nature plants

LETTERS

https://doi.org/10.1038/s41477-018-0289-4

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

Caroline Belser[1,7], Benjamin Istace[1,7], Erwan Denis[1,7], Marion Dubarry[1,7], Franc-Christophe Baurens[2,3], Cyril Falentin[4], Mathieu Genete[5], Wahiba Berrabah[1], Anne-Marie Chèvre[4], Régine Delourme[4], Gwenaëlle Deniot[4], France Denoeud[6], Philippe Duffé[4], Stefan Engelen[1], Arnaud Lemainque[1], Maria Manzanares-Dauleux[4], Guillaume Martin[2,3], Jérôme Morice[4], Benjamin Noel[1], Xavier Vekemans[5], Angélique D'Hont[2,3], Mathieu Rousseau-Gueutin[4], Valérie Barbe[1], Corinne Cruaud[1], Patrick Wincker[6] and Jean-Marc Aury[1]*

Plant genomes are often characterized by a high level of repetitiveness and polyploid nature. Consequently, creating genome assemblies for plant genomes is challenging. The introduction of short-read technologies 10 years ago substantially increased the number of available plant genomes. Generally, these assemblies are incomplete and fragmented, and only a few are at the chromosome scale. Recently, Pacific Biosciences and Oxford Nanopore sequencing technologies were commercialized that can sequence long DNA fragments (kilobases to megabase) and, using efficient algorithms, provide high-quality assemblies in terms of contiguity and completeness of repetitive regions[1–4]. However, even though genome assemblies based on long reads exhibit high contig N50s (>1 Mb), these methods are still insufficient to decipher genome organization at the chromosome level. Here, we describe a strategy based on long reads (MinION or PromethION sequencers) and optical maps (Saphyr system) that can produce chromosome-level assemblies and demonstrate applicability by generating high-quality genome sequences for two new dicotyledon morphotypes, *Brassica rapa* Z1 (yellow sarson) and *Brassica oleracea* HDEM (broccoli), and one new monocotyledon, *Musa schizocarpa* (banana). All three assemblies show contig N50s of >5 Mb and contain scaffolds that represent entire chromosomes or chromosome arms.

The plant genome epic started with the genomes of two model plants: *Arabidopsis*[5] for dicotyledons and rice[6] for monocotyledons in 2000 and 2005, respectively. Their genome sequences, based on the BAC approach and Sanger sequencing, are of high quality and are still today among the best assemblies of plant genomes. With the introduction of Illumina sequencing technology, more than 200 plant genomes have now been sequenced, but most have poor contiguity (Fig. 1) and are composed of thousands of scaffolds. Generally, the gene space is relatively complete and correctly assembled, but regions rich in transposable elements are more fragmented or even under-represented. In addition, the dynamics of transposable elements is largely unknown and this knowledge gap is mostly due to the difficulties in assembling repeated elements from genome sequences obtained using short-read technologies. However, with the development of long-read sequencing technologies, such as Oxford Nanopore Technology (ONT) and Pacific Biosciences (PACBIO), the situation is changing radically and these technologies hold great promise for obtaining high-quality assemblies[1–4]. From 105 plant genomes, we observed that, even with long-read strategies, there is still surprisingly high heterogeneity in terms of assembly contiguity. Even today, only a few plant species have a genome assembly with high contiguity. For example, only six species have an assembly with a contig N50 of > 5 Mb—rice[6,7], *Arabidopsis*[5], woodland strawberry[8], *Schrenkiella*[9], *Brachypodium*[10] and *Rosa*[11]—and they all have a small genome size (see Methods and Supplementary File 2). Here, using our sequencing strategy, we are able to add three more species to this list: *Brassica rapa*, *Brassica oleracea* and *Musa schizocarpa*.

Brassica crops include important vegetables for human nutrition and vegetable oil production. Furthermore, they underwent several paleo-polyploidy events, making their current genomes important models for understanding polyploid plants. The recent sequencing of several hundred *B. rapa* and *B. oleracea* genotypes highlighted the fact that similar morphotypes appeared independently in these two species after a whole-genome triplication event and through parallel selection of paralogous genes[12]. This whole-genome triplication contributed to their diversification into heading and tuber-forming morphotypes[12] (see Methods). It is now accepted that the variability between two morphotypes of the same *Brassica* species is high, showing the importance of having several reference assemblies for a given species. *Musa* spp. include dessert and cooking bananas and are essential staple crops in many tropical and subtropical countries and the most popular fruit in industrialized countries. Cultivars are derived from hybridization between *Musa* species and subspecies, and as such are particularly interesting for studying reticulate evolution. In this context, we decided to sequence two unsequenced morphotypes of *Brassica* (Supplementary Table 1) and the previously unknown genome of *M. schizocarpa*. The genomes of three relatives of these plants

[1]Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France. [2]CIRAD, UMR AGAP, Montpellier, France. [3]AGAP, Université Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France. [4]IGEPP, INRA, Agrocampus Ouest, Université Rennes 1, BP35327, Le Rheu, France. [5]Université Lille, CNRS, UMR 8198—Evo-Eco-Paleo, Lille, France. [6]Génomique Métabolique, Genoscope, Institut de biologie François Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay, Evry, France. [7]These authors contributed equally: Caroline Belser, Benjamin Istace, Erwan Denis, Marion Dubarry. *e-mail: jmaury@genoscope.cns.fr

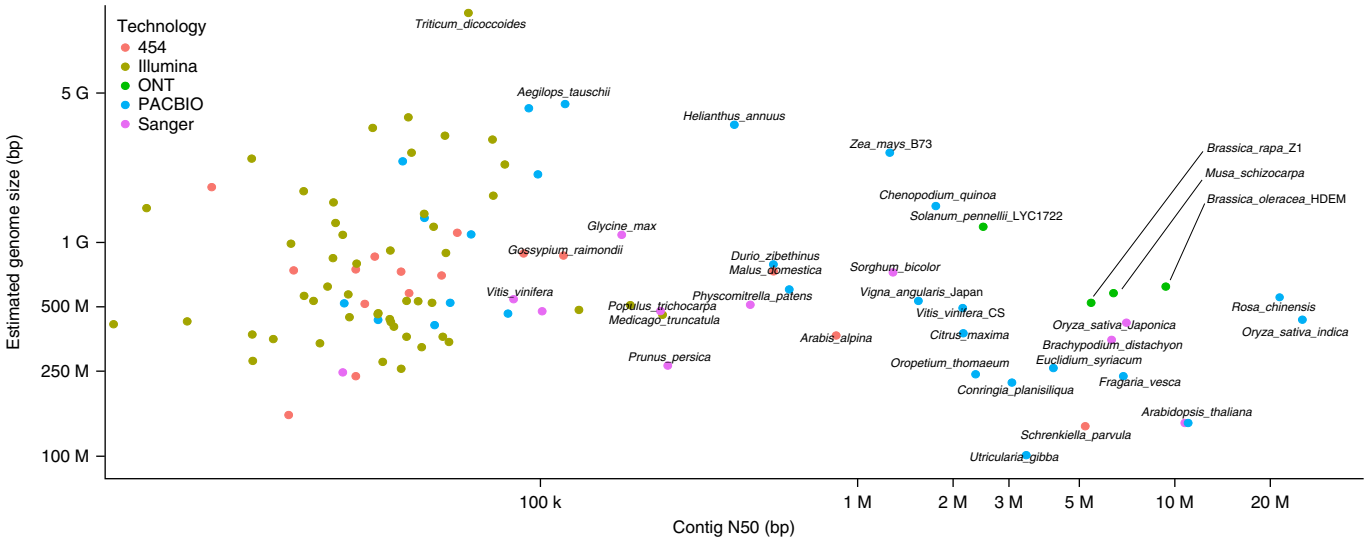

Fig. 1 | Comparison of contig N50 and genome sizes of 105 existing plant genome assemblies. The dots were coloured according to the main sequencing technology used: 454, Illumina, ONT, PACBIO and Sanger.

(*B. rapa* Chiifu[13,14], *B. oleracea* To1000 (ref. [15]) and *Musa acuminata* Pahang-HD[16,17]) have already been sequenced using short-read strategies, but the resulting assemblies are fragmented (contig N50 of <50 kb) and contain a high proportion of unknown or missing bases (22–30%; Supplementary Fig. 1A). For example, the initial sequence of the *B. rapa* genome lacked near half of the expected genome content (273 of 529 Mb). Even though a recent release[13] improved the situation by adding PACBIO long reads, the assembly is still highly fragmented.

Here, we de novo sequenced the genomes of *B. rapa* (Z1 genotype, estimated size of 529 Mb), *B. oleracea* (HDEM genotype, estimated size of 630 Mb) and *M. schizocarpa* (estimated size of 587 Mb) with a strategy combining MinION, which is a portable sequencer commercialized by the Oxford Nanopore company, optical maps produced using the Saphyr system (BioNano Genomics)[18] and short reads from an Illumina sequencer. First, we generated between 38× and 79× nanopore long reads containing a significant proportion of reads longer than 50 kb, representing between 4.4× and 8.2× coverage (Supplementary Table 2). The resulting long-read assemblies showed high contiguity (less than 1,000 contigs with N50s between 3.8 and 7.3 Mb) that facilitated the use of long-range information provided by the optical maps. The final assemblies had N50s between 5.5 and 9.5 Mb at the contig level and N50s between 15.4 and 36.8 Mb at the scaffold level (Table 1). Compared with existing assemblies, the contig N50s of our assemblies are between 100-times and 450-times higher, whereas the scaffold N50s are, in general, lower, but the published assemblies were built using genetic maps (Supplementary Fig. 1B). When adding a genetic map for one of our genomes, *B. oleracea*, we were able to deliver an assembly composed of 129 scaffolds, with the nine chromosomes representing 95.3% of the assembly. Importantly, more than 98% of the markers were in accordance in both the assembly and the genetic map (see Methods). Here, we are able to anchor 528.8 Mb, a significant improvement when compared with the 446.8 Mb of the published release. We decided to use comparative genomics based on available related genomes to produce anchored versions of the *B. rapa* and *M. schizocarpa* genomes. However, although we did not compare these final assemblies with existing references as they were not obtained de novo, we did submit these versions to public repositories as they represent valuable resources for the scientific community (Fig. 2 and Supplementary Table 12).

One-quarter of the chromosomes were composed of a single scaffold and 66% of the chromosomes were assembled into one or two scaffolds, representing either the complete chromosome or a chromosome arm. For example, chromosome 7 of the banana assembly is spanned by a single scaffold that harbours telomeric repeats at both extremities and a high density of centromeric repeats in a 4-Mb region (Supplementary Fig. 2), representing a real improvement compared with the available reference. As observed on this particular chromosome, long-read assembly generated a mix of large and small contigs, proving the importance of combining long reads with long-range information to decipher the chromosome architecture.

We then performed gene prediction on our three genome assemblies using existing annotations of closely related species (see Methods). We annotated 46,721, 61,279 and 32,809 genes for *B. rapa*, *B. oleracea* and *M. schizocarpa*, respectively (Table 1), consistent with the available gene sets and the evolutionary history of these genomes. Transposable elements and, more generally, transposable element-rich regions are under-represented in short-read assemblies, and as expected, in the long-read assemblies, we detected a higher proportion of bases accounting for long interspersed nuclear element (LINE), long terminal repeat retrotransposon and DNA transposon families and the average sizes of the detected transposable elements were higher (see Methods and Supplementary Fig. 10). For example, we predicted 14.95%, 37.95% and 59.95% more complete copies of Copia elements in our assemblies (*B. rapa*, *B. oleracea* and *M. schizocarpa*, respectively) than the reference genomes (Supplementary Table 15). The gene content was mostly the same between the short-read and long-read assemblies, but the long-read assemblies improved the completeness of the transposable element catalogue as well as the genomic context of these transposable element-rich regions (Fig. 3). Generally, genes inserted in transposable element-rich regions are hard to anchor on the chromosomes, and again, our long-read assemblies made it possible to anchor a higher proportion of genes, more than 98% for the three assemblies (Supplementary Table 12).

Read length is a key factor in improving the assembly of transposable element-rich regions and, as a consequence, the assembly contiguity. Recent plant assemblies based on PACBIO sequencing show lower N50s at the contig level (except for the *Rosa chinensis* assembly) owing to the difficulty of sequencing long DNA

Table 1 | Statistics of the genome assemblies

	B. oleracea		*B. rapa*		*Musa* spp.		
	To1000	**HDEM**	**Chiifu**	**Z1**	***M. acuminata***	***M. schizocarpa***	***M. schizocarpa***
Ref.	Parkin et al.[15]	This study MinION	Cai et al.[13]	This study MinION	Martin et al.[17]	This study MinION	This study PromethION
Estimated genome size (Mb)	630	630	529	529	523	587	587
Number of scaffolds (≥2 kb)	1,428	140	86,852	335	24	227	199
Cumulative size	473,834,292	554,975,960	391,410,456	401,923,810	450,848,473	525,280,193	519,202,252
N50 (L50)	48,366,697 (5)	29,516,207 (8)	33,885,992 (5)	15,385,215 (8)	37,593,364 (6)	36,762,082 (6)	36,841,820 (6)
N90 (L90)	39,822,476 (9)	13,883,733 (17)	30,058 (212)	1,671,465 (31)	29,070,452 (11)	9,697,206 (15)	19,788,892 (14)
Maximum size	64,984,695	48,260,371	54,546,898	38,870,275	46,622,217	52,742,985	52,101,276
Number of Ns	42,740,102 (9.02%)	9,958,104 (1.79%)	23,665,136 (6.04%)	32,963,474 (8.2%)	45,326,459 (10.05%)	7,793,997 (1.48%)	7,582,583 (1.46%)
Number of contigs (≥500 bp)	51,566	264	21,717	627	19,265	379	329
Cumulative size	435,858,618	545,017,856	348,573,990	368,960,336	405,516,558	517,486,196	511,619,669
N50 (L50)	22,128 (5,797)	9,491,203 (19)	55,952 (1,902)	5,519,976 (17)	43,237 (2,363)	6,493,909 (24)	9,983,208 (17)
N90 (L90)	4,448 (21,523)	2,202,317 (59)	13,025 (6,630)	184,937 (211)	9,026 (10,326)	1,047,001 (84)	1,020,486 (67)
Maximum size	163,976	26,712,175	327,235	22,127,468	602,020	18,138,554	27,023,771
Number of genes	59,225	61,279	41,019	46,721	36,542	32,809	ND
Number of exons per multi-exon genes (average:median)	5.54:4	5.47:4	6.15:5	5.94:4	6.05:5	6.19:5	ND
BUSCO (complete)	95.1%	95.8%	96.3%	96.6%	86.8%	92.3%	ND

Statistics of HDEM, Z1 and *M. schizocarpa* assemblies and gene content compared with existing reference genomes (To1000, Chiifu and *M. acuminata*). If unspecified, the unit is base pair. ND, not determined.

fragments. We compared the recent *Vigna angularis*[19], *Vitis vinifera*[1], *Citrus maxima*[20], *Arabidopsis thaliana*[21], *Fragaria vesca*[8] and *R. chinensis*[11] PACBIO data with our three ONT data sets and observed a higher proportion of long reads (>50 kb) in the ONT data (Supplementary Fig. 3 and Supplementary Table 17). Moreover, the PACBIO coverage was higher (between 125× and 283×), suggesting the need for higher coverage to obtain a sufficient number of long reads to perform high-contiguity genome assemblies. The nine genomes have estimated genome sizes of 130–630 Mb, and the best assemblies (in terms of contiguity) were the ones produced with the longest reads. Interestingly, the second-most contiguous (contig N50 of 9.5 Mb) assembly was obtained with a long-read data set that had the smallest coverage (~36×) and the longest reads (reads N50 of 31 kb), showing the higher effect of read length over coverage and confirming the possibility of producing high-quality assemblies with 30× long reads coverage[4]. This low-coverage requirement will surely be reduced thanks to the ongoing improvement of the nanopore technology and of protocols for DNA extraction, which still represents a real challenge for numerous plant species.

MinION is a low-cost sequencer, but the current throughput, although sufficient to sequence eukaryotic genomes, is still a limitation to reducing sequencing costs. Our chromosome-scale assemblies cost on average US$14,071 (see Methods) for an average genome size of 582 Mb. However, ONT is currently launching a high-throughput platform named PromethION that promises to lower the cost of sequencing eukaryotic genomes. Here, we sequenced the *M. schizocarpa* genome using a single flow cell that produced 17.6 Gb of data with a comparable read N50 size (26 kb versus 24 kb with MinION; Supplementary Table 2). We assembled this PromethION data set using the same protocol (based on long reads, short reads and two optical maps) and obtained an assembly of the same quality (Table 1). This first attempt using the PromethION device reduced the sequencing cost from US$16,300 to US$6,500 for the banana genome.

To highlight the importance of using these new *Brassica* assemblies as reference genomes, we aligned the resequencing data of 199 *B. rapa* and 119 *B. oleracea* accessions[12] on both the existing references and our genomes. These 318 *Brassica* accessions represent various morphotypes (Supplementary Table 1), with 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*). However, we surprisingly observed that we were able to map a higher proportion of reads on our assemblies (0.61% more for *B. rapa* and 2.77% more for *B. oleracea* on average) for all of the accessions regardless of the morphotype, except for the Chinese cabbage accessions of *B. rapa* (Supplementary Fig. 13 and Supplementary Table 18). Furthermore, as expected, the proportion of uniquely mapped reads was lower on our assemblies, suggesting that repeats were collapsed in the reference genomes (Supplementary Table 19). We expertized the reads that could not be mapped to the reference genomes and detected 1.14 Mb and 1.54 Mb (in HDEM and Z1, respectively) of genic regions that were specific to our accessions or missing from the reference chromosomes (see Methods). These results promote our new genome assemblies as prime references for resequencing analysis.

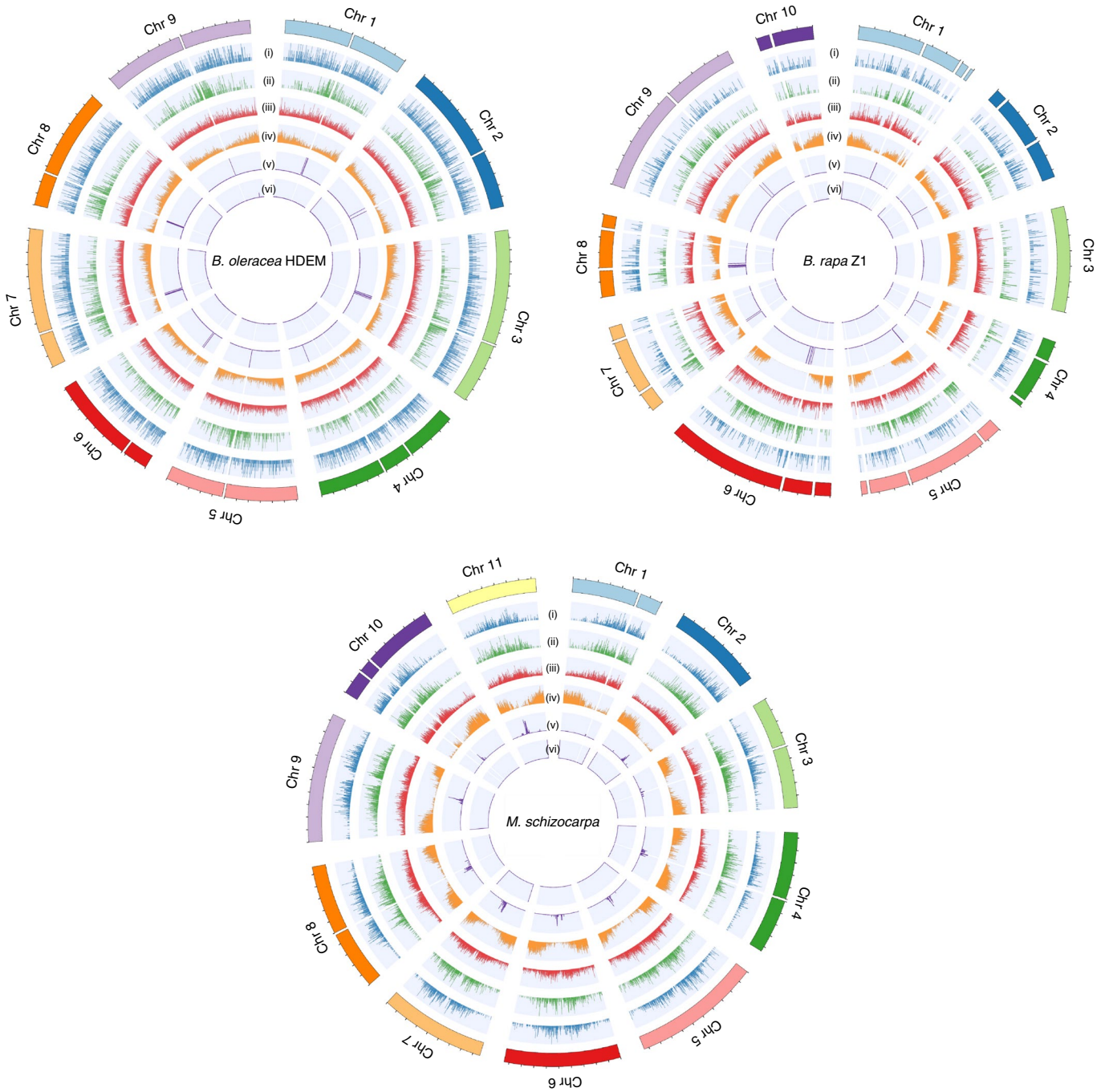

Fig. 2 | Circular representation of anchored scaffolds of *B. oleracea* HDEM, *B. rapa* Z1 and *M. schizocarpa* genome assemblies. Rings represent: density of Copia elements (i), density of Gypsy elements (ii), density of transposable elements (iii), gene density (iv), centromeric repeats positions (v) and telomeric repeats positions (vi). Chr, chromosome.

Even if the *B. oleracea* and *B. rapa* pair of genomes are highly conserved, we detected several differences at the gene level. A comparison of orthologous proteins from Z1–Chiifu and HDEM–To1000 pairs revealed a higher conservation between *B. oleracea* morphotypes (median identity per cent of 99.2% and 98.9%, respectively; Supplementary Fig. 16). Similarly, when looking at the gene-order conservation, we identified more translocation events of gene blocks in the Z1–Chiifu pair (23 against 1; see Methods and Supplementary Figs. 7–9).

We searched for the Flowering Locus C (*FLC*) genes, which are known to be responsible for vernalization and flowering time. Copy number variations of this gene family seem to affect the flowering time. Here, we found, as expected in a broccoli morphotype, the *FLC1*, a partial *FLC2* (the disrupted *FLC2* allele in cauliflower was associated with early flowering) and *FLC3* genes and, interestingly, we annotated an *FLC5* gene (reported as specific to the cauliflower morphotype[22]) and three tandemly duplicated copies of *FLC1*, as suggested in a previous study[23]. In comparison, we found four *FLC* genes in *B. rapa* Z1, as expected (Supplementary Fig. 17). Furthermore, we investigated the *S*-locus, a 30–150-kb region that is strongly enriched in transposable elements[24,25], which causes major difficulties when attempting to assemble it in its entirety using short

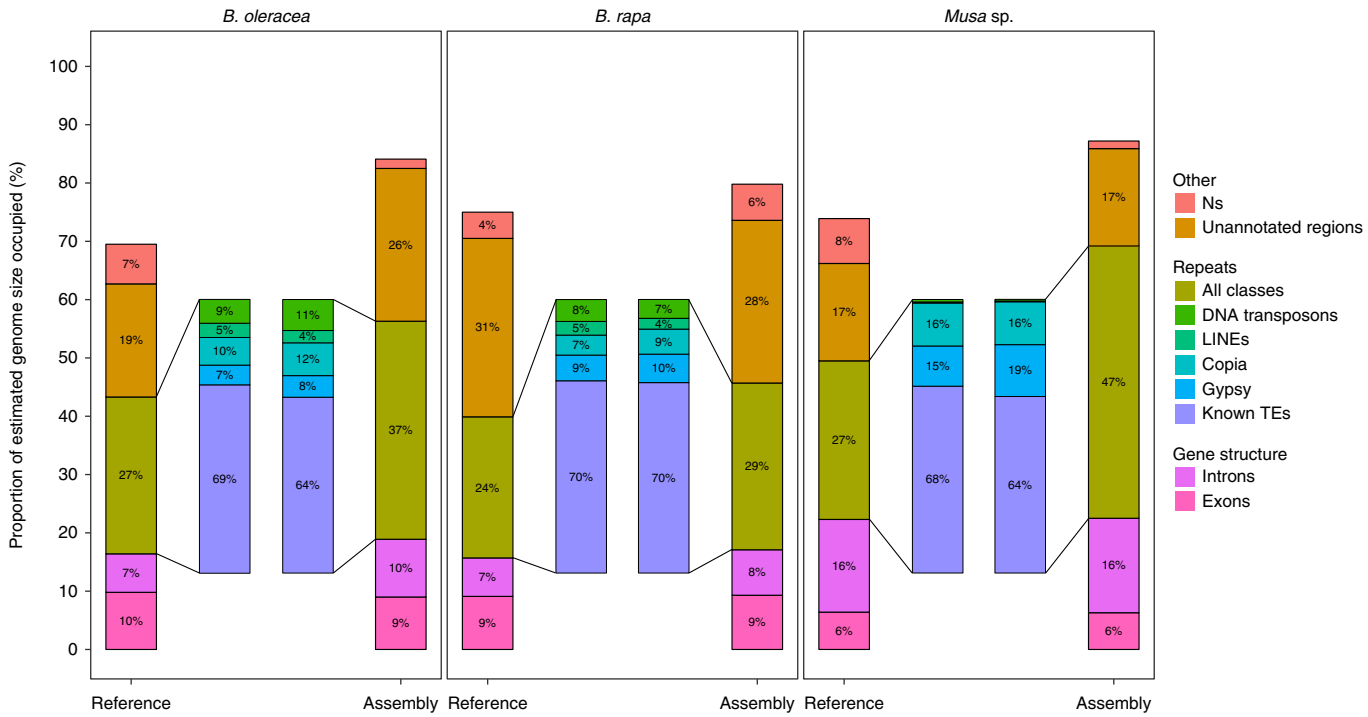

Fig. 3 | Base annotation of the three ONT genomes and the corresponding current references (*B. rapa*, *B. oleracea* and *Musa* species). Genomic regions were classified into several categories: repeats, introns, exons, gaps (Ns) and unannotated regions. Repeated elements were divided into five classes: DNA transposons, LINEs, Copia, Gypsy and other known transposable elements (TEs).

reads[26]. For example, the *S*-locus in the *Raphanus sativus* genome assembly is spread over several contigs[27], whereas that of *Brassica nigra*[28] is spread over 2.2 Mb, suggesting problems with the assembly. We identified a full *S*-locus in *B. rapa* Z1 spanning a 48-kb region syntenic to the *S*-locus region (53 kb, chromosome A07) in the Chiifu assembly. For *B. oleracea* HDEM, the *S*-locus spanned a 102-kb region similar to the *S*-locus region (71 kb, chromosome C06) of To1000 (see Methods and Supplementary Fig. 18).

Whole-genome comparison of *M. schizocarpa* with *M. acuminata* revealed high variability in the centromeric regions (Supplementary Fig. 4). Both versions[16,17] of the *M. acuminata* genome were more fragmented and were anchored using a genetic map. Centromeric regions have a low recombination rate; thus, sequences originating from these regions are always difficult to order and orient correctly, although they represent an essential component of the genomic landscape. This observation highlights the importance of having large contigs to locate centromeres and the richness of the information provided by an optical map compared with a conventional genetic map. Furthermore, we examined disease resistance-like genes (*R*-genes), which are organized in clusters and are generally difficult to assemble correctly. We compared the proportion of undetermined nucleotides for three orthologous *R*-gene clusters on both the *M. acuminata* and the *M. schizocarpa* genomes (see Methods). We found a clear difference between the two genome assemblies (6.5% and 0% of undetermined bases for *M. acuminata* and *M. schizocarpa*, respectively), showing again the importance of long reads for resolving complex regions. A comparison of orthologous proteins showed a high conservation level, both at the base level (median identity per cent of 98.0% between orthologues) and the synteny level (five translocation events, due to assembly errors in the first *M. acuminata* assembly[17]) (see Methods and Supplementary Figs. 7 and 14).

We demonstrated that combining three technologies (ONT, BioNano Genomics and Illumina) can lead to high-quality and relatively low-cost genome assemblies when using the PromethION device (around US$6,000 for 500–600 Mb genomes). We present high-quality genomes for three plant genomes (two Brassicaceae and one banana species) and, when compared with existing reference genomes, our assemblies provide a real improvement, especially in regions enriched in transposable elements. We annotated the three assemblies and observed similar gene content in the gene-rich regions, but a more complete catalogue of transposable elements was produced and the *S*-loci of the two *Brassica* could be entirely annotated thanks to the high quality of the assemblies in these regions. Further improvements are still needed to enable extraction of high-molecular-weight DNA for all plants, and systematic errors in the nanopore long reads mean that Illumina sequences are still required to polish assemblies. Today, optical maps[8] or chromosome conformation[29–32] capture is still mandatory to propose chromosome-scale assemblies for large plant genomes. Even though extraction of high-molecular-weight DNA could remain a challenge, one can imagine that read lengths will increase, enabling the assembly of complete chromosomes with long-read sequencing in coming years.

Methods

***Brassica* morphotypes.** Among the large diversity described for both *B. rapa* and *B. oleracea*, domestication gave rise around 500 years ago to highly contrasted morphotypes for tubers, heads or seeds. Among the six distinct genetic groups identified for *B. rapa*[12], the genotype 'Chiifu' used for the reference genome[14] is a heading type and belongs to the clade that is the most divergent to the last *B. rapa* common ancestor. On the contrary, the genotype sequenced in this study (Z1) is a sarson type (oilseed type) (Supplementary Table 1) and is much closer to the *B. rapa* root. Similarly, the genotype used in this study (HDEM: broccoli) belongs to a different morphotype than the *B. oleracea* 'To1000' reference genome[15] (Chinese kale) (Supplementary Table 1).

Plants. *B. rapa* ssp. *trilocularis* Z1 and *B. oleracea* ssp. *botrytis italica* HDEM seeds were sown in Fertiss blocks (Fertil). Plantlets were grown under a 16-h light/8-h night photoperiod in a greenhouse at 20 °C for 10–12 days. Prior to harvest, the plants were either dark treated for 5 days or not treated after the 10–12-day culture.

M. schizocarpa (ITC926) in vitro plants were obtained from the International Musa Germplasm Collection hold at the International Transit Centre (ITC, Leuven, Belgium). The plants were grown under natural light in a greenhouse at about 25 °C (minimum: 15 °C, maximum: 50 °C), in 1 l pots until they reached a height of 50 cm and had five fully expanded leaves. Harvesting was performed in a period comprised between 8 months and 1 year after the initiation of culture.

DNA extraction. *M. schizocarpa DNA extraction.* High-molecular-weight plant DNA extraction was performed using a modified mixed alkyl trimethyl ammonium bromide (MATAB) procedure[33,34]. A total of 2 g freshly harvested leaves was ground in liquid nitrogen with a mortar and pestle and immediately transferred to 12 ml of 74 °C prewarmed extraction buffer containing 100 mM Tris-HCl, pH 8, 20 mM EDTA, 1.4 M NaCl, 2% w/v MATAB, 1% w/v PEG6000 (polyethylene glycol), 0.5% w/v sodium sulfite and 20 mg l^{-1} RNAse A. Crude extracts were maintained for 20 min at 74 °C, extracted with an equal volume of chloroform-isoamylalcohol (24:1) and transferred to clean tubes. DNA was recovered by centrifugation after adding 10 ml isopropanol. DNA precipitates were briefly dried, washed with 2 ml of 70% ethanol and resuspended in 1 ml sterile water. Extract quality was evaluated using pulse field gel electrophoresis for size estimation and spectrophotometry (A260/A280 and A260/A230 ratios) for purity estimation. DNA samples with a fragment size above 50 kb, a A260/A280 ratio close to 2 and a A260/A230 ratio above 1.5 were kept.

B. rapa and B. oleracea DNA extraction. For some of the nanopore sequencing runs, *B. rapa* Z1 and *B. oleracea* HDEM DNA extracts were prepared according to the protocol used for optical maps (see Optical maps).

For the other sequencing runs, 1 cm^2 first young leaves were harvested and the mid-ribs were removed. The samples were placed on aluminium foil on ice. Then, 2.5 g of each genotype was ground in liquid nitrogen with a mortar and pestle for 1 min. The ground materials were homogenized with 10 ml pre-heated CF lysis buffer (MACHEREY-NAGEL) supplemented with 4 mg proteinase K in 50-ml tubes containing phase-lock gel and incubated for 45 min at 56 °C. Next, 10 ml saturated phenol (25:24:1) was added to the samples and the tubes were placed on a rotator at 40 r.p.m. for 10 min to get a fine emulsion. The samples were centrifuged for 24 min at 4,500*g* (Acc3/Dec3) and the aqueous phases were poured into new 50-ml tubes containing phase-lock gel. Subsequently, 10 ml chloroform-octanol (24:1) was added to each sample and the tubes were placed on a rotator at 40 r.p.m. for 10 min to get a fine emulsion. The samples were centrifuged for 24 min at 4,500*g* (Acc3/Dec3) and the aqueous phases were poured into new 50-ml tubes. The DNA was precipitated by adding 4 ml of 5 M NaCl and 30 ml of cold 100% isopropanol. After 3 h at 4 °C, the DNA was removed in one piece with a hook produced by melting a glass capillary in a blue flame. The DNA pellets were submerged in 50-ml tubes containing 70% ethanol and transferred to new tubes to evaporate the remaining isopropanol at 37 °C (in an oven). The dried DNA was resuspended with 3 ml transposable element 10/1 buffer. The extract quality was evaluated using field inverted gel electrophoresis with the Pippin pulse system (Sage Sciences). DNA samples with a fragment size above 50 kb were kept and run on BluePippin (Sage Sciences).

Illumina sequencing. DNA (1.5 µg) was sonicated to a 100–1,500-bp size range using a Covaris E220 sonicator (Covaris). The fragments were end repaired and 3′-adenylated, and Illumina adapters were added using the Kapa Hyper Prep Kit (Kapa Biosystems). The ligation products were purified with AMPure XP beads (Beckmann Coulter Genomics). The libraries were quantified by quantitative PCR using the KAPA Library Quantification Kit for Illumina Libraries (Kapa Biosystems), and the library profiles were assessed using a DNA High Sensitivity LabChip kit on an Agilent Bioanalyzer (Agilent Technologies). The libraries were sequenced on an Illumina HiSeq2500 instrument (Illumina) using 250 base length read chemistry in paired-end mode.

Nanopore sequencing. *MinION.* Most of the libraries were prepared according to the following protocol, using the Oxford Nanopore SQK-LSK108 kit. Genomic DNA or DNA previously fragmented to 50 kb with a Megaruptor (Diagenode S.A.) was first size selected using BluePippin (Sage Sciences). The selected DNA fragments were end repaired and 3′-adenylated with the NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs). The DNA was then purified with AMPure XP beads (Beckmann Coulter) and ligated with sequencing adapters provided by ONT using the Blunt/TA Ligase Master Mix (NEB). After purification with AMPure XP beads, the library was mixed with running buffer with fuel mix (ONT) and library loading beads (ONT) and loaded on MinION R9.4 or R9.5 SpotON Flow Cells.

PromethION. Libraries were prepared following the Oxford Nanopore '1D Genomic DNA by ligation (Kit 9 chemistry)—PromethION' protocol. Genomic DNA was first repaired and end prepped with NEBNext FFPE Repair Mix (New England Biolabs) and the NEBNext Ultra II End Repair/dA-Tailing Module (NEB). The DNA was then purified with AMPure XP beads (Beckmann Coulter) and ligated with sequencing adapters provided by ONT using concentrated T4 DNA ligase 2 M U ml^{-1} (NEB). After purification with AMPure XP beads (Beckman Coulter) using dilution buffer (ONT) and wash buffer (ONT), the library was mixed with sequencing buffer (ONT) and library loading beads (ONT) and loaded on the PromethION flow cells.

Optical maps. For *B. rapa* Z1 and *B. oleracea* HDEM, 1 cm^2 first young leaves were harvested and the mid-ribs were removed. The samples were placed on aluminium foil on ice. Then, 5 g of each genotype was ground in liquid nitrogen with a mortar and pestle for 2 min. The ground materials were homogenized in 50 ml NIBTM (10 mM Tris-HCl, pH 8, 10 mM EDTA, pH 8, 80 mM KCl, 0.5 M sucrose, 1 mM spermine tetrahydrochloride, 1 mM spermidine trihydrochloride and 2% (w/v) PVP40), the pH was adjusted to 9.4 and the solution was filtered through a 0.22-µm filter (NIB) and supplemented with 0.5% Triton X-100 (NIBT) and 7.5% 2-mercaptoethanol (NIBTM). The nuclei suspensions were filtered through cheese cloth and Mira cloth and centrifuged at 1,500*g* for 20 min at 4 °C. The pellets were suspended in 1 ml NIBTM and adjusted to 20 ml with NIBTM. The nuclei suspensions were filtered again through cheese cloth and Mira cloth and centrifuged at 57*g* for 2 min at 4 °C. The supernatants were kept and centrifuged at 1,500*g* for 20 min at 4 °C. The pellets were suspended in 1 ml NIBT and adjusted to 20 ml with NIBT. To wash the pellets, the last steps were repeated three times with 50 ml NIBT and a final time with 50 ml NIB. The pellets were suspended in residual NIB (approximately 200 µl), transferred to a 1.5-ml tube and centrifuged at 1,500*g* for 2 min at 4 °C. The nuclei were suspended in cell suspension buffer from CHEF Genomic DNA Plug Kits (Bio-Rad), and melted 2% agarose from the same kit was added to reach a 0.75% agarose plug concentration. Plug lysis and DNA retrieval were performed as recommended by Bionano Genomics.

For *M. schizocarpa*, a young cigar leaf was harvested and the mid-rib was removed. Only the yellow part was used. Then, 2 g of the leaf segment was cut to 2 × 2 cm and fixed in 2% formaldehyde according to the Bionano Genomics protocol, except the fixation step was performed in a vacuum bell. After the 100-µm and 40-µm filter steps, the nuclei suspension was centrifuged at 60*g* for 2 min at 4 °C. The supernatant was filtered again through a 40-µm filter. The last centrifugation was repeated and the supernatant was filtered again through a 40-µm filter. The nuclei suspension was centrifuged at 400*g* for 15 min at 4 °C. The pellet was suspended in residual buffer and adjusted to 35 ml. The nuclei suspension was centrifuged once more at 400*g* for 15 min at 4 °C. The pellet was suspended in residual buffer and adjusted to 35 ml. The nuclei suspension was centrifuged at 200*g* and the supernatant was kept and centrifuged at 400*g* for 10 min at 4 °C. The nuclei were suspended in homogenisation buffer plus (Bionano Genomics), and melted 2% agarose from CHEF Genomic DNA Plug Kits (Bio-Rad) was added to reach a 0.82% agarose plug concentration. Plug lysis was performed with Bionano lysis buffer adjusted to pH 9 and supplemented with 0.4% 2-mercaptoethanol. DNA retrieval was performed as recommended by Bionano Genomics.

The Nicks, Labels, Repairs and Stains labelling (BspQI) protocols were performed for *B. rapa* Z1, *B. oleracea* HDEM and *M. schizocarpa* according to Bionano Genomics with 600, 300 and 191.2 ng DNA, respectively. For the direct label and stain (DLS) labelling (DLE-1), *M. schizocarpa* DNA was concentrated by evaporation at room temperature. All DLS labelling was performed with 750 ng DNA. Chip loading was performed as recommended by Bionano Genomics.

Quality control of raw reads. *Illumina data.* After Illumina sequencing, an in-house quality control process was applied to the reads that passed the Illumina quality filters. The first step discarded low-quality nucleotides ($Q < 20$) from both ends of the reads. Next, Illumina sequencing adapters and primer sequences were removed from the reads. Then, reads shorter than 30 nucleotides after trimming were discarded. These trimming and removal steps were achieved using in-house-designed software based on the FastX package[35]. The last step identified and discarded read pairs that mapped to the phage phiX genome, using SOAP[36] and the phiX reference sequence (GenBank: NC_001422.1). This processing resulted in high-quality data and improvement of subsequent analyses.

Nanopore data. The nanopore long reads were not cleaned; we used the raw reads for each genome assembly. Taxonomic assignation was performed using Centrifuge[37] for each data set to detect potential contamination.

Long-read genome assemblies. We used the Ra[38], SMARTdenovo[39] and wtdbg assemblers with all nanopore raw (or corrected) reads or subsets of raw (or corrected) reads composed of either the longest reads or those selected by the Filtlong[40] software (Supplementary Tables 3–5), as it has been proven that downsampling the read coverage can be beneficial for the assembly phase[4]. We also tried to use Canu[41] but could not get to the final assembly stage owing to the high computational requirements. Moreover, we could not select any subsets of reads for *B. oleracea*, as the sequencing depth was too low to subsample the sequencing data. Ra, wtdbg and Filtlong were used with default parameters. We used the following options as inputs to SMARTdenovo: '-c 1' to generate a consensus sequence, '-J 5000' to remove sequences smaller than 5 kb and '-k 17' to use 17-mers, as this is advised by the developers for large genomes.

Then, we selected the 'best' assembly for each organism, based on contiguity metrics, such as N50 or cumulative size. For all organisms, the

Ra assembler produced the most contiguous assembly. The best *B. oleracea* and *B. rapa* assemblies were obtained using all of the reads and had contig N50s of 7.3 Mb and 3.8 Mb, respectively. By contrast, the best *M. schizocarpa* assembly (N50 of 4.0 Mb) was obtained using a 30× subset of reads generated by Filtlong.

A high-quality consensus was needed for both aligning the optical map onto the contigs and annotating genes. As nanopore reads contain systematic errors in homopolymeric regions, we polished the consensus of the selected assembly three times with the nanopore reads as input to the Racon[42] software and then three additional times using Illumina reads as input to the Pilon[43] tool (Supplementary Tables 6–8). Both tools were used with default parameters. The polishing process significantly improved the number of complete BUSCOs detected in all organisms. The percentage of complete BUSCOs went from 74.2% to 97.3% for *B. oleracea*, from 79.7% to 97.8% for *B. rapa* and from 53.8% to 93.4% for *M. schizocarpa*.

Long-range genome assemblies. Two enzymes (BspQI and DLE-1) were used to generate optical maps and both maps were produced using a single chip for each genome. The DLE-1 map was generated using the new DLS technology, which significantly improved the contiguity of the optical maps (N50s are 6–15-times higher using DLS; Supplementary Table 10). Genome map assemblies for the three species were generated using Bionano Solve Pipeline version 3.1.1 and Bionano Access version 1.0a. A rough assembly was first performed with the following parameters: -i 0 -V 0 -A -z -u -m (pipelineCL.py). This first result was used as a reference for a second assembly, launched with the following parameters (as recommended by the supplier): -y -r (rough assembly cmap) -V 0 -m. We filtered out molecules smaller than 180 kb and molecules with less than nine labelling sites (Supplementary Tables 9 and 10). The nanopore contigs were then organized using the two Bionano maps (DLE and BspQI), with the scaffolding procedure provided by BioNano Genomics, and negative gap sizes were checked with an internal procedure that fused overlapping contigs and greatly improved the contig size (see the 'Resolution of negative gaps' section).

Construction of a high-density *B. oleracea* genetic map and validation and anchoring of our *B. oleracea* assembly. To construct a *B. oleracea* genetic map, an F2 population (95 progenies) was obtained from a cross between Richelain (*B. oleracea* ssp. *capitata*) and HDEM (*B. oleracea* var. *botrytis italica*). This population was genotyped using the Illumina 60 K array and a genetic map was constructed using the CarthaGène software[44]. A total of 6,528 markers were genetically mapped, totalling 817.3 cm. The sequence contexts of all single-nucleotide polymorphism markers that were genetically mapped were blasted against our *B. oleracea* HDEM assembly to validate the quality of our assembly and to help with the ordering and orientation of scaffolds. Of these 6,528 markers, 5,449 were physically anchored on the HDEM assembly and, more specifically, onto the 20 largest scaffolds (out of 140), representing 96.96% of the whole assembly. The genetic and physical positions were discordant for only 95 markers (1.74%) due to an inaccurate position on the genetic map (of a few centimorgans in almost all cases). In most cases, only two scaffolds per pseudomolecule were obtained, with one end of each scaffold corresponding to a centromere region (Supplementary Fig. 12).

Resolution of negative gaps. We inspected several regions of the optical map where two nanopore contigs were joined and found in several cases that the nanopore contigs overlapped (based on the optical map), and this overlap was not managed by the hybrid scaffolding procedure (Supplementary Fig. 11). In these cases, the workflow decided not to fuse the two contigs and added a 499-bp gap. We checked all 499-bp gaps and aligned both 30-kb flanking regions with BLAT[45]. The two flanking contigs were joined if one alignment (score of >3,000) was detected. This procedure resolved several negative gaps and improved the contig N50 (Supplementary Table 11).

Transposable element annotation. Transposable elements and, more generally, transposable element-rich regions are under-represented in short-read assemblies. To investigate this aspect, we performed transposable element detection for our three genomes and the three available reference assemblies. Transposable elements were annotated using RepeatMasker[46] (with default parameters, taxon viridiplantae for *M. schizocarpa* and eudicotyledons for *Brassica*) and transposable element libraries. The transposable element database generated in ref. [47] was used to annotate the *B. oleracea* and *B. rapa* genomes. The transposable element database for the *M. schizocarpa* annotation came from an *M. acuminata* study[17]. We masked 34.43%, 37.82% and 51.09% of the genomes of *B. rapa, B. oleracea* and *M. schizocarpa*, respectively (Supplementary Tables 15 and 16).

Gene prediction. Gene prediction was done using proteomes from homologous species. For *B. rapa*, we used the following three proteomes: *B. rapa* (UP000011750), *Brassica napus* (UP000028999) and *A. thaliana* (UP000006548). For *B. oleracea*, we used the proteomes of *B. oleracea* (UP000032141), *B. napus* and *A. thaliana*. For *M. schizocarpa*, we used the proteomes of *M. acuminate* (UP000012960), *Oryza sativa* (UP000059680) and *Phoenix dactylifera* (UP000228380).

Low complexity in protein sequences was masked with the SEG algorithm. Low complexity in genomic sequences was masked using the DustMasker algorithm[48]. Tandem repeats were masked using Tandem Repeat Finder[49]. The transposable elements detected by RepeatMasker were also masked for the gene prediction step, as described in the 'Transposable element annotation' section.

The proteomes were aligned to the genomes in two steps. First, BLAT[45] (default parameters) was used to quickly localize corresponding putative genes of the proteins on the genome. The best match and matches with a score of ≥90% (70% for *A. thaliana* proteins) of the best-match score were retained. Second, the alignments were refined using Genewise[50] (default parameters), which is more precise for intron–exon boundary detection. Alignments were kept if more than 80% of the length of the protein was aligned to the genome.

We integrated the protein homologies using a combiner called Gmove[51] to predict gene structures. This tool can find coding sequences based on the protein mapping structures. It is easy to use with no need for a pre-calibration step. Briefly, putative exons and introns, extracted from the alignments, were used to build a simplified graph by removing redundancies. Then, Gmove extracted all paths from the graph and searched for open reading frames consistent with the protein evidence. A selection step was applied to all candidate genes, essentially based on gene structure.

Finally, we used the pan-genomes of *B. oleracea*[22] and *B. napus* to complete the gene catalogue. We aligned these two protein databases using the same workflow and integrated the results using Gmove. The final gene catalogues of *B. oleracea* and *B. rapa* are composed of the first Gmove results with the predicted genes (based on pan-genomes) that do not overlap with any previous annotation.

Using this pipeline, 46,721, 61,279 and 32,809 gene models were predicted for *B. rapa*, *B. oleracea* and *M. schizocarpa*, respectively. We assessed the completeness of the annotation using BUSCO[52] (embryophyta data set) and detected a similar (or higher) proportion of complete genes when compared with existing gene annotations (Table 1 and Supplementary Table 13). Moreover, we computed the annotation evaluation distance for the three genomes using the existing gene catalogues as reference annotation (Supplementary Table 14).

Comparison with available plant assemblies. We downloaded a selection of 102 plant genome assemblies by retrieving assemblies organized at the chromosome level and recently published genomes. We computed the usual metrics (cumulative size, NX, LX, average size, number and size of gaps) at the scaffold level. Where NX is defined as the sequence length of the shortest sequence at X% of the total genome length, and LX is defined as the smallest number of contigs whose length sum produces NX. Contigs were generated by fragmenting scaffolds at each N and metrics were computed from the resulting contig fasta files. Expected genome size and chromosome number information was obtained from the Kew website (https://www.kew.org) or from scientific publications. This information is available in Supplementary File 2.

Comparison of ONT and PACBIO data sets. Six PACBIO data sets were downloaded from the European Bioinformatics Institute-European Nucleotide Archive and metrics were computed from the whole data set with the following coverage: 127× (*V. angularis*, PRJDB3778), 231× (*V. vinifera*, PRJNA316730), 224× (*R. chinensis*, PRJNA413292), 125× (*C. maxima*, PRJNA318855), 168× (*F. vesca*, PRJNA383733) and 283× (*A. thaliana*, PRJNA237120). Likewise, metrics were computed from the whole ONT data sets: 79× (*B. rapa*), 51× (*M. schizocarpa*) and 36× (*B. oleracea*). All of the standard metrics are available in Supplementary Table 17.

Comparison with reference genomes. We compared the three genome assemblies with corresponding reference genomes using nucmer from the mummer4 package[53] and dot[54], an interactive dot plot viewer, to generate genome–genome alignment dot plots (Supplementary Figs. 4–6).

We downloaded the 199 *B. rapa* and 119 *B. oleracea* accessions from the NCBI website (PRJNA312457; Supplementary Table 1), raw reads were then mapped using bwa mem (default parameters) to the chromosomes of the reference and our nanopore assemblies for the two species. The proportion of mapped reads for each accession was computed from the BAM output files (Supplementary Tables 18 and 19 and Supplementary Fig. 13). We retrieved the 65.7 million and 188 million Illumina reads that could not be mapped to the reference genomes (To1000 or Chiifu). We were able to localize 52% and 39% of these reads to our assemblies, respectively. By screening the coverage along the HDEM and Z1 chromosomes, we found 2,735 and 3,508 regions (larger than 1 kb and with a coverage of at least 10×), respectively, representing 5.26 Mb and 7.61 Mb and spread over all of the chromosomes. We localized and annotated these regions as coding exon, intron or transposable element using bedtools (Supplementary Figs. 14 and 15 and Supplementary Table 20), and observed that nearly 20% of the bases (representing 1.14 Mb and 1.54 Mb for HDEM and Z1, respectively) were annotated as genic regions (intron + exon).

We computed best reciprocal hits using blastp between reference annotations and our gene predictions. We compared the gene order (synteny) for each couple of genome and searched for syntenic clusters of at least ten genes that are not located on the same chromosome. We found 1, 23 and 5 translocations between the *B. oleracea*, *B. rapa* and *Musa* (assembly V1) genomes involving 25, 951 and 658

genes, respectively. Synteny visualizations were performed using the MCScan tool, obtained from https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version) (Supplementary Figs. 7–9).

Estimation of sequencing costs. We computed the sequencing costs for each genome using the following public prices: US$500 and US$2,200 for each MinION and PromethION flow cell, US$2,146 for one-third of a HiSeq2500 flow cell (2×250 bp), US$175 for nanopore long-read library preparation, US$51 for Illumina PCR-Free library preparation, US$1,500 for a single BioNano Genomics chip and US$474 for the library preparation for two optical maps (BspQI and DLE-1). Given these costs, we estimated sequencing costs of US$13,621, US$12,271 and US$16,321 to generate the chromosome-scale assemblies of *B. rapa*, *B. oleracea* and *M. schizocarpa*, respectively, based on the MinION device. The sequencing cost for the *M. schizocarpa* assembly based on the PromethION device was estimated to be US$6,546.

Analysis of *FLC* genes. The *FLC* genes were searched using blastp and a database of *FLC* genes composed of predicted genes from a previous study[23] and the pan-genome of *B. oleracea*[22]. Five and three *FLC* genes were found in the gene catalogue of HDEM and Z1, respectively. The sixth and fourth genes were missing in the HDEM and Z1 annotation and were retrieved from the genomic sequence and added to the gene catalogue. Finally, we reported three *FLC1*, one disrupted *FLC2*, one *FLC3* and one *FLC5* genes for *B. oleracea* HDEM and one *FLC1*, one *FLC2*, one *FLC3* and one *FLC5* genes for *B. rapa* Z1. Phylogenic trees were built using Phylogeny.fr[55], a free, simple-to-use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences (Supplementary Fig. 17).

Identification and analysis of the self-incompatibility locus in the two *Brassica* genomes. Self-incompatibility emerged as an evolutionary strategy to foster genetic diversity and exchange in plant species. In Brassicaceae, self-incompatibility is controlled by a single multi-allelic locus (named the *S*-locus). We first identified the self-incompatibility alleles within each of the *Brassica* genomes with an unpublished *S*-locus genotyping pipeline using raw Illumina reads from shotgun sequencing of each individual (Z1 and HDEM) and a database of all available sequences of *SRK* (the self-incompatibility gene expressed in the pistil) from GenBank. Briefly, this pipeline (named NGSgenotyp) uses Bowtie2 to align raw reads against each reference sequence from the database and produces summary statistics with SAMtools (v1.4). We found that Z1 shared the same class II *S*-haplotype (*B. rapa S-60*) as Chiifu, whereas HDEM had a different class I *S*-haplotype (*B. oleracea S-13*) than To1000 (*B. oleracea S-28*). Using the full *SRK* sequences of these two *S*-alleles from GenBank, we localized the *S*-locus within each of the two *Brassica* assemblies with a blast search. We also identified the two other *S*-locus genes, *SCR/SP11*, the pollen-expressed gene, and *SLG*, a pistil-expressed *SRK* paralogue, within each assembly. Note that HDEM has full coding sequences of *SRK* and *SCR*, whereas To1000 apparently lacks the *SCR* gene as well as the first two exons of *SRK*. Then, we analysed the synteny in the *S*-locus and in flanking regions by comparative analyses (using blast) of the annotated assemblies with the corresponding *S*-locus regions in the *B. rapa* and *B. oleracea* reference genomes. Analysis of the two pairs of *S*-locus sequences revealed high sequence homology in genic and intergenic regions for the Z1–Chiifu pair as they share the same *S*-allele, whereas low overall homology was found for the HDEM–To1000 pair, which has distinct *S*-alleles. Finally, we produced a figure representing the annotated genes within the *S*-locus and its flanking regions using R and used mVista to estimate the degree of sequence homology in the *S*-locus region between our two *Brassica* assemblies and the two reference genomes (Supplementary Fig. 19).

Analysis of assembly completeness of *R*-gene clusters. Three orthologous disease resistance-like genes clusters identified in DH-Pahang assembly version 1 (ref. [16]) were searched in DH-Pahang version 2 (ref. [17]) and in the assembly of *M. schizocarpa* with gene identifiers and blast search, respectively. Cluster boundaries were manually refined based on the gene annotation of these two genomes and the proportion of N was calculated for each region (Supplementary Table 21). The proportions of N reported in results were calculated as the N sum in the three clusters divided by their cumulated size.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The genome assemblies, gene predictions and genome browsers are freely available at http://www.genoscope.cns.fr/plants. The Illumina, MinION and PromethION data, the assemblies and the annotations are available in the European Nucleotide Archive under the following projects: PRJEB26620 (*B. rapa*), PRJEB26621 (*B. oleracea*) and PRJEB26661 (*M. schizocarpa*). Germplasm for these genomes will be made freely and publicly available to the entire community. *M. schizocarpa* germplasm is available at Bioversity International Transit Center under ITC number ITC0926. *B. rapa* ssp. *trilocularis* (genotype Z1) is available at the Plant Genetic Resources of Canada and *B. oleracea* ssp. *italica* (genotype HDEM) is available at the Biological Resource Center BrACySol, Rennes, France. All supporting data are included in the Supplementary Information.

Received: 23 May 2018; Accepted: 24 September 2018;
Published online: 2 November 2018

References

1. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
2. Jiao, W. B. & Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant. Biol.* **36**, 64–70 (2017).
3. Michael, T. P. et al. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* **9**, 541 (2018).
4. Schmidt, M. H. et al. De novo assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* **29**, 2336–2348 (2017).
5. Arabidopsis Genome Initiative Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
6. International Rice Genome Sequencing Project The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
7. Du, H. et al. Sequencing and de novo assembly of a near complete *indica* rice genome. *Nat. Commun.* **8**, 15324 (2017).
8. Edger, P. P. et al. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *Gigascience* **7**, 1–7 (2018).
9. Dassanayake, M. et al. The genome of the extremophile crucifer *Thellungiella parvula*. *Nat. Genet.* **43**, 913–918 (2011).
10. International Brachypodium Initiative Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
11. Raymond, O. et al. The *Rosa* genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
12. Cheng, F. et al. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nat. Genet.* **48**, 1218–1224 (2016).
13. Cai, C. C. et al. *Brassica rapa* genome 2.0: a reference upgrade through sequence re-assembly and gene re-annotation. *Mol. Plant* **10**, 649–651 (2017).
14. Wang, X. W. et al. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039 (2011).
15. Parkin, I. A. et al. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol.* **15**, R77 (2014).
16. D'Hont, A. et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213–217 (2012).
17. Martin, G. et al. Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* **17**, 243 (2016).
18. Lam, E. T. et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
19. Sakai, H. et al. The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Sci. Rep.* **5**, 16780 (2015).
20. Wang, X. et al. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nat. Genet.* **49**, 765–772 (2017).
21. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
22. Golicz, A. A. et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* **7**, 13390 (2016).
23. Schranz, M. E. et al. Characterization and effects of the replicated flowering time gene *FLC* in *Brassica rapa*. *Genetics* **162**, 1457–1468 (2002).
24. Goubet, P. M. et al. Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in *Arabidopsis*. *PLoS Genet.* **8**, e1002495 (2012).
25. Shiba, H. et al. Genomic organization of the *S*-locus region of *Brassica*. *Biosci. Biotechnol. Biochem.* **67**, 622–626 (2003).
26. Bachmann, J. A., Tedder, A., Laenen, B., Steige, K. A. & Slotte, T. Targeted long-read sequencing of a locus under long-term balancing selection in *Capsella*. *G3 (Bethesda)* **8**, 1327–1333 (2018).
27. Kim, D., Jung, J., Choi, Y. O. & Kim, S. Development of a system for *S* locus haplotyping based on the polymorphic *SLL2* gene tightly linked to the locus determining self-incompatibility in radish (*Raphanus sativus* L.). *Euphytica* **209**, 525–535 (2016).
28. Yang, J. H. et al. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* **48**, 1225–1232 (2016).
29. Jarvis, D. E. et al. The genome of *Chenopodium quinoa*. *Nature* **542**, 307–312 (2017).
30. Jiao, W. B. et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.* **27**, 778–786 (2017).

31. Reyes-Chin-Wo, S. et al. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nat. Commun.* **8**, 14953 (2017).
32. Teh, B. T. et al. The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat. Genet.* **49**, 1633–1641 (2017).
33. Gawel, N. J. & Jarret, R. L. A modified CTAB DNA extraction procedure for *Musa* and *Ipomoea*. *Plant Mol. Biol. Rep.* **9**, 262–266 (1991).
34. Risterucci, A. M. et al. A high-density linkage map of *Theobroma cacao* L. *Theor. Appl. Genet.* **101**, 948–955 (2000).
35. Engelen, S. & Aury J. M. Fastxtend tool (Genoscope/CEA, 2015); http://www.genoscope.cns.fr/fastxtend/
36. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
37. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
38. Vaser, R. et al. Ra assembler. v. git commit 65bedfe (Faculty of Electrical Engineering and Computing, University of Zagreb, 2017); https://github.com/rvaser/ra
39. Ruan, J. et al. SMARTdenovo assembler. v. git commit 3d9c22e (Agricultral Genomics Insititute, China, 2015) ; https://github.com/ruanjue/smartdenovo
40. Wick, R. et al. Fitlong tool. v. git commit 8d81024 (University of Melbourne, 2017); https://github.com/rrwick/Filtlong
41. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
42. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
43. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
44. de Givry, S., Bouchez, M., Chabrier, P., Milan, D. & Schiex, T. CARHTA GENE: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* **21**, 1703–1704 (2005).
45. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
46. RepeatMasker Open-4. 0 (Institute for Systems Biology, 2013); http://www.repeatmasker.org
47. Chalhoub, B. et al. Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950–953 (2014).
48. Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).
49. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
50. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
51. Dubarry, M. et al. Gmove a tool for eukaryotic gene predictions using various evidences (poster). *F1000Res.* **5**, 681 (2016).
52. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
53. Marcais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
54. Nettstad M. Dot (DNA Nexus, 2017); http://github.com/dnanexus/dot
55. Dereeper, A. et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* **36**, W465–W469 (2008).

Acknowledgements

This work was supported by the Genoscope, the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) and France Génomique (ANR-10-INBS-09-08). We are grateful to ONT for early access to the MinION device through the MinION Access Programme and we thank their staff for technical help. Work by X.V. and M.G. is supported financially by Région Hauts-de-France, the Ministère de l'Enseignement Supérieur et de la Recherche (CPER Climibio) and the European Fund for Regional Economic Development.

Author contributions

C.F., G.D., F.-C.B., E.D. and C.C. extracted the DNA. C.C. and A.L. optimized and performed the sequencing. E.D., W.B. and V.B. generated the optical maps. P.D., R.D. and M.M.-D. generated the genetic map for the *B. oleracea* HDEM accession. B.I., C.B. and J.-M.A. performed the genome assemblies. G.M. performed the anchoring of the *M. schizocarpa* scaffolds. C.F., J.M. and M.R.-G. performed the anchoring of the *B. oleracea* scaffolds. M.D. and J.-M.A. performed the anchoring of the *B. rapa* scaffolds. M.D. and B.N. performed the gene prediction for the genome assemblies. B.I., C.B., M.D., F.D., J.-M.A. and S.E. performed the bioinformatic analyses. X.V. and M.G. performed the *S*-locus annotation of the two Brassicaceae genomes. B.I., C.B., M.D. and J.-M.A. wrote the article. A.D., A.-M.C., P.W. and J.-M.A. supervised the study.

Competing interests

The authors declare no competing interests. B.I., S.E., C.C., P.W. and J.-M.A. are part of the MinION Access Programme and J.-M.A. received travel and accommodation expenses to speak at ONT conferences.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0289-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.-M.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): AURY Jean-Marc

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐ Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Sequencing data were obtained using software available with sequencing machines: RTA (1.18.64) and bcl2fastq from Illumina (2.17.1.14 and 2.19.1.403), MinKnow (1.5.18 --> 1.13.1), Metrichor (2.43.1 and 2.45.3) and Albacore (1.0.4 --> 2.1.10) from Oxford Nanopore. Optical maps were analyzed using the software provided by BioNano Genomics (Bionano Solve Pipeline version 3.1.1 and Bionano Access version 1.0a), the vendor of the system. Genome assembly and gene prediction were performed using open-source tools : fitlong (git commit 8d81024), Ra (git commit 65bedfe), smartdenovo (git commit 3d9c22e), racon (2018.3.2), pilon (1.22), blat (v36), blast (2.2.26), genewise (2.2.0), trf (4.09), RepeatMasker (4.0.5), gmove (v2) and are all referenced in the method section.
Data analysis	Commercial software were not used and we rely on circos (v0.66) and R (3.3.1) for statistical analyses and figure generation.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The genome assemblies, gene predictions and genome browsers are freely available at http://www.genoscope.cns.fr/plants. The Illumina, MinION and PromethION data, the assemblies and the annotations are available in the European Nucleotide Archive under the following projects: PRJEB26620 (B. rapa), PRJEB26621 (B. oleracea) and PRJEB26661 (M. schizocarpa). Germplasm for these genomes will be made freely and publicly available to the entire community. Musa schizocarpa germplasm is available at Bioversity International Transit Center under ITC number ITC0926. Brassica rapa ssp trilocularis (genotype Z1) is available at Plant Genetic Resources of Canada, PGRC and Brassica oleracea ssp italica (genotype HDEM) is available at the Biological Resource Center BrACySol, Rennes, France.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the comparison of plant genome assemblies, we selected genomes with a chromosome-scale assembly that were available through the NCBI website and added the recently published genomes (amborella, betula, B. rapa v2, camptotheca, cephalotus, coffea, fragaria, musa, oropetium, physcomitrella, quercus, solanum nanopore, vigna pacbio, vitis pacbio) that were not available from the NCBI website (using phytozome or the consortium website).
Data exclusions	No data exclusions in this manuscript
Replication	No replication in this manuscript
Randomization	No randomization in this manuscript as genomes assemblies were not allocated into experimental groups.
Blinding	No blinding in this manuscript as the data were not allocated into groups

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Germplasm for these genomes will be made freely and publicly available to the entire community. Musa schizocarpa germplasm is available at Bioversity International Transit Center under ITC number ITC0926. Brassica rapa ssp trilocularis (genotype Z1) is available at Plant Genetic Resources of Canada, PGRC and Brassica oleracea ssp italica (genotype HDEM) is available at the Biological Resource Center BrACySol, Rennes, France.