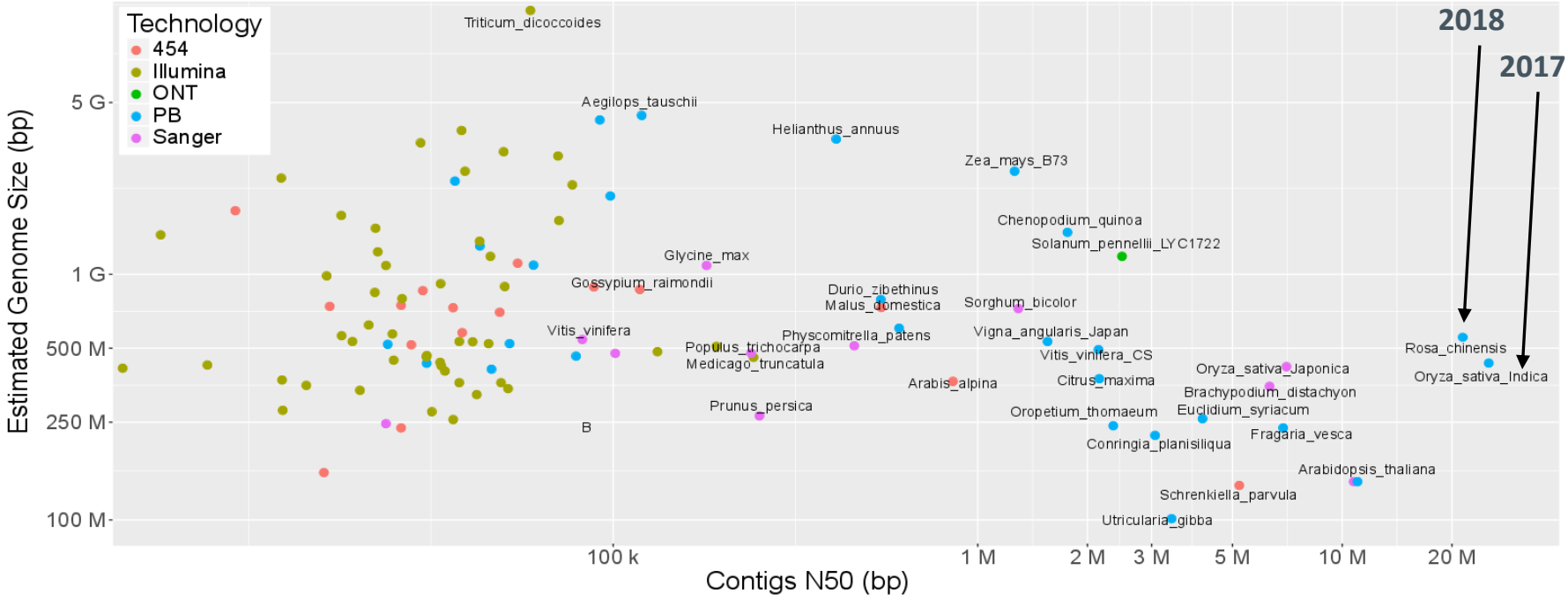
- Large scale genomic projects focused on *Brassica* and *Musa* genomes
- *Brassica* includes important vegetables for human nutrition and are important models for understanding polyploid plants
- The variability between two morphotypes of the same *Brassica* species is high
- *Musa* spp are essential 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



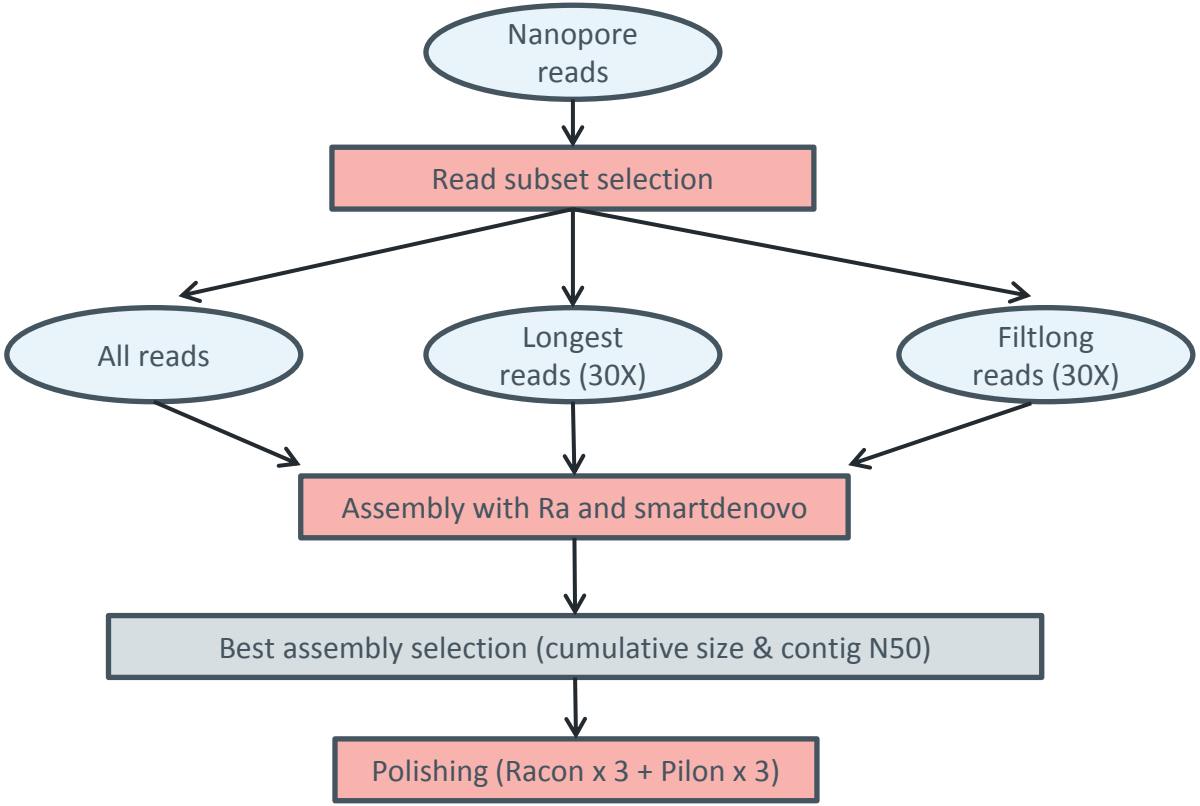
Genome assembly of plant genomes using long and short reads

So far, 2 Brassica and 6 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
# 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

Chromosome-scale assemblies

Organization of nanopore contigs using optical maps

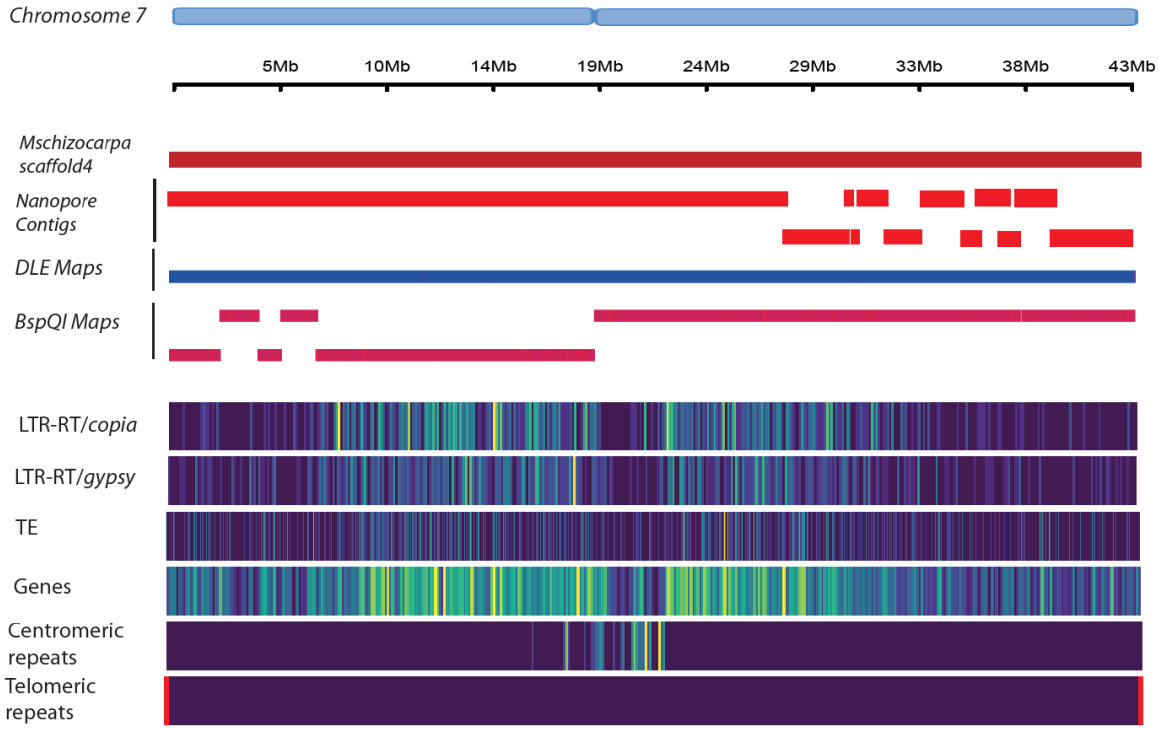
BIONANO GENOMICS

Bionano Direct Label and Stain (DLS) technology

	Brassica rapa *ssp* Z1	***Brassica oleracea*** *ssp* HDEM	***Musa schizocarpa***
# scaffolds	335	140	227
Cumul. Size (N's)	402 Mb (8.2%)	555 Mb (1.8%)	525 Mb (1.5%)
N50	15.4 Mb	29.5 Mb	36.8 Mb
Contig N50	5.5 Mb	9.5 Mb	6.5 Mb
% chromosomes in ≤3 scaffolds	9 /10	8 / 9	11 / 11

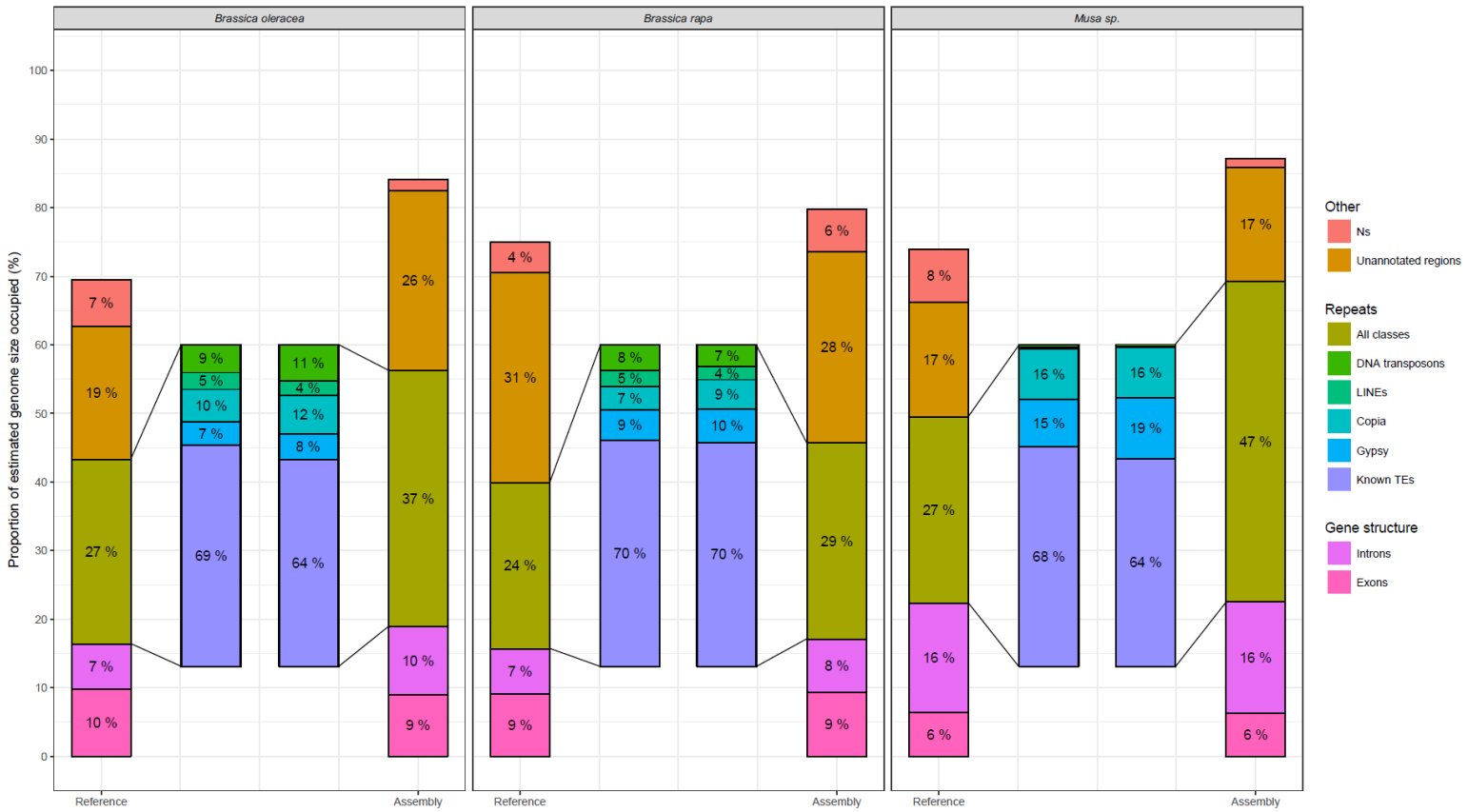
Chromosome-scale assemblies

Schematic view of chromosome 7 from banana genome assembly



Chromosome-scale assemblies

Comparison of existing references with long-read assemblies



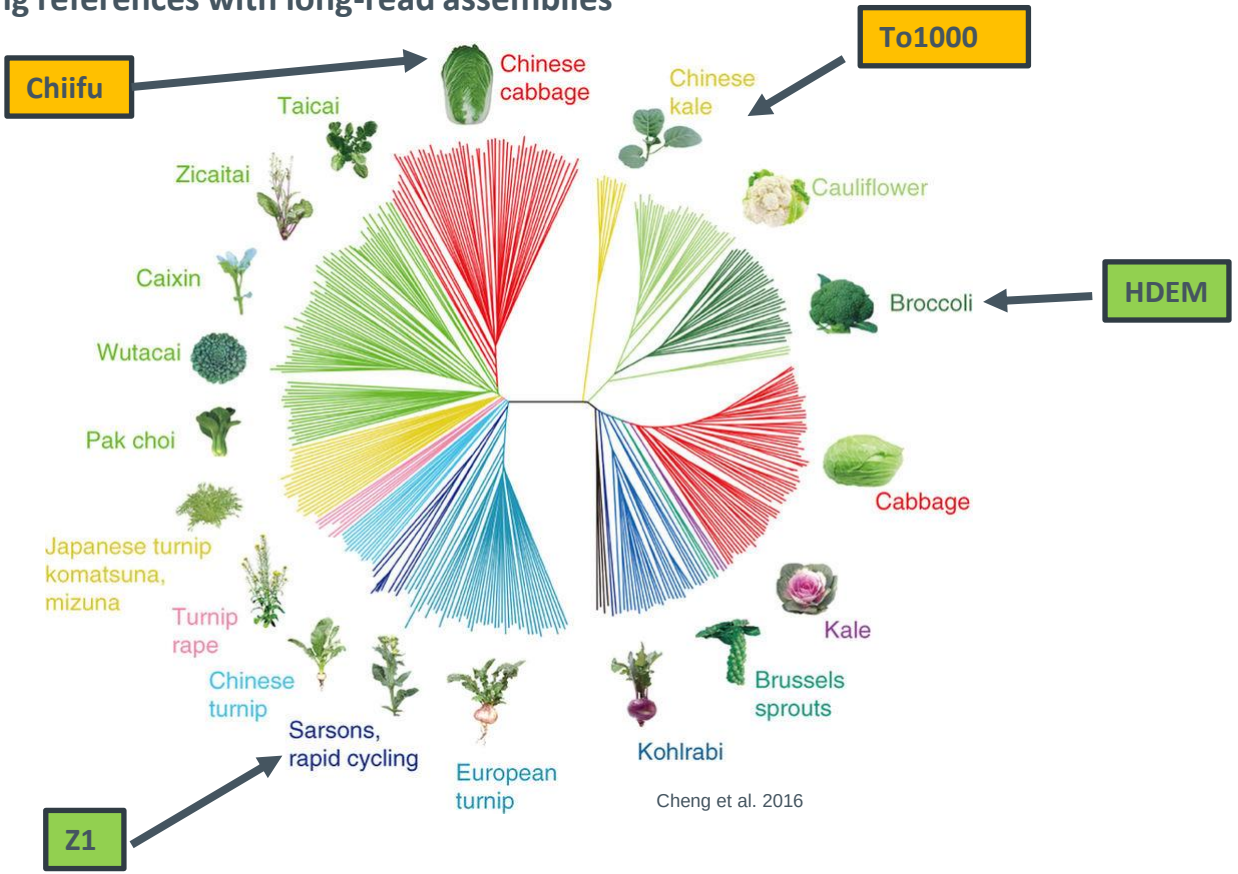
Chromosome-scale assemblies

Comparison of existing references with long-read assemblies

	Brassica oleracea		***Brassica rapa***		***Musa sp.***	
	To1000	HDEM	Chiifu	Z1	*Musa acuminata*	*Musa schizocarpa*
Reference	Liu et al.	This study	Cai et al.	This study	D'hont et al.	This study
Estimated genome size	630	630	529	529	523	587
# chromosomes	9	9	10	10	11	11
Cumulative size	446,885,882	528,860,695	330,820,566	357,074,948	397,008,016	496,921,565
% of anchored bases	91.46%	95.29%	84.52%	88.84%	88.06%	94.60%
Max size	64,984,695	73,711,317	54,546,898	57,670,803	44,889,171	54,858,060
# of N's	39,344,992 (8.8%)	5,972,482 (1.12%)	13,940,645 (4.21%)	27,917,589 (7.81%)	33,488,183 (8.43%)	6,816,353 (1.37%)
Number of genes	59,225	58,874	41,019	45,049	36,542	32,371
% of anchored genes	91.39%	98.22%	96.56%	97.28%	91.98%	98.46%

Chromosome-scale assemblies

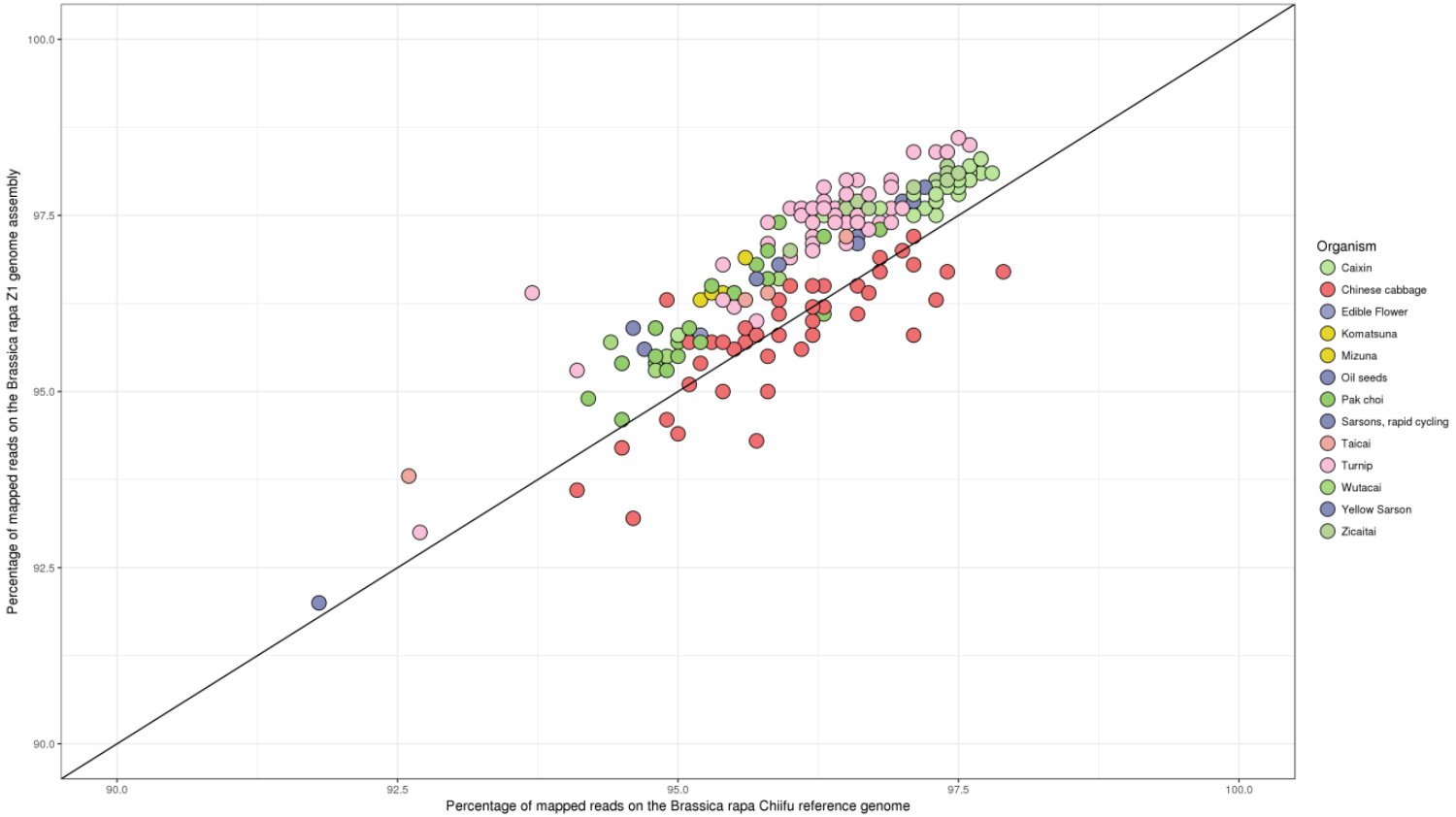
Comparison of existing references with long-read assemblies



Cheng et al. 2016

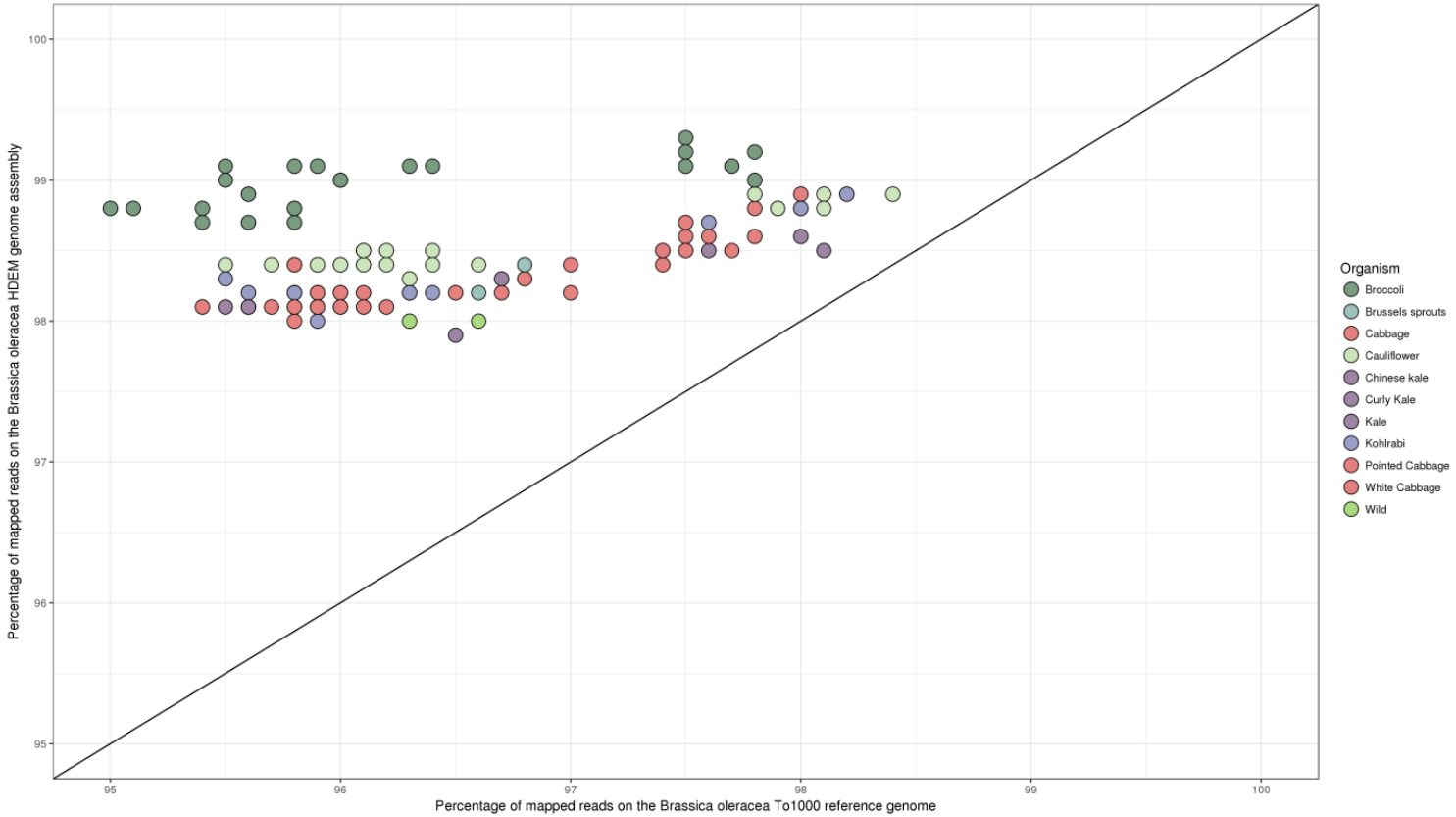
Chromosome-scale assemblies

Comparison of existing references with long-read assemblies



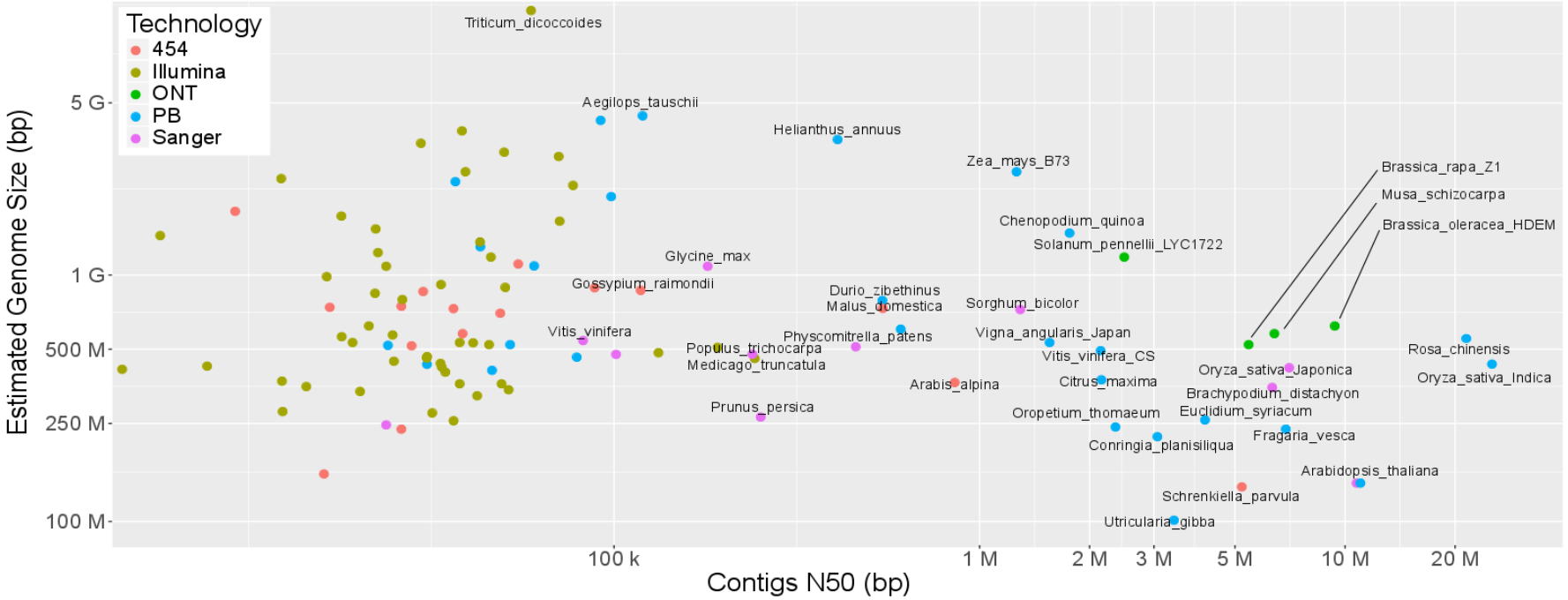
Chromosome-scale assemblies

Comparison of existing references with long-read assemblies



Continuity of current plant genome assemblies

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



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

Conclusion

- Already >40 sequenced eukaryotic genomes (200Mb-1500Mb ; plants, brown algae, insects, …) and currently working on optical maps and genome assemblies
- Heterozygous genomes/regions are still complicate to manage for actual assemblers
- Today error rate is acceptable for de novo sequencing projects, 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

Acknowledgments

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

R&DBioSeq Team

www.genoscope.cns.fr/rdbioseq

jmaury@genoscope.cns.fr

@J_M_Aury

Chromosome-scale assemblies

PacBio and Nanopore sequencing data