Approches bioinformatiques pour l'exploration des génomes et de la biodiversité

Caroline Belser

THESE DE DOCTORAT

NNT : 2022UPASB073

Approches bioinformatiques pour l'exploration des génomes et de la biodiversité

Bioinformatics approaches for genome and biodiversity exploration

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 567 : sciences du végétal : du gène à l'écosystème (SEVE)
Spécialité de doctorat : sciences du végétal
Graduate School : BioSphERA - Biologie, Société, Ecologie & Environnement, Ressources, Agriculture & Alimentation. Référent : Faculté des sciences d'Orsay

Thèse par VAE préparée dans l'unité de recherche Génomique métabolique (Université Paris-Saclay, Univ Evry, CNRS, CEA), sous la direction de **: Benhamed Moussa**, professeur Université Paris Cité.

Thèse soutenue à Paris Saclay, le 16 décembre 2022, par

Caroline Belser Menguy

Composition du Jury

Florian Frugier Directeur de recherche, CNRS	Président du jury
Aline Probst Directrice de recherche, CNRS	Rapporteur & Examinatrice
Mohammed Bendahmane Directeur de recherche, INRAE	Rapporteur & Examinateur
Hélène Neyret-Kahn Ingénieure de recherche, Institut Curie	Examinatrice
Olivier Martin Directeur de recherche, INRAE	Examinateur
Moussa Benhamed Professeur, Université Paris Cité	Directeur de thèse

université PARIS-SACLAY | ÉCOLE DOCTORALE Sciences du végétal : du gène à l'écosystème (SEVE)

Titre : Approches bioinformatiques pour l'exploration des génomes et de la biodiversité.

Mots clés : séquençage, assemblage, génome, plantes

L'avènement des nouvelles technologies de séquençage puis du séquençage longues lectures a ouvert la voie vers la production de génomes de référence à un coût abordable. Les outils bioinformatiques, tout comme les techniques de laboratoire, ne cessent d'évoluer et il est important de tester et mettre en place toutes ces technologies de pointe.
Ce mémoire de thèse présente, au travers de différentes publications, les développements méthodologiques mis en place dans le cadre d'assemblages de génomes de référence. Les longues lectures obtenues grâce au séquençage avec la technologie Oxford Nanopore nous ont permis d'améliorer la continuité de nos assemblages et de reconstituer plus fidèlement des régions souvent invisibles avec des méthodes traditionnelles, notamment les régions contenant des séquences répétées ou des clusters de gènes dupliqués. La technologie de carte optique proposée par Bionano Genomics, et le séquençage de banques Hi-C nous ont permis d'amener nos assemblages à l'échelle des chromosomes et d'accéder à la structure de régions dites complexes comme les centromères.
J'ai pu mettre en évidence l'importance de ces nouveaux assemblages pour différents génomes de plantes, pour lesquels ils constituent une ressource très précieuse pour de futures recherches sur l'histoire évolutive des espèces, sur les adaptations à de nouvelles conditions de vie ou sur les gènes de résistance aux pathogènes.
Ces développements vont être adaptés pour répondre aux besoins de nouveaux projets de production de génomes de référence à grande échelle et dont le but est de préserver l'information génétique de la biodiversité.

Title : Bioinformatics approaches for genome and biodiversity exploration.

Keywords : sequencing, assembly, genome, plants

Abstract : The emergence of new sequencing technologies, as long-read sequencing, has paved the way for the production of reference genomes at an affordable cost. Bioinformatics tools, as well as laboratory techniques, are constantly evolving and it is important to test and implement state-of-the-art technologies.
This thesis presents, through different publications, the methodological developments implemented in the framework of reference genome assemblies. The long reads obtained using the Oxford Nanopore sequencing technology allowed us to improve the contiguity of our assemblies and to accurately reconstruct regions that were resistant to older sequencing technologies, as regions containing repeated sequences or clusters of duplicated genes. The optical map technology proposed by Bionano Genomics and the sequencing of Hi-C libraries, allowed us to produce assemblies at the chromosome scale and to access the structure of complex regions such as centromeres.
I have been able to demonstrate the importance of these enhanced assemblies with different plant genomes, for which they provide a valuable resource for future research on evolutionary history, adaptations to new living conditions or resistance genes.
These developments will be adapted to meet the needs of future large-scale projects aimed at producing reference genome and preserving the genetic information of biodiversity.

1 REMERCIEMENTS

Ce mémoire de thèse par Valorisation des Acquis de l'Expérience regroupe le travail réalisé ces 7 dernières années sur les méthodes d'assemblage de génomes. Il a pour vocation de présenter tout le travail de recherche et développement effectué en partenariat avec les différentes équipes du Genoscope mais également avec nos collaborateurs.

Soutenir cette thèse est pour moi l'achèvement de mon parcours pour le moins atypique et l'obtention du titre de Docteur représente une mise en adéquation entre mes activités et mes compétences.

Je remercie Mme Delarue qui, en acceptant mon dossier de thèse par VAE, m'a offert la chance de présenter mon travail.

Je remercie M. Benhamed qui a gentiment accepté de me suivre dans cette démarche en tant qu'accompagnateur VAE et m'a apporté conseils et soutien.

Je suis très reconnaissante envers Mme Probst et M. Bendahmane d'avoir accepté de juger mon travail de thèse en tant que rapporteur.

Je remercie Mme Neyret-Kahn, M. Martin et M. Frugier de m'avoir fait l'honneur de participer au jury de cette thèse.

Tout ceci n'a été possible que parce que je suis entourée au sein du LBGB par une équipe soudée. Je remercie tout d'abord Jean-Marc qui a accepté que je rejoigne cette équipe après l'obtention de mon Master en 2013. Merci pour ta confiance. Un grand merci à Corinne DS, qui m'a mis le pied à l'étrier en étant ma maître de stage de Master. Tu as su trouver le juste milieu entre ma soif d'apprendre et mon besoin d'autonomie. Depuis tu es là pour m'aider et me soutenir et pas que pour le travail ! Un grand merci à France qui est une source de soutien et d'inspiration. Tu me tires vers le haut. Un grand merci pour ta précieuse relecture du mémoire. Nous partageons toutes les trois le même bureau depuis des années et je n'aurais pas pu mieux tomber. Nous pouvons tout partager, notre vie personnelle comme professionnelle et c'est si précieux! Ensuite merci à Stéfan qui m'a épaulée dans mes premières activités et dans mon apprentissage du awk! A mon petit Fred que j'embête très souvent! Je remercie également Benjamin I., mon binôme de choc plein de talent! Toujours prêt à aider, à discuter. Mais je n'oublie pas tous les autres membres de l'équipe. Ils participent à cette dynamique de bien des façons.

Ensuite je pense à ma première famille du Genoscope, le Laboratoire de Séquençage. Merci à toute l'équipe avec qui j'ai partagé plein de moments de vie et beaucoup de fous rires! Je pense tout particulièrement à Céline, à Odette, à Laurie, à Emilie et à Elodie! Merci aussi pour les footings le midi qui vident la tête, nos discussions sur nos familles, nos galères, nos vacances! Un merci tout particulier également à Corinne C.

qui a été là pour moi à un moment où je remettais beaucoup de choses en question et qui m'a tant appris. Merci à Karine et Julie qui m'ont fait confiance en me confiant certaines responsabilités. J'apprécie de travailler avec vous trois sur tous nos développements liés à la plateforme, j'apprécie nos longues et interminables discussions qui vont bien au-delà du travail! Merci également à Adriana pour son énergie, ces disvussions et ces collaborations toujours fructueuses.

Je remercie également Patrick Wincker, qui m'a embauchée au Genoscope en 2000 et qui m'a fait confiance toutes ces années. Merci de m'avoir permis de réaliser mon Master et maintenant cette thèse, de m'avoir permis d'évoluer dans mes activités.

Je dois remercier aussi tous les collaborateurs qui nous apportent des projets passionnants, comme les équipes de l'IGEPP de Rennes et du CIRAD de Montpellier.

Bien sûr, je ne peux pas finir sans penser à ma famille, à mon mari et à mes filles. J'avais à cœur de leur montrer que tout est possible même si le chemin est sinueux, et qu'il n'est jamais trop tard.

"Dans un voyage ce n'est pas la destination qui compte mais toujours le chemin parcouru, et les détours surtout."

Philippe Pollet-Villard

2 RESUME EN FRANÇAIS

L'avènement des nouvelles technologies de séquençage puis du séquençage longues lectures a ouvert la voie vers la production de génomes de référence à un coût abordable. Les outils bioinformatiques, tout comme les techniques de laboratoire, ne cessent d'évoluer et il est important de tester et mettre en place toutes ces technologies de pointe.

Ce mémoire de thèse présente, au travers de différentes publications, les développements méthodologiques mis en place dans le cadre d'assemblages de génomes de référence. Les longues lectures obtenues grâce au séquençage avec la technologie Oxford Nanopore nous ont permis d'améliorer la continuité de nos assemblages et de reconstituer plus fidèlement des régions souvent invisibles avec des méthodes traditionnelles, notamment les régions contenant des séquences répétées ou des clusters de gènes dupliqués. La technologie de carte optique proposée par Bionano Genomics, et le séquençage de banques Hi-C nous ont permis d'amener nos assemblages à l'échelle des chromosomes et d'accéder à la structure de régions dites complexes comme les centromères.

J'ai pu mettre en évidence l'importance de ces nouveaux assemblages pour différents génomes de plantes, pour lesquels ils constituent une ressource très précieuse pour de futures recherches sur l'histoire évolutive des espèces, sur les adaptations à de nouvelles conditions de vie ou sur les gènes de résistance aux pathogènes.

Ces développements vont être adaptés pour répondre aux besoins de nouveaux projets de production de génomes de référence à grande échelle et dont le but est de préserver l'information génétique de la biodiversité.

3 RESUME EN ANGLAIS

The emergence of new sequencing technologies, as long-read sequencing, has paved the way for the production of reference genomes at an affordable cost. Bioinformatics tools, as well as laboratory techniques, are constantly evolving and it is important to test and implement state-of-the-art technologies.

This thesis presents, through different publications, the methodological developments implemented in the framework of reference genome assemblies. The long reads obtained using the Oxford Nanopore sequencing technology allowed us to improve the contiguity of our assemblies and to accurately reconstruct regions that were resistant to older sequencing technologies, as regions containing repeated sequences or clusters of duplicated genes. The optical map technology proposed by Bionano Genomics and the sequencing of Hi-C libraries, allowed us to produce assemblies at the chromosome scale and to access the structure of complex regions such as centromeres.

I have been able to demonstrate the importance of these enhanced assemblies with different plant genomes, for which they provide a valuable resource for future research on evolutionary history, adaptations to new living conditions or resistance genes.

These developments will be adapted to meet the needs of future large-scale projects aimed at producing reference genome and preserving the genetic information of biodiversity.

4 SOMMAIRE

Table des matières

5 LISTE DES FIGURES

6 GLOSSAIRE

Basecalling : transformation en base d'un pic de chromatogramme ou de modification de courants électriques.

Carte optique : elle est obtenue grâce à la technologie Bionano Genomics. Elle contient la position des sites de coupure d'une enzyme de restriction et la distance entre deux sites de coupure adjacents le long d'un génome.

Cluster : amplification avant séquençage d'un seul et unique fragment d'ADN. On obtient une amplification clonale.

Contig : dérive du mot "contiguous". Il représente une séquence d'ADN continue obtenue à partir d'un sous-set de séquences chevauchantes ou partageant les mêmes k-mers.

Flow cell : support du séquençage. Ce terme est employé pour la technologie Illumina et Oxford Nanopore.

Hybrid scaffold : un hybrid scaffold est un scaffold obtenu grâce à la technologie Bionano Genomic. Les contigs sont ordonnés et orientés les uns par rapport aux autres grâce à une carte optique.

k-mers : sous-chaîne de longueur k contenue dans une séquence.

Metabarcoding : Séquençage grâce aux nouvelles technologies de séquençage de pool d'amplicons obtenus par amplification d'un gène marqueur sur un échantillon contenant un mélange d'espèces (prélèvements environnements, microbiotes …).

Scaffold : un scaffold est composé de contigs et de trous ("gaps"). Les contigs y sont ordonnés et orientés les uns par rapport aux autres et séparés par des gaps.

7 ABREVIATIONS

ADN : Acide Désoxyribonucléique
ARN : Acide Ribonucléique
BAC : Bacterial Artificial Chromosome
BiSCoT : Bionano Scaffolding Correction Tool
BNG : Bionano Genomics
CCS : Circular Consensus Sequence
DLS : Direct Label and Stain
DLE-1 : Direct Label Enzyme -1
dNTP : désoxyribonucléotide triphosphate
ddNTP : didésoxyribonucléotide triphosphate
emPCR : PCR en émulsion
ERGA : European Genome Reference Atlas
Hi-C : Chromosome Conformation Capture
LTR : Long Terminal Repeat
LTR-R : Long terminal repeat retrotransposon
NGS : Next generation Sequencing
NLR : Nucleotide-binding and leucine-rich-repeat proteins
NLRS : Nick - Label - Repair and Stain
OLC : Overlap–layout–consensus
ONT : Oxford Nanopore Technology
PacBio : Pacific Biosciences
PCR : Polymerase Chain Reaction
PTP : PicoTiter Plate
SBS : Sequencing by Synthesis
SMRT : Single Molecule Real Time Sequencing
TADs : Topologically associated domains
TE : Transposable Element
TGS : Third Generation Sequencing
VGP : Vertebrate Genome Project
YAC : Yeast Artificial Chromosome

8 PARCOURS PERSONNEL

8.1 DE 2000 A 2012 : LE LABORATOIRE DE SEQUENCAGE

Ma carrière au Genoscope a débuté en 2000 au sein du laboratoire de séquençage en qualité de technicienne supérieure de laboratoire. Le Genoscope réalisait alors le séquençage du chromosome 14 humain dans le cadre du Human Genome Project. Je prenais part aux activités de production de séquences, de développements méthodologiques et à la mise en place des nouvelles technologies. J'ai ainsi travaillé sur les premières technologies de séquenceurs (Licor puis ABI3730 de chez Perkin Elmer) puis participé à la mise en place des nouvelles technologies de séquençage (Genome Sequencer de chez Roche 454, SOLID de chez Life Technologies et Genome Analyzer de chez Illumina). J'ai participé aux premières formations menées par les distributeurs, aux tests de mise en place des protocoles et réalisé les formations du personnel de notre plateforme de séquençage. J'ai également pu prendre part à l'organisation de la plateforme, au suivi de la production et des projets. Mon rôle était d'assurer le flux de production, de veiller au respect des bonnes pratiques et de former le personnel aux nouveaux protocoles. J'ai commencé à interagir avec les bio-informaticiens grâce au suivi des projets. Ceci m'a donné l'envie de développer mes compétences en bioinformatique pour participer aux activités de recherche du Genoscope. Je me suis alors inscrite au Master de Biologie Informatique de Paris 7 dont le programme me permettait d'acquérir les compétences que je visais.

8.2 DEPUIS 2013 : RECHERCHE ET DEVELOPPEMENT POUR LE LABORATOIRE DE SEQUENCAGE

En 2013, après l'obtention de mon master, j'ai intégré le Laboratoire de Bioinformatique pour la Génomique et la Biodiversité (LBGB) sous la direction de Jean-Marc Aury. Ce laboratoire regroupe plusieurs activités : la gestion du flux de données (développement et évolution des pipelines réalisant les analyses qualité des séquences), l'évaluation des technologies de séquençage, le développement méthodologique autour de l'assemblage et de l'annotation des génomes et enfin la

génomique comparative pour l'analyse des génomes produits. Je me suis tout de suite impliquée auprès du service de recherche et développement du laboratoire de séquençage. C'est une activité transversale par rapport à celles du LBGB qui me permet d'être en relation avec tous les acteurs et de garder une vision globale.

Le laboratoire de séquençage se doit d'adapter ses protocoles et ses techniques de séquençage au fur et à mesure des évolutions technologiques mais également des demandes liées aux projets réalisés sur la plateforme. Les nouvelles procédures sont validées par l'analyse des données produites lors de la phase de test. Le schéma des expériences à conduire, ainsi que le choix des organismes modèles à utiliser lors de ces tests font l'objet de discussions préliminaires avec la plateforme de séquençage. Les données sont ensuite analysées pour mettre en évidence d'éventuels biais introduits par l'utilisation de kits commerciaux, vérifier la robustesse et la reproductibilité des expériences, tester les limites des différents protocoles etc. Il en est de même pour toute nouvelle technologie de séquençage. Ces tests sont toujours réalisés en les mettant en perspective des projets de séquençage et peuvent conduire à l'écriture d'articles scientifiques. A titre d'exemple, ce fut le cas pour une étude menée avec le Dr. Adriana Alberti sur l'impact des méthodes de préparation de banques de séquençage à partir des ARN bactériens (Alberti, A., Belser, C., et al. Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* 15, 912 (2014)). J'ai ainsi pu mettre à profit mes compétences acquises précédemment en participant au design des expériences mais également mes nouvelles compétences en bio-informatique en réalisant l'ensemble des analyses.

Mes activités se sont ensuite progressivement organisées autour de deux grandes thématiques : l'amélioration des assemblages de génomes et l'évaluation de la qualité et l'analyse dans le cadre de projets de génomique environnementale (metabarcoding).

8.3 Depuis 2015 : amelioration de la continuite des assemblages, mise en place des technologies pour la reconstruction des chromosomes

Mon premier axe de recherche consiste à évaluer de nouvelles méthodes destinées à l'amélioration des assemblages de génomes. En effet, le Genoscope est impliqué dans de nombreux projets ayant pour but de produire des génomes de référence (plantes, insectes, coraux, poissons…). L'obtention d'un génome de référence de haute qualité est une étape indispensable pour l'analyse de ces derniers en recherche fondamentale (phylogénie, génomique comparative) mais également dans des domaines appliqués (évolution de ces espèces, amélioration des espèces cultivées, résistance aux pathogènes, conservation des espèces).

Pour rester au niveau des standards de qualité des génomes qui évoluent sans cesse, il faut mettre en place les technologies de pointe très rapidement, les évaluer et les utiliser au profit des projets réalisés au Genoscope.

Toujours en partenariat avec la plateforme de production, j'ai participé à la mise en place de deux méthodes permettant l'amélioration de la continuité des assemblages: la génération de cartes optiques, réalisées grâce à l'instrument Saphyr commercialisé par Bionano Genomics, et le séquençage de banques Hi-C (chromosome conformation capture sequencing).

Pour ces travaux, j'ai, notamment, encadré en 2016 un stagiaire de Master 2 de biologie informatique (Université Paris 7). Son stage consistait à évaluer des outils alternatifs à ceux proposés par la société Bionano Genomics.

La mise en place des cartes optiques a permis l'obtention de génomes de référence à l'échelle des chromosomes. Les articles suivants décrivent les stratégies mises en place dans le cadre de deux projets de séquençage : Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps (Belser et al. *Nature Plants* 4, 879–887 (2018)) et Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing (Belser et al.

Commun Biol 4, 1047 (2021).). Ces articles sont présentés dans le mémoire de thèse.

Le premier des deux articles a été l'un des premiers à mettre en avant l'utilisation des cartes optiques basées sur l'enzyme de restriction DLE-1 pour l'obtention d'un génome à l'échelle du chromosome, devenu le standard aujourd'hui.

J'ai coécrit un article présentant une comparaison des méthodes Hi-C et cartes optiques pour l'obtention d'un génome de la plante Brassica rapa (Istace B, Belser C, et al. Sequencing and Chromosome-Scale Assembly of Plant Genomes, Brassica rapa as a Use Case. *Biology*. 2021; 10(8):732.). Cet article est également présenté dans le mémoire de thèse.

Cependant, des développements technologiques et méthodologiques sont encore nécessaires pour parvenir au même niveau de résultats pour des génomes plus complexes (génome de grande taille, génomes polyploïdes, …)

Depuis quelques années, de nombreuses initiatives visant à caractériser la biodiversité ont émergé. Ces initiatives ont pour objectif de fournir un génome de haute qualité pour chaque espèce. Le Genoscope prend part à certaines de ces initiatives telles que European Reference Genome Atlas (ERGA) et a même proposé un programme à l'échelle du territoire français, le programme ATLASea (an atlas of marine genomes). Ces projets vont permettre de cataloguer l'information génétique de la biodiversité à des fins de protection, d'observation et de conservation des espèces. Je vais être personnellement impliquée dans la réalisation de ces grands programmes, demandant une grande adaptabilité et la mise en place de protocoles spécifiques à tous ces organismes.

8.4 Depuis 2016 : etude de la biodiversite grace au metabarcoding

L'étude de la biodiversité à plus grande échelle est une des grandes thématiques de mon équipe de recherche mais également du Genoscope. Répertorier les espèces qui peuplent un environnement

donné est permis grâce au Metabarcoding. Celui-ci consiste à amplifier et séquencer une région cible simultanément pour l'ensemble des espèces présentes dans un échantillon environnemental. Cette région cible peut être la partie hypervariable des gènes ribosomaux codant pour la sous-unité 16S, 18S, la région intergénique Internal transcribed spacer 2 (ITS2), la gène chloroplastique rbcl etc.. Mon deuxième axe de travail consiste donc à mettre en œuvre des méthodes bio-informatiques permettant la validation, les pré-traitements et l'analyse des données de Metabarcoding.

Les projets eDNAbyss (projet France Génomique « Et pourquoi pas les Abyss? ») et Tara Pacific (projet France Génomique Tara expéditions - Tara Pacific) ont pour but d'explorer la biodiversité d'écosystèmes marins. Le Metabarcoding y est réalisé à partir de prélèvements d'eau, de sédiments, de coraux ou encore de poissons. Dans le cadre de ces projets, j'ai développé un pipeline d'analyse permettant de détecter d'éventuelles contaminations, introduites par exemple par les réactifs de séquençage, et éliminant ainsi des biais dans l'analyse de ces données. Ce pipeline s'appuie sur des outils libres d'accès, il traite tous les échantillons produits au Genoscope dès la fin du séquençage et produit des rapports sur le niveau de contamination des différents échantillons.

Ces développements sont valorisés dans une publication qui présente toutes les méthodes expérimentales et bioinformatiques utilisées pour générer les séquences du projet Tara Pacific (Belser C, Poulain J Poulain, et al. Integrative omics framework for characterization of coral reef ecosystems from the Tara Pacific expedition. 2022. https://doi.org/10.48550/arXiv.2207.02475 - article accepté, en cours de publication dans *Scientific data*).

J'ai ensuite développé un pipeline d'analyse permettant, à partir des séquences issues du Metabarcoding, de déterminer la diversité taxonomique d'un échantillon. Cette méthode est, par exemple, utilisée pour déterminer la composition bactérienne de sols contaminés à la chlordécone dans le cadre d'une étude menée sur la remédiation naturelle de ce pesticide hautement toxique et persistant[1].

Des développements complémentaires ont été réalisés conjointement

avec une étudiante en alternance (Master 2 Bioinformatique Modélisation et Statistique - Université de Rouen) que j'ai encadré de septembre 2021 à septembre 2022. Par la suite, je souhaite encadrer un étudiant en Master 2 afin d'évaluer et développer une méthode d'analyse du Metabarcoding généré grâce à la technologie Oxford Nanopore.

Je participe également cette année au dépôt d'un dossier pour l'appel à projet ANR. Ce projet a pour but l'étude de la biodiversité marine des grands fonds et des côtes françaises. J'y suis leader de groupe de travail pour l'analyse et l'interprétation des données de Metabarcoding.

8.5 2022 : PRESENTATION DU DOCTORAT PAR VAE

Dans le cadre de mon mémoire de thèse, j'ai souhaité présenter mes réalisations autour de l'amélioration des assemblages. Elles sont le fruit d'un intense et passionnant travail collaboratif, avec les membres de mon équipe, de l'équipe de production de séquences mais également avec les collaborateurs porteurs de projets. Ces projets collaboratifs m'ont permis d'acquérir et de développer de multiples compétences à la fois techniques et scientifiques. Par exemple, la mise en place de nouvelles technologies en interaction avec les équipes de production a exacerbé mon goût pour le développement de nouvelles approches au service de l'amélioration des ressources génomiques. Cette expérience s'accompagne d'une montée en compétences qui me permet aujourd'hui d'appréhender tous les enjeux restant à maîtriser. Les compétences techniques initiales, essentielles pour mener des travaux de recherche de façon autonome, ont été acquises lors de mon Master et grâce à différentes formations indispensables pour maîtriser tous les outils nécessaires (formations au langages Python et R). Je travaille à présent sur des techniques de pointe permettant d'obtenir la séquence des génomes avec une qualité encore inatteignable il y a quelques années. Ces résultats novateurs m'ont conduite à réaliser des analyses de génomique comparative orientées vers le monde de la recherche. J'ai ainsi mené ces sept dernières années des travaux de recherche, allant de la définition de la question scientifique, en passant par la mise en œuvre de méthodes innovantes et jusqu'à la valorisation de ces travaux à travers des présentations orales lors de congrès

scientifiques ou encore la rédaction d'articles scientifiques.

L'obtention de ce doctorat est pour moi une étape clé qui valorise l'expérience que j'ai acquise durant ces dernières années mais surtout me permet d'être en adéquation avec mes activités actuelles dans le laboratoire. J'espère ainsi que la lecture de ce mémoire vous permettra d'évaluer l'évolution de mon parcours et de mes compétences.

9 RESULTATS DE RECHERCHE

9.1 INTRODUCTION

9.1.1 Histoire de la génétique

L'histoire de la génétique et par la suite de la génomique a été jalonnée par de nombreuses découvertes qui ont progressivement conduit à élucider l'origine du support de l'hérédité puis à déchiffrer des génomes entiers de nombreux organismes de l'arbre du vivant (Figure 1).

Dès l'Antiquité, les Hommes ont constaté l'existence d'une grande variabilité entre les espèces, composées d'individus qui se ressemblent, et de la transmissibilité des caractères des parents vers leur progéniture. Hippocrate (460-377 av J.C) émet l'hypothèse que l'embryon se forme par le mélange de la semence du père et de la mère et que ces semences proviennent de toutes les parties du corps. Aristote (427-346 av J.C) affirme que la semence paternelle est une substance génératrice de forme alors que la semence maternelle est juste nourricière. Il y aurait une lutte entre les deux semences. Si la semence du père l'emporte, l'enfant sera un garçon.

D'autres théories émergent sur l'origine de l'embryon. La théorie de la préformation soumet l'idée que l'être vivant préexiste avant sa conception et voit l'embryon comme un être vivant « miniature » où tous les organes sont déjà présents. Elle rejette l'apparition de la complexité des organismes au cours du développement. La théorie de l'épigenèse, quant à elle, propose que l'embryon d'un être vivant se développe par multiplication et différenciation cellulaire progressive. L'embryon se construit peu à peu, par complications successives, par épigenèse. Aristote, partisan de l'épigenèse, soutient aussi que des facteurs génétiques ont un rôle prépondérant dans le développement d'un organisme.

À la fin du XVIIe siècle, la théorie préformiste est soutenue par l'Église: Dieu étant le créateur de toute chose, il a, dès le commencement, créé tous les animaux, toutes les plantes et tous les hommes amenés à

peupler le monde jusqu'à la fin des temps. Les enfants à naître existent donc déjà, minuscules mais totalement formés, dans leurs géniteurs; ces enfants eux-mêmes abritent, dans cet état minuscule, leurs enfants et, par emboîtements successifs, toutes les générations suivantes.

Des scientifiques de premier plan sont alors partisans du préformisme, tout comme Diderot et d'Alembert qui présentent cette hypothèse comme étant la plus crédible dans leur *Encyclopédie*.

Pendant tout le XVIIIe siècle, la polémique entre épigénèse et préformation sera féroce. Elle prendra fin au XIXe siècle suite à la découverte du rôle de la cellule, déjà envisagé par Buffon dans son *Histoire naturelle générale et particulière*.

Les prémices de la génétique remontent au XIXe avec, en 1859, la publication par Darwin de sa théorie sur l'évolution avec son texte *l'Origine des espèces*, et en 1865 celle des lois de Mendel. Darwin suggère que toutes les espèces vivantes sont en perpétuelle transformation et que l'évolution permet l'apparition de nouvelles espèces ou la disparition d'autres par le biais de la sélection naturelle. De son côté, le moine Grégor Mendel réalise des croisements de lignées homozygotes de pois et observe la formation des hybrides. Il démontre que les facteurs héréditaires existent par paires et il introduit les concepts majeurs de dominance et de récessivité.

En 1869, Johann Friedrich Miescher découvre dans le noyau des cellules une substance riche en phosphate : la nucléine renommée ensuite en acide désoxyribonucléique (ADN).

En 1902, Walter Sutton et Theodor Boveri développent chacun de leur côté la théorie chromosomique qui identifie les chromosomes comme étant les porteurs de l'information génétique.

Thomas Morgan teste ensuite les lois de Mendel sur la mouche drosophile et s'intéresse à l'hérédité. En 1911, il découvre que les chromosomes sont le support des gènes[2] et construit avec Alfred Sturtevant les premières cartes de localisation des gènes sur les chromosomes, les cartes génétiques[3]. L'unité de fréquence de recombinaison utilisée en cartographie génétique a été baptisée en son honneur le centimorgan.

Après la deuxième Guerre Mondiale, les découvertes s'accélèrent. En 1942, Conrad Waddington propose le terme d'épigénétique, en lien avec l'épigenèse, pour étudier les mécanismes de spécialisation des cellules au cours du développement. Ce généticien désigne l'épigénétique comme le lien entre les caractères observables (phénotypes) et l'ensemble des gènes (génotypes). Oswald Avery en 1944 démontre que l'ADN est le support biochimique de l'hérédité grâce à ses travaux sur la transformation bactérienne chez les pneumocoques[4]. En 1953, Watson et Crick publient la structure en double hélice de l'ADN[5], s'inspirant des rapports non publiés de Rosalind Franklin qui fut une pionnière de la biologie moléculaire. Rosalind Franklin, grâce à ses travaux sur la diffraction aux rayons X, réussit à distinguer la forme B de la forme A de l'ADN. Des expériences de transplantation nucléaire initiées par Briggs et King [6] en 1952, puis par Gurdon[7] en 1962, apportent la preuve expérimentale que toute l'information génétique présente dans le zygote se retrouve dans toutes les cellules différenciées de l'organisme. Les différences entre les phénotypes cellulaires ne peuvent s'expliquer que par des différences dans l'expression des gènes suivant les types cellulaires, impliquant des mécanismes de régulation. En 1961, François Jacob, Jacques Monod et André Lwoff découvrent le fonctionnement des gènes[8]. Ceux-ci ne sont pas exprimés de manière constante au fil du temps, mais ils sont régulés pour répondre aux besoins de notre organisme. Le code génétique est finalement décrypté en 1963 par Nirenberg[9], Khorana[10] et Ochoa[11]. La table de correspondance entre les codons potentiels et les 20 acides aminés universellement répandus dans le monde vivant est établie.

Par la suite, différentes techniques émergent avec pour but de déchiffrer l'enchaînement exact des nucléotides composant une séquence d'ADN. Le premier génome, génome du phage φX174, fut ainsi séquencé en 1977[12].

Histoire de la génétique

-460-377 AV JC

Hippocrate émet l'hypothèse que l'embryon se forme par le mélange de la semence du père et de la mère et que ces semences proviennent de toutes les parties du corps

-427-346 AV JC

Aristote affirme que la semence paternelle est une substance génératrice de forme alors que la semence maternelle est juste nourricière.

1751-1772

Encyclopédie de Diderot et d'Alembert. Ils soutiennent la théorie préformiste..

1859

Darwin publie la théorie de l'Evolution.

1865

Lois de Mendel.

1869

Miescher découvre la nucléine.

1902

Walter Sutton et Theodor Boveri identifient les chromosomes comme étant les porteurs de l'information génétique.

1911

Thomas Morgan découvre que les chromosomes sont le support des gènes et construit avec Alfred Sturtevant les premières cartes génétiques.

1942

Conrad Waddington propose le terme d'épigénétique, en lien avec l'épigenèse.

1944

Oswald Avery démontre que l'ADN est le support biochimique de l'hérédité.

1953

Watson et Crick publient la structure en double hélice de l'ADN, s'inspirant des rapports non publiés de Rosalind Franklin.

1953

Frédérick Sanger publie la séquence de l'insuline.

1961

François Jacob, Jacques Monod et André Lwoff découvrent le fonctionnement des gènes.

1963

Le code génétique est décrypté par Nirenberg, Khorana et Ochoa.

1975

Sanger et Coulson développent la méthode "plus and minus".

1977

le génome du phage φX174 est décrypté par la méthode Sanger.

1979

Staden propose la méthode du "shotgun sequencing".

1990

Lancement du Human Genome Project.

2000

Publication du génome d'Arabidopsis.

2001

Publication de la première version du génome humain.

2022

Le consortium Telomere to Telomere publie une version complète du génome humain.

Figure 1 : **Histoire de la génétique**

9.1.2 Le séquençage

9.1.2.1 Les débuts du séquençage et la méthode Sanger

La première molécule séquencée fut une protéine, l'insuline, en 1953 par l'équipe de Frédérick Sanger grâce à une méthode de chromatographie[13,14]. Les protéines étaient aléatoirement fragmentées et les fragments étaient lus individuellement. Les séquences obtenues étaient chevauchantes, permettant la reconstitution d'une séquence dite "consensus". En 1965, les 76 bases de l'Acide Ribonucléique (ARN) de transfert de l'Alanine de *Saccharomyces cerevisiae* furent séquencées[15]. L'ARN fut fragmenté grâce à des RNAses, les fragments furent séparés par chromatographie et leur séquence déduite de leur produits de dégradation. La première molécule d'ADN séquencée fut celle des extrémités complémentaires de 12 bases de long du cos-site du phage Lambda. Entre 1972 et 1976, différentes équipes publient des séquences de gènes de phages obtenues à partir du séquençage de leur ARN[16-19].

En 1975, Sanger et Coulson développèrent la méthode "plus and minus" et l'appliquèrent à de courtes régions du phage φX174[12]. Un fragment d'ADN, un primer, de l'ADN polymérase et les quatre désoxynucléotides (dont un est marqué au ^{32}P) sont placés dans quatre réactions indépendantes. Des fragments de taille différente sont produits, tous commençant par le primer. Chaque réaction est ensuite séparée en deux, une moitié engagée dans la réaction "minus" et une moitié engagée dans la réaction "plus". Dans la réaction "minus", l'ADN polymérase et trois désoxyribonucléotide triphosphate (dNTP) permettent d'allonger le fragment jusqu'au prochain nucléotide manquant. Dans la réaction "plus", seulement un nucléotide et la T4 polymérase sont ajoutés. L'activité exonucléase de la T4 polymérase permet de dégrader le fragment depuis l'extrémité 3' jusqu'au nucléotide ajouté à la réaction. Le produit des 8 réactions est déposé sur gel de polyacrylamide. Les photos des gels sont obtenues grâce aux rayons X et l'ordre de migration des fragments détermine l'ordre des nucléotides (Figure 2).

En 1977, le génome du phage φX174 est décrypté par F. Sanger grâce à cette méthode. C'est le premier génome ADN jamais séquencé. La

même année, Sanger met au point une nouvelle méthode qui portera son nom mais aussi appelée "chain termination sequencing[20]. Elle se base sur l'extension d'un fragment à partir d'une amorce et d'un fragment d'ADN modèle. Dans quatre réactions indépendantes, on ajoute trois dNTP et un 2,3-dideoxynucleoside triphosphate (ddNTP). Quand celui-ci est incorporé, l'élongation s'arrête par manque du groupe 3'-hydroxyl. Les fragments sont déposés sur gel et la séquence est lue suivant la taille de ceux-ci (Figure 3). Parallèlement, Maxam et Gilbert proposèrent une méthode de séquençage purement chimique qui ne perdura pas à cause de la complexité technique et de l'utilisation de produits chimiques dangereux[21].

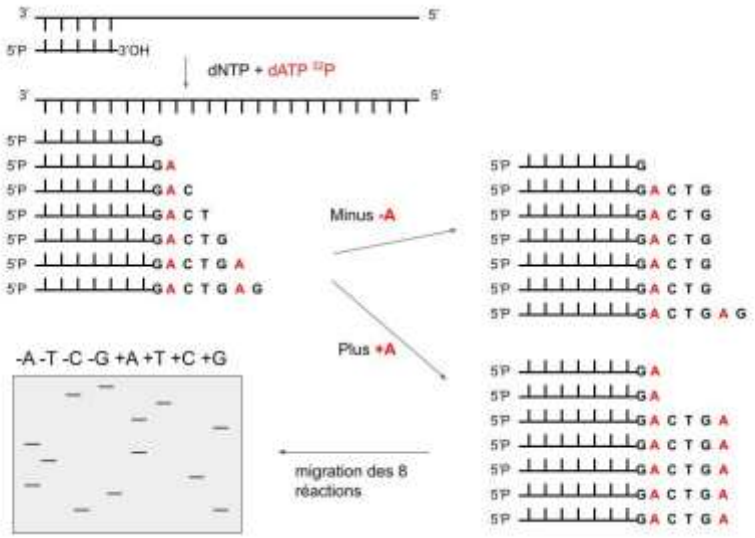


Figure 2 : ***Réaction "Plus and Minus"****. Un fragment d'ADN est amplifié en présence de 3 désoxynucléotides et d'un désoxynucléotide marqué au ^{32}P (4 réactions différentes, dans l'exemple ddATP marqué au ^{32}P). Les fragments obtenus sont de tailles différentes. Chaque réaction est ensuite séparée en deux : la moitié est engagée dans la réaction "minus" et la moitié dans la réaction "plus". Dans la réaction "minus", seulement trois désoxynucléotides sont ajoutés (ici pas de dATP) et les fragments sont allongés jusqu'au nucléotide manquant. Dans la réaction "plus", l'activité exonucléase de la T4 Polymérase digère les fragments jusqu'au nucléotide ajouté (ici dATP). Les huit réactions sont finalement déposées sur gel de polyacrylamide et la séquence est lue du bas du gel vers le haut.*

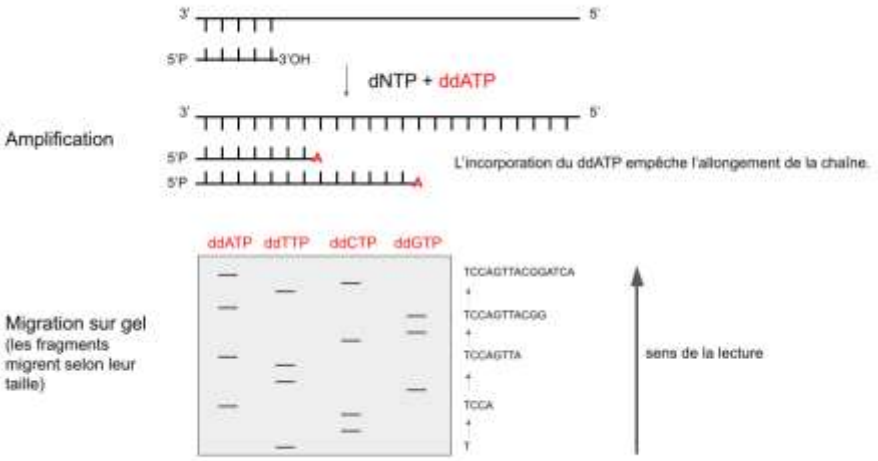


Figure 3 : ***Séquençage par méthode de Sanger****. Un fragment d'ADN génomique est amplifié en présence de désoxynucléotides et d'un didésoxynucléotide (4 réactions différentes, dans l'exemple ddATP). La réaction d'élongation s'arrête si le didésoxynucléotide est incorporé. Statistiquement, on retrouvera des fragments arrêtés après chaque position. Les produits des 4 réactions sont déposés sur gel de polyacrylamide. La séquence peut ainsi être lue du bas du gel vers le haut.*

9.1.1.1 Les premières générations de séquenceurs

Des industriels ont profité d'avancées en physique, en informatique et en nanotechnologies pour développer des instruments de première génération permettant de faire migrer les fragments d'ADN obtenus par méthode Sanger[22]. La méthode de migration a évolué de gels de polyacrylamide coulés entre deux plaques de verre à des gels enfermés dans des capillaires. Les ddNTP portaient un fluorochrome spécifique. Celui-ci était excité par un laser et la détection de la fluorescence émise était réalisée par des caméras. Ces instruments permettaient de lire des fragments de maximum 1000 bases. Cela marqua le début de l'automatisation et de l'augmentation du débit. La société Applied Biosystems a commercialisé les séquenceurs 3730 qui ont équipé tous les grands centres de séquençage[23]. Ils permettaient de séquencer en

parallèle 96 fragments d'ADN. L'excitation des fluorochromes par le laser conduisait à l'émission de fluorescence captée par une caméra et à la production d'un chromatogramme (Figure 4). Chaque couleur représentait une base.

Figure 4 : ***Séquenceur ABI3730*** *: a) photo du séquenceur b) capillaires (en jaune) placés dans l'instrument et à l'intérieur desquels migrent les fragments d'ADN c) chromatogramme obtenu après séquençage*

Deux méthodes furent employées pour séquencer des génomes entiers à partir des années 80/90. La première est une méthode dite "BAC à BAC" et la seconde "shotgun sequencing", proposée par Staden en 1979[24]. Pour la première méthode, il convient de fragmenter chaque chromosome en fragments d'environ 150000 bases et d'en déterminer l'ordre grâce à l'élaboration d'une carte physique. Chaque fragment est ensuite inséré dans un BAC (Bacterial Artificial Chromosome). Celui-ci est ensuite fragmenté aléatoirement et séquencé. La séquence de chaque BAC est ensuite reconstituée. Connaissant l'ordonnancement de chaque BAC, la séquence du chromosome est ensuite déduite. Entre les contigs, des bases indéterminées (N) sont placées. Pour résoudre ces séquences inconnues, une PCR est réalisée en utilisant des primers ancrés de part et d'autre de cette séquence sur les extrémités des contigs. Le produit de PCR est ensuite lui-même séquencé. Cette méthode est très fiable mais très coûteuse en temps comme en réactifs. Elle a été utilisée en 1981 pour séquencer le génome du virus de la mosaïque du chou-fleur[25]. La deuxième méthode s'affranchit de l'élaboration de la carte physique. Le génome est aléatoirement fragmenté. Des vecteurs bactériens de différentes tailles (plasmides, YACs, BACs) sont utilisés pour cloner de longs fragments d'ADN et constituer une banque génomique. Les fragments insérés sont séquencés et leurs séquences

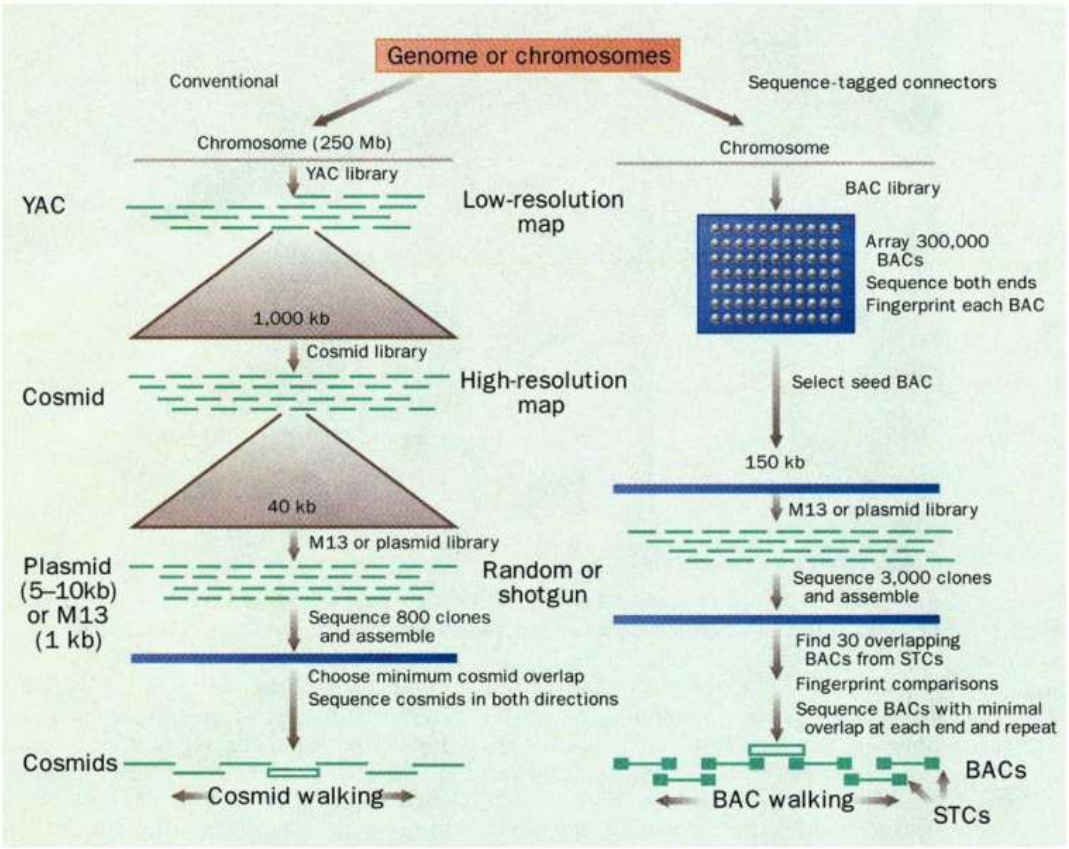
sont assemblées grâce à leur chevauchement pour reconstituer des séquences plus longues appelées contigs. Cette méthode est plus rapide et moins coûteuse mais l'absence de carte physique rend plus difficile la résolution des chromosomes.

9.1.1.2 Le Human Genome Project

Le Human Genome Project fut lancé en 1990 avec pour but de décoder l'intégralité du génome humain avant 2005. Le consortium était composé de 20 laboratoires américains, anglais, japonais, français, allemands et chinois. Les financements provenaient des gouvernements des pays impliqués et de donateurs. L'ADN a été cloné dans des BACs et des Yeast Artificial Chromosomes (YACs) avant séquençage. Le coût du séquençage a été énorme, environ 2.7 milliards de dollars mais le génome fut publié une première fois en 2001 avec un taux de complétion d'environ 92%[26]. Le projet a été déclaré terminé en avril 2003 et une nouvelle version plus complète du génome fut publiée en 2004[27]. En France, le Genoscope (Centre National de Séquençage) a participé au Human Genome Project en réalisant le séquençage du chromosome 14[28].

Parallèlement, Craig Venter réalisa le séquençage de son propre génome[29,30] par une méthode dérivée du shotgun, appelée STC pour Sequence-tagged connectors[31]. Pour limiter le coût et le temps d'exécution, les BACs étaient sélectionnés pour n'en séquencer qu'un nombre minimal (Figure 5).

Il a été estimé que les génomes obtenus n'étaient pas complets et que les régions manquantes étaient localisées dans les régions hétérochromatiques connues pour être hautement condensées et donc difficilement atteignables. Par la suite, de nouvelles versions ont été publiées mais il faudra attendre 2022 pour que le consortium international Telomere To Telomere réalise une version quasi parfaite du génome humain en combinant toutes les technologies modernes[32] (voir Discussion). Cette version aura certes demandé bien moins de fonds que la première mais le coût engagé pour atteindre ce niveau de complétion reste très élevé.



The conventional sequencing approach and the newly proposed sequence-tagged connectors (STC) approach. The bacterial artificial chromosome (BAC) clones in the STC approach could be sequenced by any cost-effective strategy.

Figure 5 : ***Méthode STC employée par Craig Venter pour le séquençage du génome humain.***

9.1.1.3 Les Nouvelles Générations de Séquenceurs

En 2005, apparurent les Nouvelles Générations de Séquençage ou NGS (Figure 6). Elles répondaient à des besoins de baisse des coûts et d'augmentation des débits. Ce sont des technologies basées sur la miniaturisation, la baisse des volumes de réactions et le séquençage en temps réel grâce à l'activité de l'ADN polymérase. Plusieurs compagnies se sont lancées dans cette course mais les deux technologies majoritairement employées furent celle de Roche 454 (Genome Sequencer) et celle d'Illumina (Genome Analyzer). Les séquenceurs commercialisés par Thermo Fisher (Ion Torrent) et par

Applied Biosystems (Solid, Sequencing by Oligonucleotide Ligation and Detection) ne rencontrèrent que peu de succès.

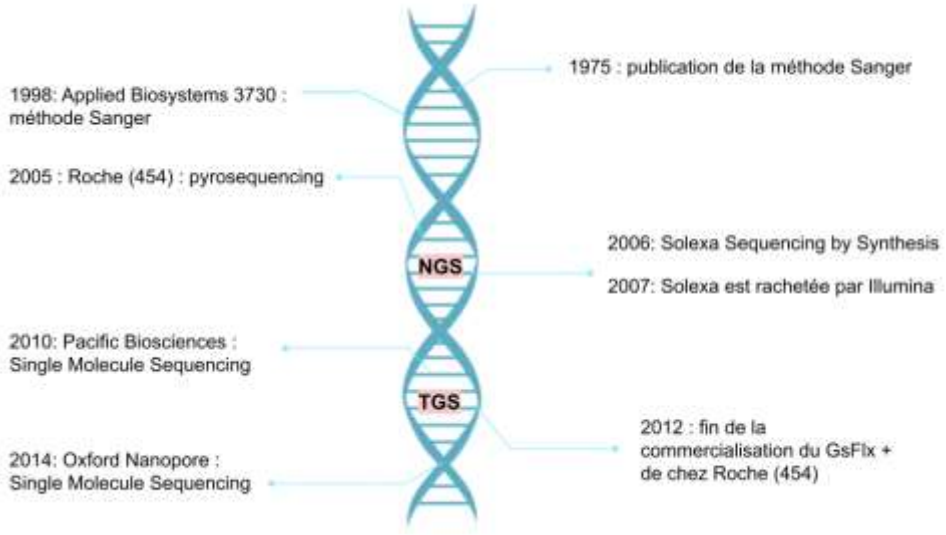


Figure 6 : ***frise chronologique avec les dates d'apparition des technologies de séquençage*** *(NGS : Nouvelles Générations de Séquençage, TGS : Troisième Génération de Séquençage).*

Pour la technologie Roche, la méthode de séquençage est nommée Pyroséquençage. L'ADN est fragmenté par nébulisation (pression d'azote appliquée sur l'ADN génomique). Les extrémités des fragments obtenus sont réparées et les adaptateurs spécifiques à la technologie sont ligués à chaque extrémité. Les fragments sont ensuite amplifiés par PCR en émulsion (emPCR). En effet, un fragment d'ADN est enfermé dans une gouttelette d'huile avec tous les réactifs nécessaires ainsi qu'une bille porteuse d'un oligonucléotide complémentaire à un des adaptateurs ligués. Le fragment d'ADN est ainsi amplifié et les fragments néo formés sont accrochés à la bille. Les gouttelettes sont ensuite rompues (cassure des émulsions) et les billes sont récupérées. Le support de séquençage ou Picotiter Plate (PTP) est une plaque comportant à sa surface des micro puits. Dans chaque puits est déposé une bille portant le fragment d'ADN amplifié, les réactifs, et des billes permettant de maintenir l'ensemble dans le puits. La plaque est ensuite placée dans le séquenceur. A chaque cycle d'amplification, un seul des quatre nucléotides est apporté. Si un nucléotide est incorporé, un pyrophosphate est libéré. Il entre dans

une suite de réactions conduisant à une émission de lumière. Il est possible d'incorporer plusieurs nucléotides à la fois. L'intensité de la lumière émise, capturée par une caméra à chaque cycle, est proportionnelle au nombre de nucléotides incorporés (Figure 7). Dans les régions contenant des homopolymères, la sensibilité de détection peut être dépassée, ce qui conduit à une mauvaise estimation du nombre de bases. Le dernier instrument, le GsFlex+ permettait de produire des séquences de 700 bases en moyenne avec un débit de 700Mb par jour. Mais les coûts élevés (dûs à la PCR en émulsion) et le débit limité ont conduit Roche à arrêter la commercialisation des réactifs en 2012.

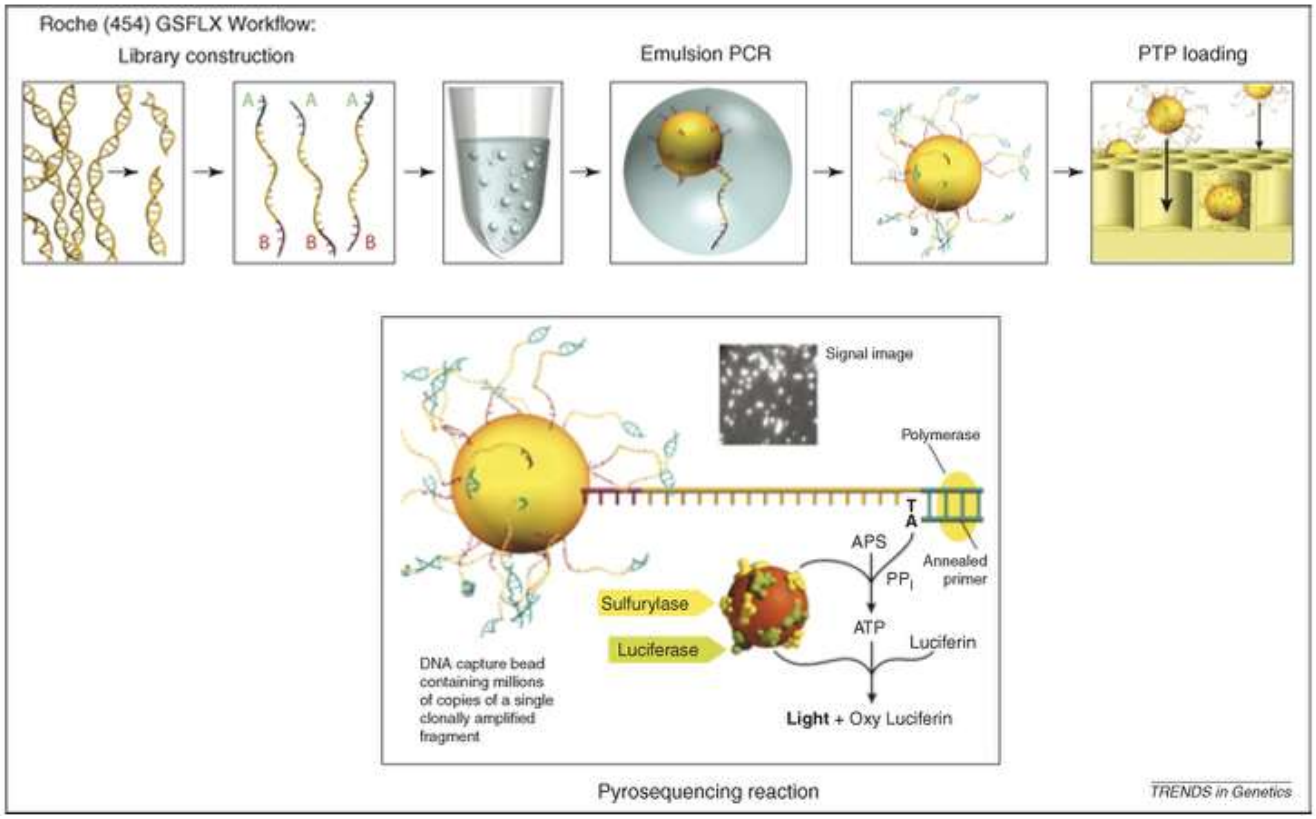
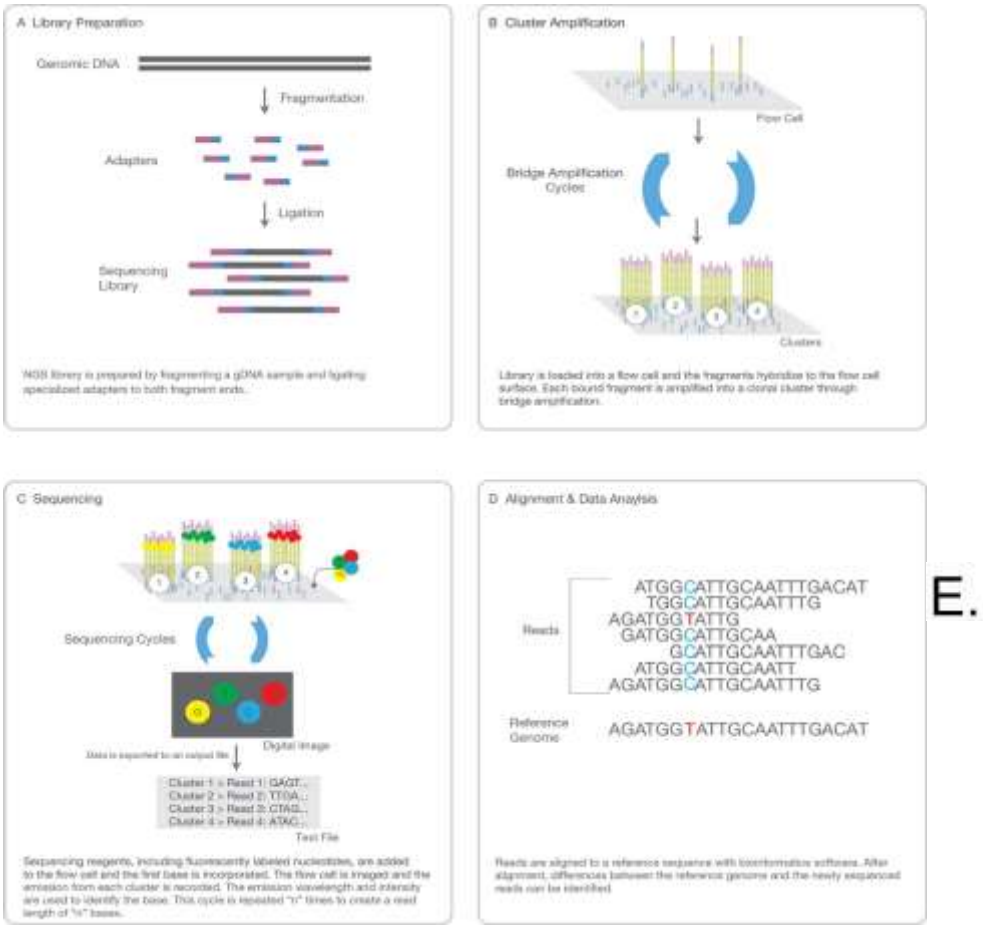


Figure 7 : ***Réaction de pyroséquençage et instrument Roche 454 GSFLX[33].***

La société Solexa commercialisa son premier instrument, le Genome Analyzer, en 2006[34]. Elle fut ensuite rachetée par Illumina en 2007. La méthode se nomme Séquençage par Synthèse (Sequencing by Synthesis ou SBS) (Figure 8). L'ADN génomique est fragmenté aléatoirement par sonication. Les extrémités des fragments sont réparés. Des adaptateurs spécifiques sont ligués en 5' et 3'. Les fragments sont amplifiés par PCR puis dénaturés pour constituer une banque de fragments simple brin. La banque simple-brin est déposée sur le support de séquençage (Flow Cell). La Flow Cell est constituée de plusieurs pistes enfermées entre deux lames de verre. Sur chaque piste, des adaptateurs complémentaires de ceux ligués aux fragments sont attachés soit directement à la lame de verre pour les premières générations d'instruments soit à la surface de micro puits pour les dernières générations. Les fragments d'ADN sont directement amplifiés sur la lame de verre (PCR par bridge) formant des clusters monoclonaux distincts, étape nécessaire pour obtenir un signal

d’intensité suffisante. La réaction de séquençage est amorcée en ajoutant les quatre nucléotides, chacun étant marqué par un fluorochrome à son extrémité 3’ qui empêche la réaction d’élongation. Un seul et unique nucléotide peut donc être incorporé à la fois. Un laser vient exciter le fluorochrome. Un signal lumineux est émis et capté par une caméra. Le fluorochrome est ensuite clivé et les nucléotides sont à nouveau ajoutés. Ces cycles sont répétés jusqu’à 150 fois pour les instruments à très haut débit (HiSeq puis NovaSeq) et 300 fois pour les instruments à faible débit (MiSeq). L’avantage de la technologie Illumina est principalement son débit qui permet de diminuer grandement les coûts. Elle élimine théoriquement les erreurs dans les homopolymères. Elle permet également de réaliser un séquençage des deux extrémités du fragment d’ADN (séquençage Paired End). Par contre, la taille des séquences obtenues est beaucoup plus petite qu’avec les séquenceurs Roche. Le NovaSeq 6000 permet aujourd’hui de générer 6 Tbases en moins de deux jours. Depuis le HiSeq 4000, Illumina a mis en place des Flow Cells ordonnées. Les pistes contiennent des micro puits à l'intérieur desquels sont placés les adaptateurs. Ceci a permis d’augmenter le nombre de clusters possibles par piste. En effet, le débit était limité par le nombre de clusters présents sur les lames de verre. Si les clusters étaient trop proches, ils étaient éliminés car leurs signaux pouvaient interférer entre eux. En séparant les clusters physiquement, les signaux détectés sont

donc bien spécifiques d'un seul et unique cluster.



E.

Figure 8 : ***Sequencing by Synthesis (Illumina)***[35]*. La préparation de la banque (A) débute par la fragmentation de l'ADN génomique, la réparation et l'ajout d'adaptateurs spécifiques aux extrémités des fragments. Ceux-ci sont ensuite amplifiés de façon monoclonale directement sur le support de séquençage (B) afin de former des clusters. Lors du séquençage, un seul nucléotide, porteur d'un fluorochrome, peut être incorporé. Le fluorochrome est excité par un laser et la fluorescence émise est captée par la caméra (C). Celle-ci prend des images de la flow cell après chaque cycle d'incorporation d'un nucléotide. (E) Image d'un séquenceur NovaSeq 6000.*

Grâce à l'avènement des nouvelles technologies de séquençage, il y a eu une augmentation considérable du nombre de génomes produits. Le coût et la rapidité d'exécution ont permis une démocratisation de

la demande. La société Illumina a amené le coût du séquençage d'un génome humain à 1000$ (Figure 9).

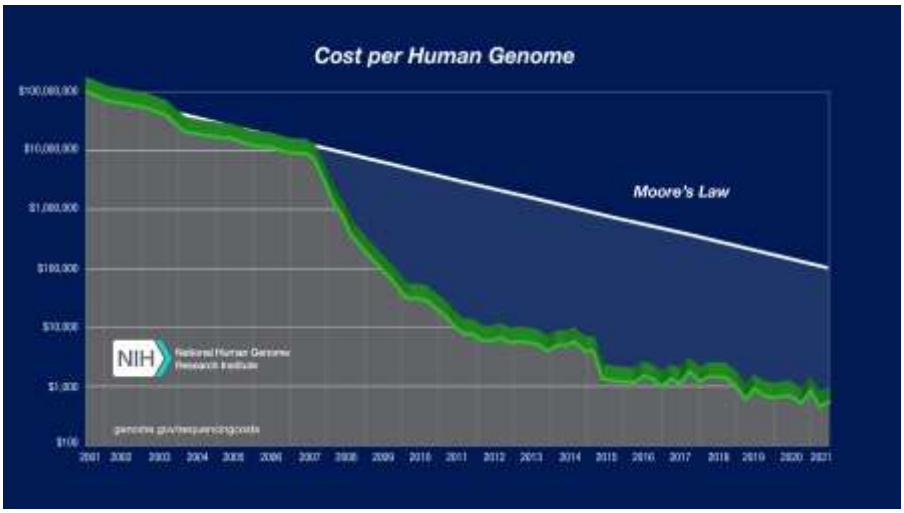


Figure 9 : ***Coût du séquençage d'un génome humain en dollars***[36]. *Avant 2007, le séquençage était réalisé selon la technologie Sanger sur des séquenceurs à capillaires. Après 2007, le séquençage est réalisé grâce aux séquenceurs NGS. La loi de Moore décrit une tendance dans le secteur du matériel informatique au doublement de la "puissance de calcul" tous les deux ans.*

9.1.2.5 Les Séquenceurs de Troisième Génération

Le séquençage longues lectures par la troisième génération de séquenceurs (Third Generation Sequencing ou TGS) répond à des difficultés posées par le séquençage courtes lectures de type Illumina. Même si ces données sont de très bonne qualité, la reconstitution des génomes est difficile, particulièrement dans des régions contenant des répétitions. Ainsi, en 2011, Pacific Biosciences (PacBio) lance sa technologie Single Molecule Real Time Sequencing (SMRT) basée sur une enzyme ADN Polymérase attachée au fond de micro-puits appelés zero-mode waveguides (ZMWs)[37]. Ils sont capables de détecter une excitation dans un volume infiniment petit, permettant le séquençage d'une seule molécule à la fois. La synthèse du brin complémentaire grâce à des nucléotides marqués est suivie en temps réel par une caméra. Il n'y a plus d'arrêt de l'élongation après chaque base. Le taux

d'erreur était élevé (environ 10%) au début de la mise en place de la technologie. Cependant, les erreurs sont réparties aléatoirement, donc si une région est séquencée plusieurs fois, les erreurs sont corrigées et la séquence consensus résultante ne présente quasiment plus d'erreurs. Récemment, PacBio a proposé la méthode HiFi pour High Fidelity mettant en avant le système CCS (pour Circular Consensus Sequence)[38,39]. Celui-ci permet de séquencer plusieurs fois le même fragment d'ADN puis de déterminer une séquence consensus. Pour cela, des adaptateurs hairpin circulaires sont ajoutés aux extrémités du fragment d'ADN (figure 10) permettant à la polymérase de lire plusieurs fois le même fragment. Cette méthode permet d'obtenir des lectures d'environ 15 Kb avec une fidélité d'environ 99,9%. La dernière génération de séquenceur PacBio, le Sequel II, permet de produire jusqu'à 4 millions de lectures HiFi en environ 30 heures.

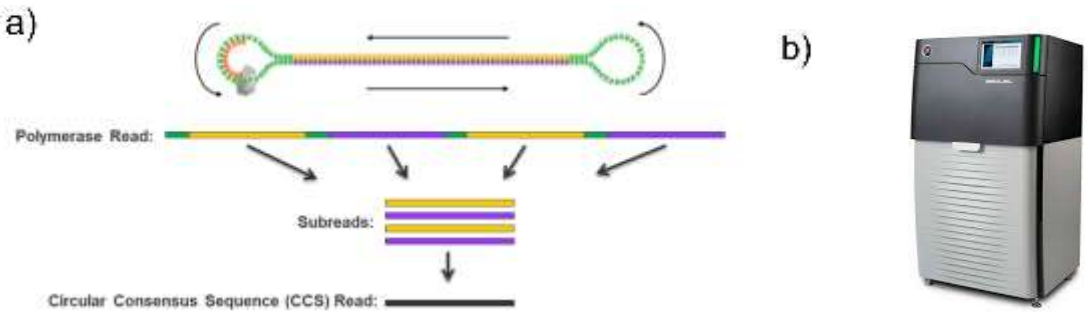


Figure 10 : ***Méthode CCS et séquenceur Sequel II (PacBio)[40]****. a) Des adaptateurs en forme de boucle sont accrochés aux extrémités des fragments d'ADN. La polymérase va ainsi synthétiser un nouveau fragment en tournant autour du fragment circularisé. Si la polymérase réalise plusieurs boucles, le fragment contiendra une concaténation de plusieurs fragments synthétisés. Chaque sous-unité est ensuite extraite et une séquence consensus est produite. Ce mécanisme permet de corriger les erreurs de séquençage non systématiques grâce à la couverture en lecture. b) Image de l'instrument Sequel 2.*

En 2014, Oxford Nanopore lance sa technologie qui n'est pas basée sur la synthèse du brin complémentaire. Le premier séquenceur porte le nom de MinION. Il permet de séquencer des fragments d'ADN très longs. Le support de séquençage est composé d'une membrane de phospholipides à l'intérieur de laquelle des pores protéiques sont insérés. Des adaptateurs spécifiques sont ligués aux extrémités du fragment d'ADN. A l'un des adaptateurs est accrochée une protéine motrice dont le rôle est d'accompagner le fragment d'ADN double brin jusqu'à l'entrée du pore. Elle permet d'ouvrir la double hélice et de ralentir le passage d'un brin à l'intérieur du pore. La présence de nucléotides à l'intérieur de la chambre de lecture perturbe le courant électrique qui parcourt la membrane. La perturbation est spécifique de la séquence de ces nucléotides. La perturbation de signal est capturée par des senseurs et convertie en séquence. Il ne semble pas y avoir de limites à la taille des fragments d'ADN séquencés. Le principal problème est de trouver la bonne méthode d'extraction permettant d'obtenir des molécules les plus longues et les plus intègres possibles. Au début de la commercialisation, le taux d'erreur était très élevé d'environ 30%, avec des erreurs systématiques dans les homopolymères. Des efforts, entre autres, sur le développement des pores et sur les logiciels de "basecalling" (qui permettent d'obtenir la séquence en base à partir du signal électrique) ont permis d'augmenter la fidélité à 99,3% (chimie Q20+[41]) se rapprochant de la qualité des lectures HiFi. Oxford Nanopore a également développé une large gamme de séquenceurs[42] et notamment une version haut débit, le PromethION qui peut générer en théorie jusqu'à 14Tb en 72 heures (avec 48 Flow Cells en parallèle) (Figure 11).

Le séquençage de troisième génération, comme celui réalisé avec la technologie Oxford Nanopore, permet en outre de séquencer des ARN et de détecter les différentes isoformes et variants d'épissage. Comme il n'y a pas besoin d'amplifier l'ADN avant séquençage, on peut détecter les bases méthylées sans passer par une préparation de banque spécifique. La figure 12 illustre les diverses applications de ces méthodes.

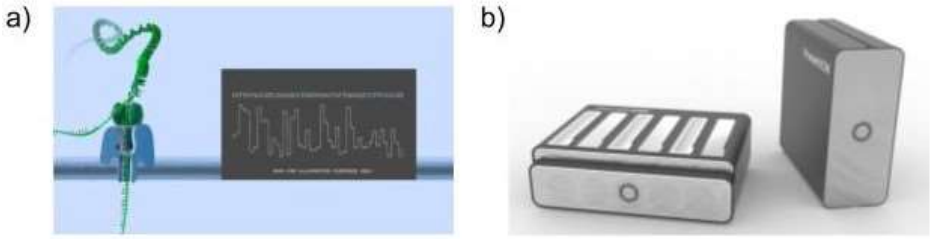


Figure 11 : ***Séquençage Oxford Nanopore[42].*** *a) La membrane de la Flow Cell contient des pores protéiques. Le fragment d'ADN est guidé par sa protéine motrice à l'entrée du pore. Cette protéine a un rôle d'hélicase, permettant au fragment de se débobiner et à un brin de pénétrer à l'intérieur du pore. Elle contrôle également la vitesse de passage dans le pore. La perturbation du courant électrique parcourant la membrane provoquée par le passage des nucléotides dans le pore est mesurée par des capteurs situés sous le pore. b) Séquenceur PromethION, modèle ultra haut débit de la gamme commercialisée par Oxford Nanopore. Il comprend 24 ou 48 emplacements pour Flow Cell.*

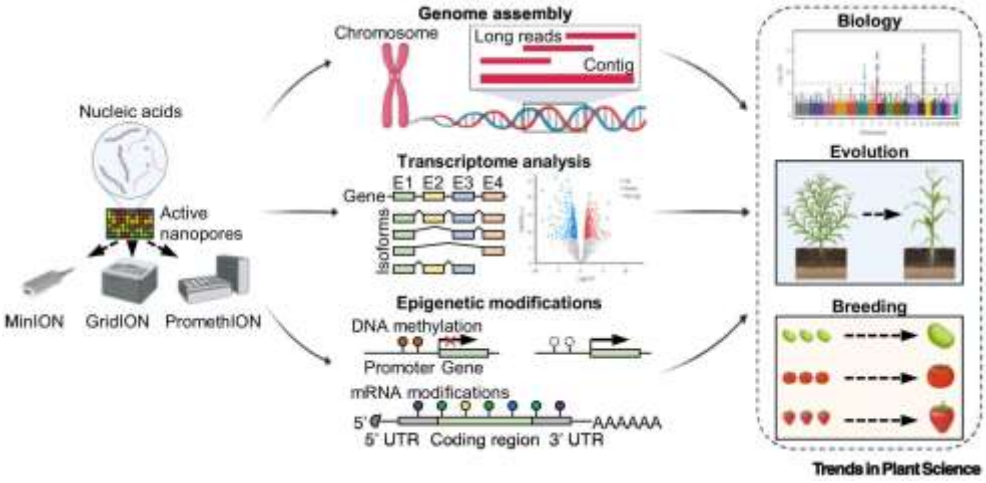


Figure 12 : ***Applications du séquençage Oxford Nanopore[43].*** *Les longues lectures obtenues grâce aux instruments Oxford Nanopore (ONT) permettent d'assembler les génomes végétaux même les plus complexes, mais aussi de caractériser les isoformes de trancrits et les niveaux d'expression de l'ARN. De plus, l'ONT peut être utilisé pour détecter les modifications épigénétiques de l'ADN et de l'ARN natifs. Les plateformes et les algorithmes ONT se développent rapidement, permettant une compréhension plus approfondie de la biologie et de l'évolution des plantes et guidant la sélection végétale.*

9.1.2 L'assemblage

Après le séquençage d'un génome, il convient de réaliser son assemblage. Assembler un génome *de novo* consiste à reconstituer sa séquence sans l'aide d'une référence et sans *à priori*. Pour cela, de nombreux algorithmes et outils sont disponibles.

9.1.2.1 Assemblage à partir de courtes lectures

Pour assembler un génome à partir des courtes lectures produites par les séquenceurs de deuxième génération, les programmes informatiques doivent résoudre trois grandes difficultés[44] : la première

est de placer chaque pièce d'un immense puzzle au bon endroit, la seconde est de pouvoir manipuler une très grande quantité de données et la troisième est de parvenir à manipuler des pièces qui présentent de très grandes similitudes voir qui sont identiques. Ils regroupent ainsi les lectures pour former une séquence plus continue appelée "contig" (Figure 13).

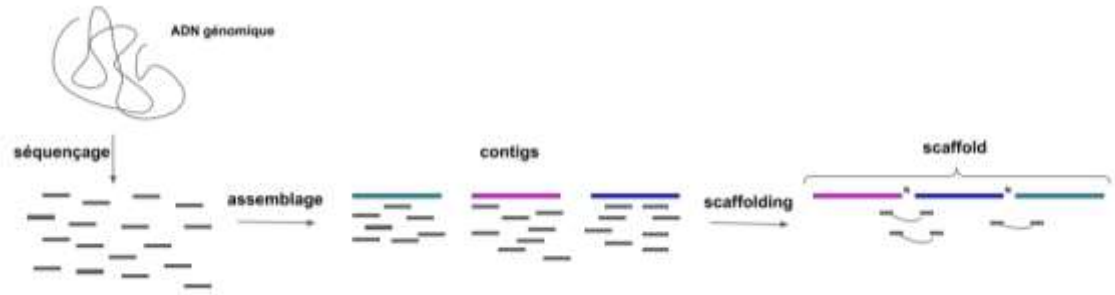


*Figure 13: **Assemblage de novo**. L'ADN génomique, préalablement extrait, est soumis à une préparation de banque de séquençage. Les lectures produites sont assemblées en contigs grâce à des outils bio-informatiques. Ces contigs sont ensuite orientés et ordonnés les uns par rapport aux autres lors du processus de scaffolding en utilisant des données de contacts distants.*

L'outil SPADES[45] principalement utilisé pour assembler les lectures produites par les séquenceurs Illumina utilise les graphes de De Bruijn[46] du nom du mathématicien néerlendais qui développa ce modèle en 1940. Il recherchait la plus courte chaîne circulaire de caractères contenant toutes les sous-chaînes possibles de même longueur k (ou k-mer), dans un alphabet donné. La solution qu'il a trouvée consistait à construire un graphe avec tous les (k - 1)-mer possibles comme nœuds. Chaque k-mer était une arête dirigée du nœud A vers le nœud B si le (k - 1)-mer du nœud A est un préfixe et celui du nœud B, un suffixe du k-mer. Il fallait ensuite trouver un chemin eulérien à travers le graphe, c'est-à-dire qui traverse chaque arête une seule fois. Sur l'exemple de la Figure 14, une séquence de 10 nucléotides a été générée avec des lectures de 6 nucléotides de long. Les lectures ont ensuite été décomposées en fragments plus petits de

taille k sur une fenêtre glissante de 1 base (dans l'exemple k est égal à 3). Un graphe de Bruijn avec les (k-1)-mers comme nœuds et les k-mers comme arêtes est construit. Un chemin eulérien est tracé à travers ce graphe, ce qui permet de reconstruire la séquence génomique originale.

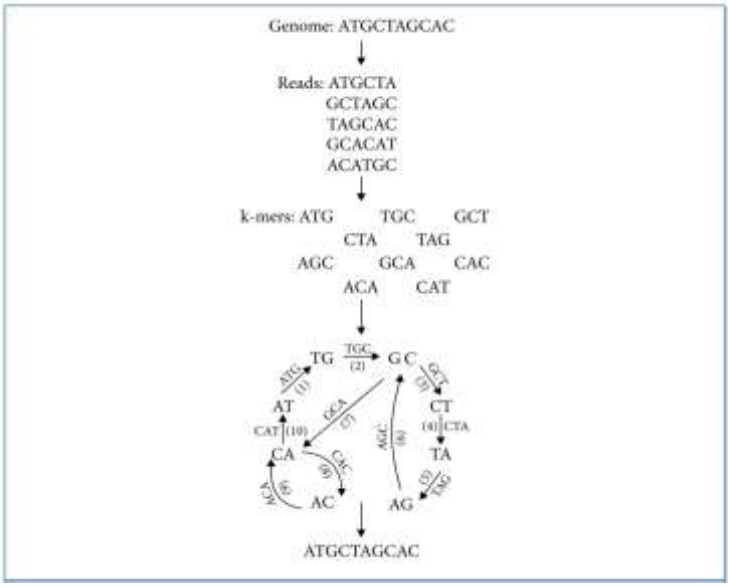


Figure 14 : ***Graphe de De Bruijn***[47]*. Le génome est découpé en fragments de petite taille. Les séquences de ces fragments sont appelées lectures ou "Reads" en anglais. Chaque lecture est ensuite découpée en mots ("k-mers"), ici de taille 3, en se décalant d'une base à la fois. Le read ATGCTA produit par exemple les k-mers ATG, TGC, GCT, CTA. Le graphe de De Bruijn est ensuite construit en reliant les k-mers qui ont des suffixes/préfix en commun. La lecture du graphe permet ainsi de reconstituer la séquence initiale.*

Les séquences répétitives, les variants, les données manquantes et les erreurs de séquençage limitent parfois l'efficacité et la précision de l'assemblage de génomes grâce à cette méthode. En effet, les séquences d'ADN répétées sont abondantes dans un large éventail d'espèces, des bactéries aux mammifères. Elles couvrent par exemple

près de la moitié du génome humain et 70% du génome du blé. Les assembleurs basés sur les graphes de De Bruijn ne sont pas capables de reconstruire fidèlement ces régions et sous-estiment le nombre de répétitions. Les répétitions créent des ambiguïtés dans le graphe, ce qui peut produire des biais et des erreurs lors de l'interprétation des résultats. Les répétitions peuvent être fusionnées, c'est-à-dire qu'une seule copie de la répétition sera présente dans la séquence reconstruite. L'assemblage du génome s'en trouvera alors fragmenté et incomplet (Figure 15). Il est également possible de créer des chimères en rapprochant deux régions chromosomiques qui ne sont pas proches l'une de l'autre[48]. De même, pour les génomes complexes, la quantité de ressources informatiques utilisées pour les calculs peut être un frein.

Ainsi, les contigs obtenus sont souvent très loin d'atteindre l'échelle des chromosomes. Pour organiser et orienter ces contigs entre eux et ainsi améliorer la continuité de l'assemblage, une étape de scaffolding peut être ajoutée (Figure 13). Pour cela, un séquençage est réalisé à partir de banques dites "Mate Pair"[49]. Il s'agit de séquencer les extrémités de grands fragments de taille connue (entre 3 et 20Kb). Le positionnement des séquences sur les contigs permet d'organiser les contigs entre eux, de connaître la distance qui les sépare (par déduction grâce à la distance attendue entre les positions des deux extrémités) et de les orienter. Les "gaps" ou trous entre deux contigs successifs sont symbolisés par des bases indéterminées (N).

Les assemblages obtenus ainsi, appelés "draft", même s'ils étaient fragmentés ou ne donnaient pas une image complète de la structure du génome, permettaient d'étudier le contenu en gènes d'un génome donné.

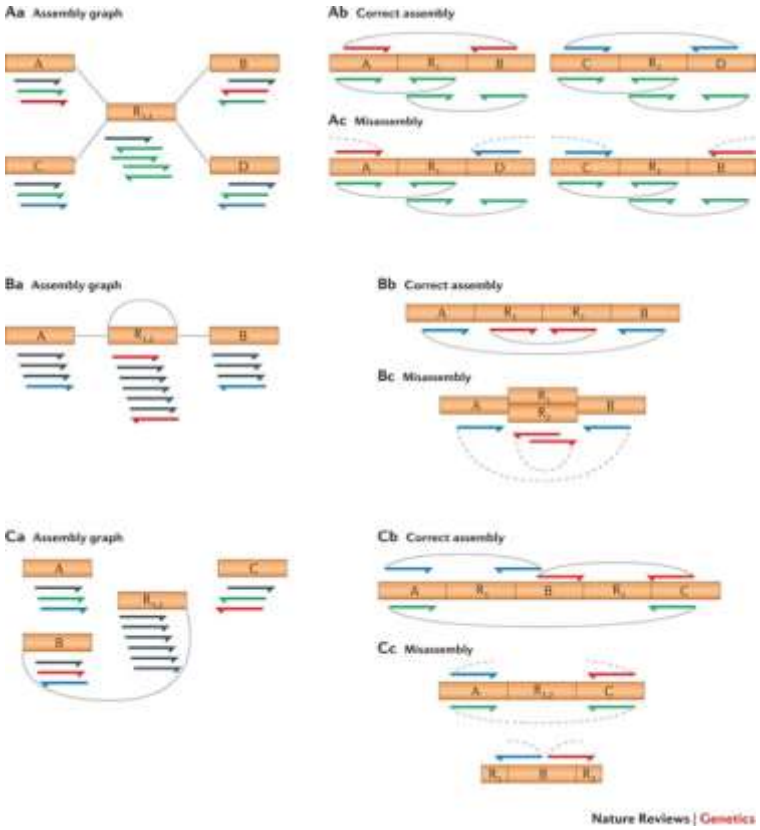


Figure 15 : ***Erreurs d'assemblage liées à la présence de répétitions***[48]*. A | Erreur d'assemblage par réarrangement causée par des répétitions. Aa | Un exemple de graphe d'assemblage impliquant six contigs, dont deux sont identiques (R1 et R2). Les flèches indiquées sous chaque contig représentent les lectures alignées. Ab | Assemblage correct en deux contigs. Ac | Deux contigs chimériques incorrectement assemblés, causés par les régions répétitives R1 et R2. Les lectures s'alignent parfaitement sur les contigs mal assemblés, mais l'orientation des paires n'est pas toujours respectée. B | Une répétition en tandem collapsée. Ba | Le graphe d'assemblage contient quatre contigs, où R1 et R2 sont des répétitions identiques. Bb | Assemblage correct, montrant un alignement correct des séquences Mate Pair avec la bonne distance. Bc | Un mauvais assemblage qui est causé par la fusion des répétitions R1 et R2. Les alignements de lecture restent cohérents, mais les distances entre les paires ne sont plus respectées. C | Une répétition collapsée. Ca | Le graphe d'assemblage contient cinq contigs, où R1 et R2 sont des répétitions identiques. Cb | Dans l'assemblage correct, R1 et R2 sont séparées par une séquence unique. Cc | Les deux copies de la répétition sont fusionnées. Le contig B est alors exclu de l'assemblage et apparaît comme un contig isolé avec des répétitions partielles sur son flanc.*

9.1.2.2 Assemblage à partir de longues lectures

L'apparition des technologies longues lectures a permis d'améliorer considérablement la qualité des assemblages produits. Ces longues lectures permettent en effet de couvrir de grandes répétitions, des régions polymorphes ou des transposons et de faciliter la reconstitution de ces régions jusque-là inaccessibles (Figure 16).

De nombreux logiciels d'assemblage ont été développés pour prendre en charge ces séquences de grande taille et pour remédier à leur problème de qualité. Il existe deux catégories d'outils : ceux qui nécessitent une étape de correction des lectures en amont de l'assemblage et ceux qui réalisent un assemblage à partir des lectures bruitées. Dans le deuxième cas, il est alors nécessaire de corriger le consensus produit grâce à des courtes lectures.

Pour la première catégorie d'assembleurs, on peut citer Canu[50], qui est capable d'assembler les lectures issues de séquençage avec les technologies PacBio ou nanopore. Il détecte les chevauchements entre les séquences, génère des séquences consensus corrigées, filtre ces lectures puis les assemble.

La deuxième catégorie a connu davantage de développements. On retrouve par exemple SMARTdenovo[51] qui utilise un pipeline d'assemblage de type Overlap Layout Consensus ou OLC (Figure 17) de longues lectures. Il a démontré sa capacité à produire des assemblages d'une continuité raisonnablement élevée à partir de lectures MinION et PacBio. On peut citer également Necat[52], Flye[53], ou wtdbg2[54].

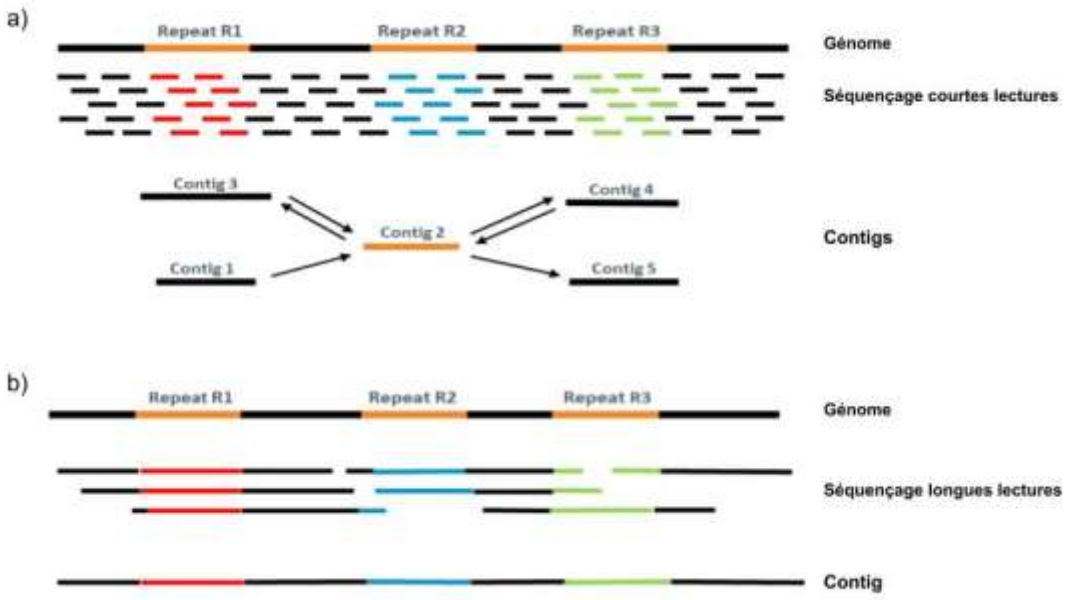


Figure 16 : ***Comparaison des assemblages obtenus à partir de courtes ou de longues lectures.*** *a) Séquençage en courtes lectures d'une région d'un génome contenant trois séquences répétées "Repeat R1/R2/R3". L'assemblage produit cinq contigs. Les régions répétées sont collapsées en un seul contig. b) Séquençage longues lectures de la même région. Les séquences obtenues peuvent traverser les régions répétées. Leur assemblage permet de reconstituer la région en un seul contig contenant les trois répétitions.*

Les contigs produits doivent être corrigés pour éliminer les erreurs restantes par des outils utilisant soit les longues lectures (Racon[55]) elles-même soit les courtes lectures (exemple : Pilon[56], Hapo-G[57])

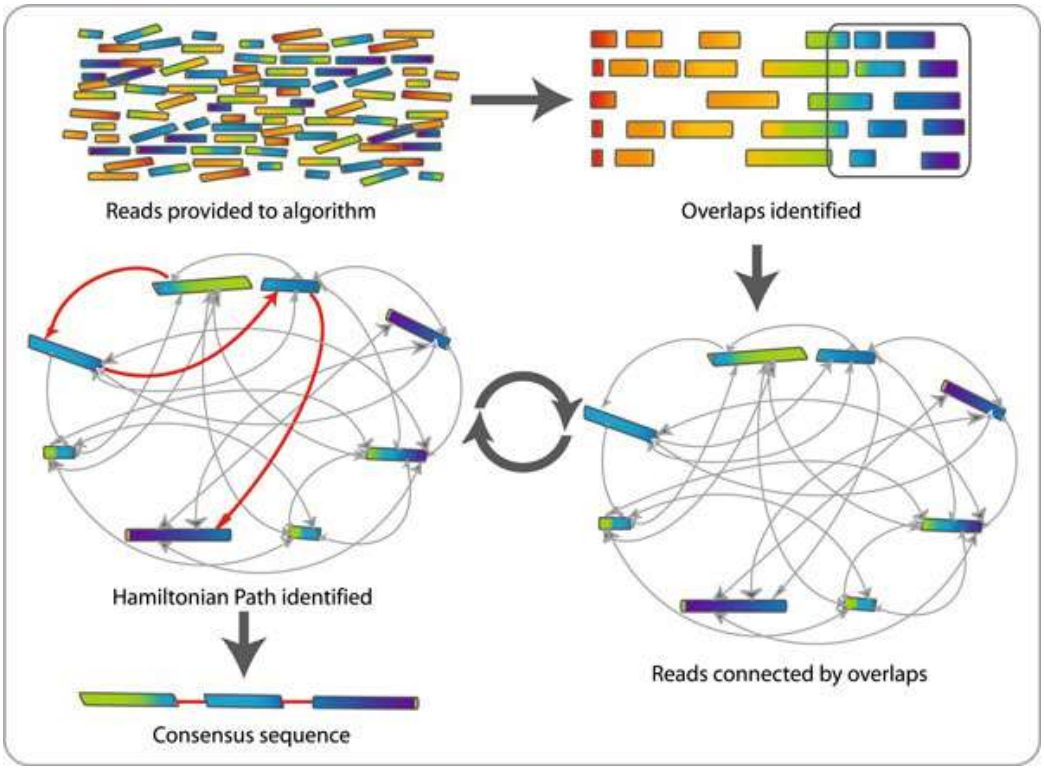


Figure 17 : ***Algorithme d'assemblage de génome Overlap–layout–consensus (OLC)[51].*** *Les lectures sont fournies à l'algorithme qui identifie les overlaps (séquences communes) entre les lectures. Un graphe est construit avec chaque lecture comme nœud et les overlaps comme bordures. L'algorithme détermine ensuite le meilleur chemin à travers le graphe (chemin Hamiltonien). Ce processus peut être effectué plusieurs fois afin de produire l'assemblage final.*

9.1.3 Comment reconstituer les chromosomes

Pour que les assemblages atteignent l'échelle des chromosomes, il est très souvent nécessaire d'ajouter des données "long range" qui apportent des informations sur de grandes distances. Les deux technologies communément employées sont les cartes optiques et la capture de la conformation 3D de la chromatine ou Hi-C (Figure 18).

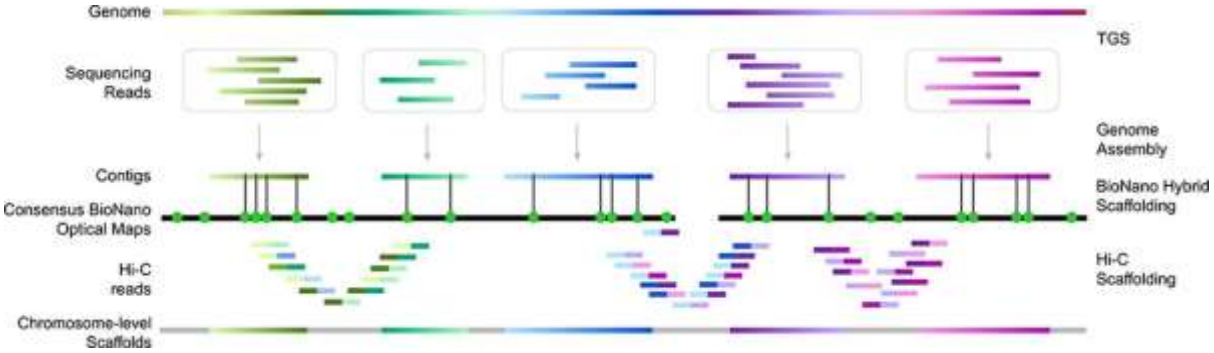


Figure 18 : ***Obtention d'assemblage à l'échelle des chromosomes[58]****. Les génomes sont séquencés grâce aux technologies longues lectures TGS (Third Generation Sequencing). Les lectures obtenues sont assemblées en contigs. Ceux-ci sont orientés et ordonnés grâce à la comparaison avec une carte optique (processus d'hybrid scaffolding). Les hybrid scaffolds obtenus sont regroupés pour atteindre l'échelle des chromosomes grâce à un scaffolding utilisant des données de séquençage Hi-C.*

9.1.3.1 Les cartes optiques

Les cartes optiques sont réalisées grâce à un instrument, le Saphyr, commercialisé par Bionano Genomics (BNG). L'enjeu principal est de réaliser une extraction d'ADN de très haut poids moléculaire la plus pure possible pour éliminer tout contaminant (polyphénols ou polysaccharides) qui inhiberait les réactions enzymatiques. Les fragments d'ADN obtenus (idéalement d'une taille supérieure à 150Kb minimum) sont soumis à l'action d'une enzyme qui reconnaît un site spécifique. L'enzyme digère le fragment uniquement sur un des deux

brins. Celui-ci est réparé et les nucléotides incorporés sont marqués avec un fluorochrome. Chaque site de coupure est appelé un label. Il existe deux méthodes de marquage : NLRS (pour Nicking - Label - Repair and Stain) qui utilise des enzymes de restriction (la plus utilisée étant BspQI) et DLS (pour Direct Label and Stain) qui utilise une enzyme (DLE-1) dont la fréquence de marquage est plus homogène mais qui n'induit pas de coupure sur le fragment d'ADN. Par la méthode NRLS, si les sites de coupure sont trop proches les uns des autres, une cassure du fragment d'ADN peut avoir lieu. Les molécules ainsi marquées sont ensuite déposées sur une flow cell et vont être linéarisées avant de migrer à l'intérieur de microcanaux. Une caméra prend des photos des microcanaux après excitation des fluorochromes par un laser (Figure 19). Les images sont converties en molécules, comportant l'information de position des labels et la distance entre chaque label.

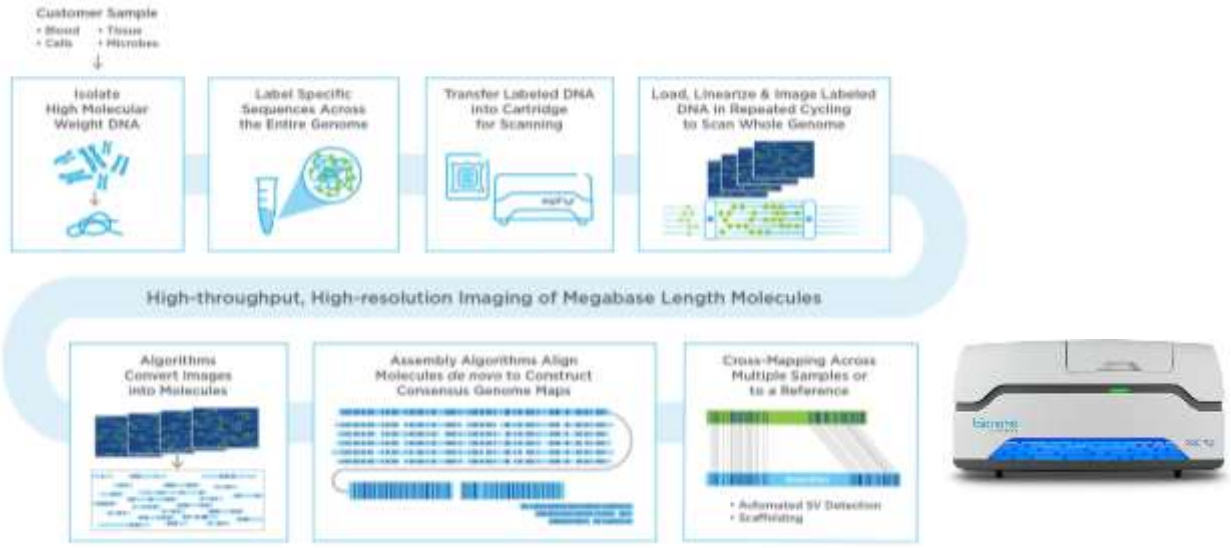


Figure 19 : ***Procédure d'obtention d'une carte optique et instrument Saphyr (Bionano Genomics)***[59].

Les molécules sont ensuite assemblées grâce à un outil dédié fourni par la société Bionano Genomics pour obtenir une carte optique. L'assemblage obtenu à partir de longues lectures est digéré

in silico avec la même enzyme de restriction. La position des labels est comparée entre la carte optique et l'assemblage digéré. Cette comparaison permet d'orienter et d'organiser les contigs les uns par rapport aux autres et d'obtenir des scaffolds qui peuvent être à l'échelle de chromosomes entiers ou de bras de chromosomes. L'avantage d'utiliser une carte optique est qu'il est possible d'organiser les régions complexes comme les centromères qui contiennent un fort taux de séquences répétées et d'estimer de façon très précise la taille des trous entre les contigs. Par contre, il peut être très difficile d'obtenir une extraction de très haut poids moléculaire de qualité suffisante.

9.1.3.2 La capture de la conformation de la chromatine ou Hi-C

L'Hi-C est une technologie permettant de capturer les contacts formés par la chromatine dans le noyau. Il mesure le degré d'interaction entre deux loci du génome. En effet, la chromatine est une structure complexe composée d'ADN et de protéines. Sa compaction permet de rapprocher spatialement des régions pourtant éloignées sur la séquence. Les interactions entre deux régions de l'ADN forment des points de contact et la distance entre ces deux régions est variable pouvant aller de quelques nucléotides à plusieurs dizaines de kilobases[60] (Figure 20). Dans leur publication de 2009, Lieberman-Aiden et al montrent la présence de territoires chromosomiques[61]. Ils ont mis en évidence un niveau supplémentaire d'organisation du génome caractérisé par la ségrégation spatiale de la chromatine ouverte et fermée pour former deux compartiments à l'échelle du génome (compartiments A et B). Ces compartiments semblent être spécifiques du type cellulaire. Les régions associées aux compartiments A/B se situent sur une échelle de plusieurs Méga-bases et sont corrélées soit à une chromatine ouverte et active sur le plan de l'expression (compartiments "A"), soit à une chromatine fermée et inactive sur le plan de l'expression (compartiments "B"). L'expression des gènes nécessite également que des régions de l'ADN interagissent entre elles formant alors des boucles ou des TADs (Topologically associated domains)[62]. Ces boucles permettent de rapprocher des séquences promotrices et activatrices par exemple. L'Hi-C tire donc parti de ces structures naturelles de la chromatine.

Il existe plusieurs protocoles commerciaux permettant de réaliser les

banques Hi-C. Ces protocoles sont tous basés sur le même type de réactions : la conformation de la chromatine est fixée dans le noyau et les contacts sont maintenus (Figure 21). L'ADN est ensuite fragmenté à l'aide d'enzymes (soit des enzymes de restriction dans le cas des kits commercialisés par Arima Genomics soit une DNAse dans le cas du kit Omni-C commercialisé par Dovetail Genomics). Les extrémités des fragments formant des contacts sont liées par un adaptateur portant une biotine. L'ADN est ensuite débarrassé des protéines et les fragments porteurs de la biotine sont purifiés grâce à des billes de streptavidine (qui a une très forte affinité pour la biotine). Les fragments sont ensuite soumis à une préparation de banque de séquençage de type Illumina en vue d'être séquencés en mode Paired End (2*150 bases). Il existe un protocole spécifique pour le séquençage nanopore : le Pore-C (Figure 22). Le principe est identique mais une concaténation de fragments de contacts est produite pour rendre la taille des fragments compatibles avec un séquençage longue lecture

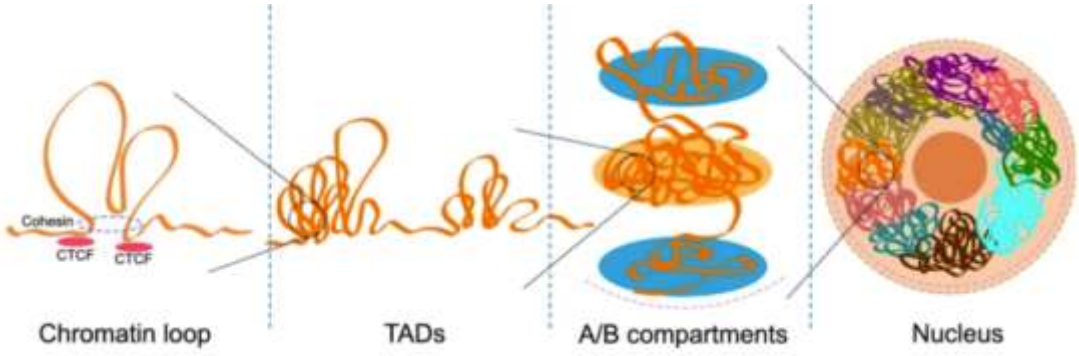


Figure 20 : ***Structure 3D du génome[60].*** *Dans le noyau, une première ségrégation spatiale sépare la chromatine ouverte de la chromatine fermée (compartiments A/B). A l'intérieur des compartiments, l'ADN forme différents types de contacts : les TADs (Topologically associated domains) et les boucles.*

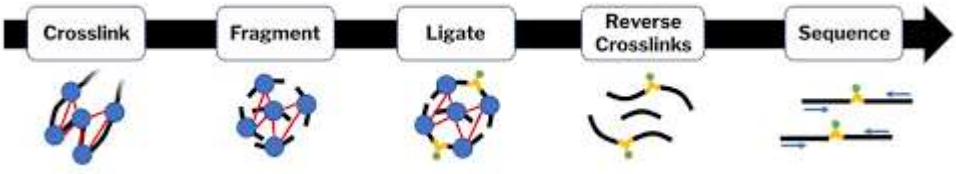


Figure 21 : ***Protocole Hi-C (Dovetail Genomics)[63].*** *La chromatine est fixée directement dans le noyau pour maintenir sa conformation. Les contacts sont maintenus. L'ADN est ensuite fragmenté et les extrémités des fragments en contact sont reliés par un adaptateur porteur d'une biotine. L'ADN est ensuite débarrassé de ses protéines et une purification des fragments porteurs de biotine est réalisée. Ceux-ci sont engagés dans une préparation de banque en vue d'un séquençage courte lecture (sur instrument de type Illumina).*

De nombreux logiciels[64,65] ont été développés pour utiliser ces lectures Hi-C afin de réaliser un scaffolding des contigs issus de longues lectures. L'assemblage obtenu est très souvent à l'échelle des chromosomes mais certaines régions posent problème aux outils, comme les centromères et les grandes répétitions en tandem. En effet, la première étape est d'aligner les lectures Hi-C sur l'assemblage et d'éliminer les lectures pouvant s'aligner à différentes positions. Cela conduit à des difficultés d'organisation de ces régions complexes. De plus, la taille des trous entre les contigs n'est pas estimée et fixée de façon arbitraire.

Les deux technologies dites "Long Range" peuvent se compléter. Il est possible d'enchaîner un hybrid scaffolding avec une carte optique et un scaffolding à l'aide d'une banque Hi-C. C'est ce qui est d'ailleurs proposé par de grands consortiums comme le Vertebrate Genome Project qui a pour but de réaliser le génome de référence d'environ 10,000 vertébrés sur les 70,000 espèces existantes[66].

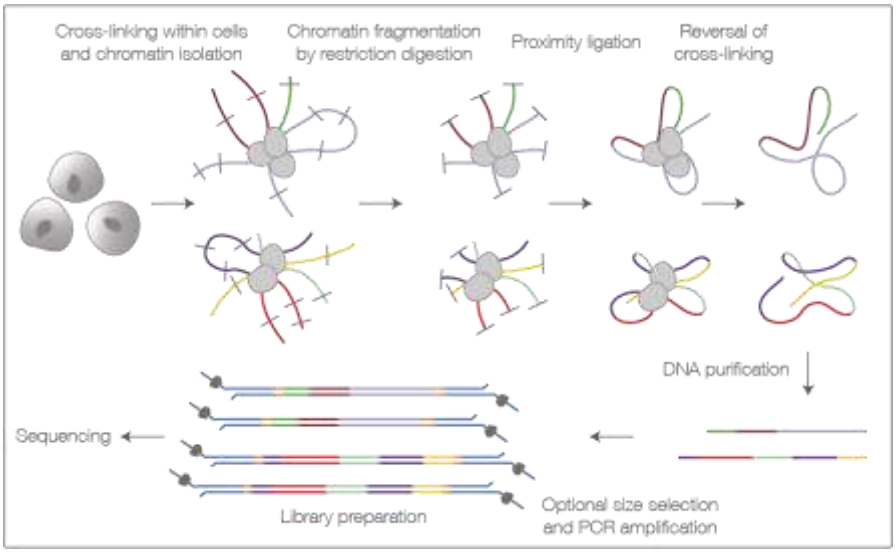


Figure 22 : ***Protocole Pore-C (Nanopore Technologies)[67].***

9.1.4 Pourquoi générer des génomes de référence ?

Les génomes de référence, ou tous génomes de grande qualité, se distinguent des versions de génome dites "draft" par leur niveau de complétion (très faible nombre de trous), leur faible nombre d'erreurs et le pourcentage élevé de séquences organisées en chromosomes. Un génome de référence correspond au génome d'un seul individu d'une espèce d'intérêt. Réaliser un (ou plusieurs) génome de référence constitue le socle de nombreuses analyses.

De façon plus générale, le séquençage d'ADN s'est imposé comme un outil essentiel pour de nombreux domaines des sciences de la vie, comme la biologie moléculaire ou la médecine. La génomique est une discipline qui regroupe un ensemble d'analyses permettant d'étudier le génome d'un individu. L'identification et le séquençage des gènes, l'étude de leurs fonctions et du contrôle de leur expression sont à l'origine de grandes avancées en médecine par exemple. Annoter les gènes sur un génome de référence permet de déterminer l'ensemble du contenu génique d'un individu. L'étude de son transcriptome, grâce au séquençage RNAseq, quant à lui, permettra de connaître l'expression de ses gènes spécifiquement dans certains tissus et dans

certaines conditions. Le séquençage de transcripts grâce au technologies longues lectures permet d'explorer les transcripts alternatifs et les transcripts de fusion[68]. L'étude des génomes de tumeurs cancéreuses, de leurs variants structuraux, du dérèglement de la régulation de leurs gènes offrent des opportunités de recherche et de soins. La détection des bases méthylées[69] grâce à la technologie Oxford Nanopore permet d'étudier les modifications d'expression des gènes induites lors du processus de cancérisation par exemple. Le séquençage de banques Hi-C permet également de comparer les compartiments A/B des génomes entre différentes conditions de vie, différentes lignées cellulaires ou différents stades du développement[70]. La chromatine très compactée est appelée hétérochromatine. L'expression des gènes y est alors réprimée en raison du manque d'accès à la séquence d'ADN. Cependant, afin d'assurer la vie de la cellule, tout le génome ne peut pas être réprimé. Le compartiment A représentant la chromatine accessible à la transcription ou euchromatine, toute modification dans sa structure a un impact sur la régulation de l'expression des gènes.

Il est parfois essentiel d'accumuler les génomes de référence de différentes variétés ou souches à l'intérieur d'une même espèce pour réaliser des études de génomique comparative, permettant de retracer l'histoire évolutive d'une espèce ou de déterminer quelles variations génomiques sont à l'origine de l'apparition de phénotypes particuliers[71]. La comparaison des génomes de référence permet également de déterminer l'existence de variants structuraux, eux-mêmes à l'origine de phénotypes différents.

De nombreuses espèces de plantes cultivées ont des génomes issus d'événements de polyploïdisation (autopolyploïdisation), ancestraux ou récents, ou sont issus d'hybridation entre différentes espèces ou sous-espèces (allopolyploïdisation). Les multiples copies de gènes homéologues, les réarrangements chromosomiques et l'amplification d'éléments répétés peuvent considérablement compliquer l'analyse du génome et la découverte de gènes par des approches classiques de génétique directe.

Chez les semenciers (entreprises qui produisent les semences pour l'agriculture), l'analyse des génomes permet de trouver les gènes

responsables de traits clés et de découvrir la distribution de la diversité génétique ainsi que sa relation avec les caractéristiques de rendement et de qualité[72,73]. Ces recherches permettent de mettre en lumière les origines de la domestication et d'identifier la variation existante, ce qui facilite la sélection de nouvelles variétés[74]. Les semenciers doivent en effet faire face aux changements climatiques, aux problèmes d'accès à l'eau et aux nouveaux pathogènes, les obligeant à améliorer les variétés actuellement cultivées.

Il est même parfois crucial de générer des génomes de référence, pour des génomes préalablement réalisés grâce à de courtes lectures (assemblages "drafts"). Les longues lectures permettent alors de compléter le génome et d'en connaître sa structure. Cela a été le cas pour le génome du maïs par exemple qui compte parmi les céréales les plus cultivées dans le monde[75].

Avec l'avènement des longues lectures, il est possible d'accéder au contenu en séquences répétées et en éléments transposables. L'accumulation de ces éléments, qui peuvent mesurer plusieurs dizaines de kilobases, a, par exemple, un rôle prépondérant dans la variation de la taille des génomes entre différentes espèces de plantes à fleurs[76]. Cette "matière noire" ("dark matter" en anglais) a été longtemps mise de côté mais de plus en plus d'études montrent le rôle de ces ADN mobiles résiduels dans l'acquisition ou la mobilisation de nouvelles fonctions pour répondre à un besoin d'acclimatation ou de résistance à des pathogènes[77].

Il est en outre possible de se focaliser sur des régions d'intérêt comme les clusters de gènes de résistance qui confèrent des résistances à des pathogènes chez les plantes par exemple[78,79].

Enfin, dans le contexte du réchauffement climatique et de la disparition d'une part de la biodiversité actuelle, la connaissance des génomes de référence est un point clé dans la conservation des espèces et de leur patrimoine génétique, et dans la compréhension des phénomènes de résilience ou d'adaptation mis en place par certaines espèces (Figure 23)[80]. De grands projets de séquençage de génomes de référence ont vu le jour, comme le Vertebrate Genome Project (VGP), l'European Genome Reference Atlas (ERGA), le Darwin Tree of life… Ils ont pour

vocation de produire les génomes de référence du plus grand nombre d'espèces à travers l'arbre du vivant et monopolisent les efforts de séquençage d'un grand nombre de plateformes de séquençage à travers le monde. Le VGP, au travers d'une série de publications dans un numéro spécial de la revue *Nature*[81], a présenté des méthodes d'assemblage et de contrôle qualité des génomes. Il impose, de fait, de nouveaux standards de qualité[82]. Le consortium exige pour tous les génomes publiés :

- une taille N50 d'au moins 1 Mb pour les contigs et 10 Mb pour les scaffolds (le N50 est la longueur minimale pour que 50 % du génome soit couvert par des contigs ou des scaffolds de cette longueur ou plus),
- que la fréquence des erreurs de séquence ne soit pas supérieure à 1 sur 10 000 bases,
- que les variants structuraux soient confirmés par plusieurs technologies
- qu'au moins 90 % de la séquence soit attribuée à des chromosomes et avec une résolution des haplotypes.

La figure 23 illustre les applications accessibles grâce à ces génomes de référence dans le cadre de la conservation des espèces. Par exemple, les efforts de conservation doivent tenir compte de la diversité génomique pour optimiser les stratégies, maintenir la viabilité des populations et préserver leur potentiel d'adaptation aux changements environnementaux.

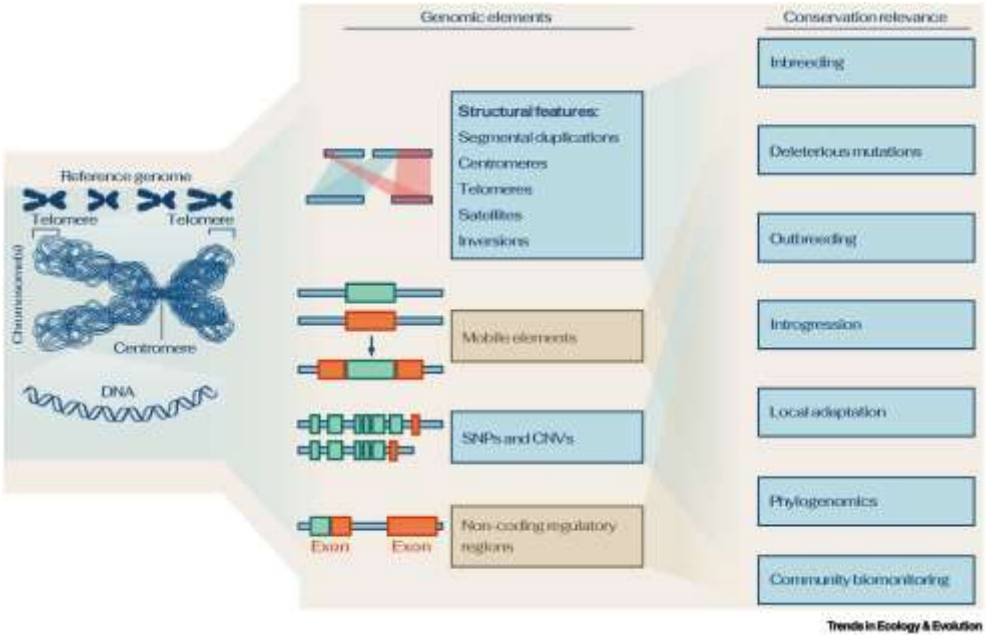


Figure 23 : ***Génomes de référence et conservation des espèces[80]****. Quelques exemples d'application des génomes de référence dans le cadre de la conservation des espèces.*

9.1.5 Défis

9.1.5.1 Obtenir des longs fragments d'ADN

Le premier défi est d'appliquer le bon protocole d'extraction afin d'obtenir les fragments d'ADN les plus longs et les plus intègres possibles tout en obtenant la préparation la plus pure. La longueur des lectures obtenues est limitée par la taille des fragments d'ADN, particulièrement pour le séquençage nanopore. Ces grands fragments sont difficiles à obtenir pour les plantes ou les organismes marins. De nombreux protocoles ont été développés mais ne sont pas toujours applicables à tous les types d'organismes, chacun ayant ses propres particularités en termes de métabolites secondaires synthétisés par exemple[83]. La présence de ces métabolites a pour conséquence d'inhiber des réactifs utilisés lors de la préparation des banques ou lors de la préparation des cartes optiques. L'accès à une quantité de matériel suffisante peut également être un frein. Il faut ainsi réaliser des adaptations pour les matériels de faible quantité initiale[84]. De la même façon, les extractions à partir d'échantillons complexes, comme

les prélèvements de sédiments, de sols ou les filtres d'eau, nécessitent aussi des adaptations spécifiques afin d'obtenir des fragments de taille compatible avec les séquenceurs de troisième génération.

9.1.5.2 Complexité des génomes

Les difficultés rencontrées pour obtenir des génomes de référence sont le plus souvent également liés à la complexité de ces génomes. Mais qu'est-ce qu'un génome complexe? C'est une notion bien difficile à définir. La complexité d'un génome n'est pas directement corrélée à la complexité de l'organisme lui-même. On admet que le génome des eucaryotes est plus complexe que le génome des procaryotes. Par exemple, la structure de leurs gènes est différente. En particulier, les génomes eucaryotes contiennent des séquences d'ADN non codantes, appelées introns et qui jouent un rôle lors de la transcription. La taille du génome n'est elle-même pas directement corrélée à la complexité de l'organisme[85] : par exemple, la taille du génome de la salamandre est bien supérieure à celle du génome humain bien qu'il s'agisse d'un organisme moins complexe[86]. Par ailleurs, elle est très variable entre différentes espèces de salamandre: elle peut varier entre 10 et 120 Gigabases alors que le génome humain contient 3 Gigabases. Il faut donc regarder la composition de ces génomes pour comprendre ce qui les rend complexes[87].

La présence d'ADN non codant en est une des principales causes. Outre la structure des gènes en introns/exons, on retrouve des familles de gènes et de pseudogènes contenant des gènes répétés un grand nombre de fois. Ceci peut être utile, comme dans le cas des histones, pour produire une grande quantité de protéines. Les membres de ces familles peuvent être également exprimés à des stades différents du développement ou dans des tissus différents. Ces gènes dupliqués sont souvent présents sous forme de clusters mais peuvent être également éparpillés dans le génome[88]. Les familles multigéniques sont issues de duplications d'un gène ancestral suivies de la divergence des différentes copies. Ceci peut avoir eu pour conséquence d'optimiser leur fonction ou au contraire de leur faire perdre leur fonction. C'est le processus de pseudogénisation.

Enfin, on retrouve des séquences répétées en multiples copies ou

éléments répétés, parmi lesquels on peut citer les satellites, les SINEs (Short interspersed Elements) et les LINEs (Long interspersed Elements). Certains de ces éléments sont dits transposables (TE) car ils sont capables de se déplacer le long du génome. L'accumulation des TE est à l'origine de différences de tailles entre les variétés d'une même espèce mais également entre espèces. Le génome du blé par exemple est composé d'environ 80% d'éléments répétés. Ce sont pour la plupart des retrotransposons de type Long terminal repeat (LTR-R), extrêmement répandus chez les plantes et qui provoquent l'expansion du génome par amplification à l'aide d'un intermédiaire ARN. Les LTR-R facilitent la création de nouveaux gènes candidats appelés rétrogènes par le biais de la rétroduplication, au cours de laquelle l'ARN messager épissé est capturé, transcrit de manière inverse, puis intégré au génome par un rétrotransposon[89]. Des études ont rapporté la capture de familles de gènes spécifiques par certains TE et ont suggéré une corrélation entre la duplication de gènes médiée par les TE et l'expansion de familles de gènes spécifiques[90]. Les protéines de type NLR (pour Nucleotide-binding and leucine-rich-repeat proteins) représentent une famille de gènes très amplifiée chez les plantes et fournissent la majorité des loci fonctionnels de résistance aux maladies[91]. Des analyses génomiques comparatives ont suggéré la possibilité d'une co-évolution des LTR-R et des NLR, en partie parce qu'ils sont souvent co-localisés[92]. Il leur est attribué des propriétés dans l'adaptation des organismes à de nouvelles conditions de vie ou de lutte contre des parasites ou maladies[79].

Le dernier niveau de complexité est relié au niveau de ploïdie du génome et à la divergence (hétérozygotie) entre chaque copie du génome. Un génome polyploïde contient plusieurs copies de son génome. Ces copies sont apparues durant l'évolution soit par autopolyploïdisation (mécanisme de multiplication d'un même génome) soit par allopolyploïdisation (hybridation de deux ou de plusieurs génomes différents). Ce phénomène d'amplification peut être suivi d'une réduction du nombre de copies pour stabiliser le génome résultant[93]. Ces amplifications ont un intérêt évolutif car elles permettent de gagner de nouvelles fonctions[94]. Ces mécanismes sont à l'origine de la grande diversité d'espèces chez les *Brassicaceae* par exemple[95].

9.1.5.3 Hétérozygotie des génomes

Le niveau de divergence entre les copies des gènes est représenté par le taux d'hétérozygotie: plus le niveau de divergence est élevé, plus le taux d'hétérozygotie est élevé. Ce taux est directement relié à la présence d'une population nombreuse avec une grande diversité génétique. Si la population est restreinte, la diversité génétique s'affaiblit.

Les génomes polyploïdes et/ou très hétérozygotes sont toujours très difficiles à résoudre et requièrent encore des adaptations technologiques. Les génomes diploïdes (contenant un exemplaire paternel et un exemplaire maternel) sont résolus par des approches spécifiques[96]. En effet, générer des génomes diploïdes nécessite de séparer les haplotypes, sans créer ce que l'on appelle des chimères d'haplotypes (les haplotypes maternels et paternels sont mélangés). Lors du processus d'assemblage, les outils peuvent faire des erreurs d'association lors de la traversée de régions homozygotes (identiques aux deux haplotypes). Pour séparer les haplotypes, il convient d'ajouter les données de séquençage des deux parents, fournies par exemple avec la méthode du TrioBinning[97]. Cette méthode permet de distinguer les k-mers uniques pour chaque parent et ainsi de séparer les contigs associés à chaque haplotype. Il est également possible de séquencer des gamètes qui contiennent un seul haplotype[98]. Enfin, la technologie Hi-C, associée à des outils comme ALLHiC[99], permet également de distinguer les contacts spécifiques de chaque haplotype. L'assembleur Hifiasm[100] associé à la nouvelle technologie HiFi proposée par la société PacBio permet de séparer les deux haplotypes mais pas forcément d'obtenir les deux versions complètes.

Pour les génomes triploïdes et au-delà il reste encore des développements méthodologiques à réaliser. Il n'y a pas de procédures standardisées et les approches restent spécifiques des génomes étudiés. Les données de séquençage Hi-C peuvent aider à séparer les haplotypes mais cela génère des chimères d'haplotypes avec une alternance d'haplotypes le long des chromosomes[101]. Le séquençage de grains de pollens, contenant chacun un génome haploïde, à faible couverture permet de séparer les contigs, issus de l'assemblage des longues lectures réalisées sur le génome entier, pour

chaque haplotype[102]. Cette méthode reste encore coûteuse.

La complexité d'un génome est donc la résultante de plusieurs facteurs, comme le nombre de gènes, la présence d'éléments répétés, le niveau de ploïdie et le taux d'hétérozygotie. Plus ces facteurs augmentent, plus le génome est complexe et difficile à reconstituer.

9.1.5.4 Assemblage des métagénomes

Enfin, le dernier défi se porte sur les espèces non cultivables et pour lesquelles il est très difficile d'obtenir suffisamment de matériel génétique pour établir la séquence de leur génome. Ces organismes sont principalement issus de prélèvements environnementaux. Une solution est de les séquencer en masse grâce aux techniques de métagénomique dans le but d'appréhender la composition en organismes d'un habitat précis. Mais il y a un réel manque de génomes de référence dans les bases de données ce qui rend difficile les analyses sur les populations. Les expériences de métagénomique réalisées à partir de ces données environnementales pallient donc à ces deux inconvénients : impossibilité de cultiver les organismes et absence de référence.

Les séquences obtenues à partir de l'ADN d'un métagénome sont séparées en paquets par des méthodes de binning. Celles-ci trient les séquences grâce à leur fréquence en k-mers. Chaque paquet ou bin est assemblé pour former un MAG (Metagenome Assembled Genome). Les assemblages sous forme de MAGs sont majoritairement incomplets ou fragmentés. L'enjeu des outils d'assemblage est de réussir à regrouper les lectures ou les contigs par organisme au sein d'un mélange qui peut être très complexe[103,104]. Ils doivent tenir compte de la différence d'abondance des espèces présentes dans le pool, de l'hétérogénéité inter et intra espèces, des différences de taille des génomes. Ces méthodes présentent des limites particulièrement dans le cas des mélanges très complexes.

De nombreux articles présentent des collections de MAGs, principalement procaryotes, généralement obtenus à partir de séquençage courtes lectures[105,106]. L'augmentation des débits et à la baisse des taux d'erreur rend dorénavant possible le séquençage de

métagénomes en longues lectures. Néanmoins, un effort sur les protocoles de conservation des échantillons environnementaux et d'extraction d'ADN devra être réalisé pour parvenir à obtenir les fragments les plus longs possibles et ainsi profiter pleinement de l'avantage des longues lectures. Le but est d'obtenir des MAGs plus complets, plus continus.

9.2 OBTENTION DE GENOMES DE REFERENCE A L'ECHELLE DES CHROMOSOMES

Les nouvelles générations de séquenceurs et leurs courtes lectures ont permis de séquencer un grand nombre de génomes de plantes mais leur assemblage était souvent très fragmenté voire incomplet. Dans cet article, nous avons appliqué le séquençage longues lectures à deux plantes dicotylédones et à une plante monocotylédone.

9.2.1 Contexte

Les *Brassicaceae* forment une famille de plantes dicotylédones présentant un grand intérêt scientifique et économique. Elles comportent, par exemple, la plante modèle de référence *Arabidopsis thaliana*, les crucifères (choux) ou le colza. Cette famille a une très grande variété de morphotypes (Figure 24) et forme un clade modèle pour la compréhension des plantes polyploïdes et pour la paléogénomique[107].

Les espèces dérivent toutes d'un ancêtre commun nommé ABK[108] : Ancestral *Brassicaceae* Karyotype (*Brassicaceae* lineages I and II) dont le génome était composé de 8 protochromosomes et de 20,037 protogenes. Après une première spéciation, il a évolué en ACK pour Ancestral Camelineae Karyotype (8 protochromosomes et 22,085 protogenes) et PCK Proto-Calepineae Karyotype (7 protochromosomes et 21,035 protogenes). Après des étapes de triplication du génome et des étapes de fractionnement, les espèces *Brassica rapa* et *Brassica oleracea* sont apparues. Le colza, *Brassica*

napus, est issu de l'hybridation de ces deux *Brassica*[109]. Il a accumulé au cours de son évolution 72 génomes ancestraux, résultat de nombreux cycles de polyploïdisation, faisant de son génome un des plus hautement dupliqués chez les plantes à fleurs (angiospermes)[110]. Ce phénomène récurrent, suivi par des restructurations du génome, a conduit à l'accumulation d'un grand nombre de gènes (101 040 gènes).

U Nagaharu, botaniste nippo-coréen, décrit pour la première fois en 1935 le triangle de U (Figure 25)[111]. Celui-ci est composé des 3 espèces diploïdes, *Brassica rapa* (AA), *Brassica nigra* (BB), et *Brassica oleracea* (CC); et de 3 espèces hybrides allotétraploïdes, *Brassica juncea* (AABB), *Brassica napus* (AACC), et *Brassica carinata* (BBCC). Comme les espèces de *Brassica* ont subi plusieurs événements de polyploïdie[112], les génomes des espèces diploïdes et allopolyploïdes sont des modèles importants pour mettre en lumière l'effet immédiat et à long terme de la polyploïdie sur la dynamique évolutive structurelle et fonctionnelle des gènes et des génomes dupliqués. Ces événements ont joué un rôle important dans leur diversification. La variabilité entre 2 morphotypes d'une même espèce de *Brassica* peut être élevée, soulignant l'importance de disposer d'une collection d'assemblages de haute qualité pour le genre *Brassica* et plus généralement pour toutes les espèces.

Connaître les génomes des progéniteurs *B. rapa* et *B. oleracea* est essentiel pour mieux appréhender la structure de *B. napus*, plante extrêmement importante pour la production d'huile dans le monde[113].

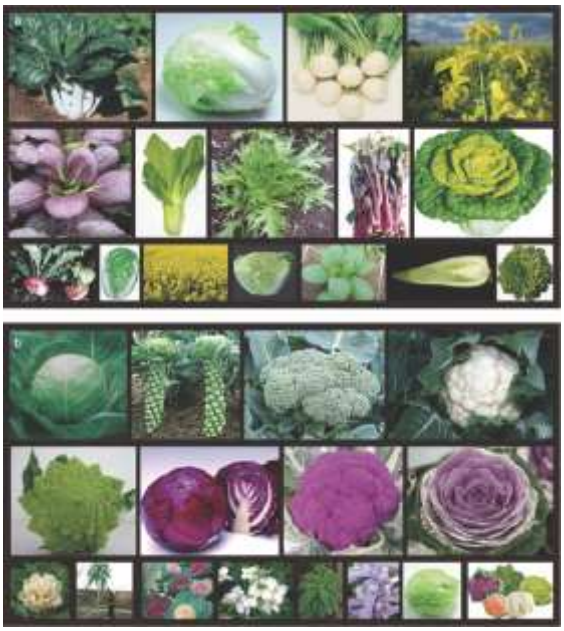


Figure 24 : ***Morphotypes de plantes Brassica***[107]*. (a) Morphotypes de B. rapa ; les deux premières lignes de gauche à droite : pak choi, B. rapa, navet, oléagineux, pak choi violet, caixin, mizuna, caitai violet et takucai ; la troisième ligne montre des morphotypes supplémentaires ou des variétés des morphotypes précédents. (b) Morphotypes de B. oleracea ; les deux premières lignes de gauche à droite : chou pommé, chou de Bruxelles, brocoli, chou-fleur, chou Romanesco, chou violet, chou-fleur violet, chou-fleur; la troisième ligne indique les morphotypes ou variétés supplémentaires.*

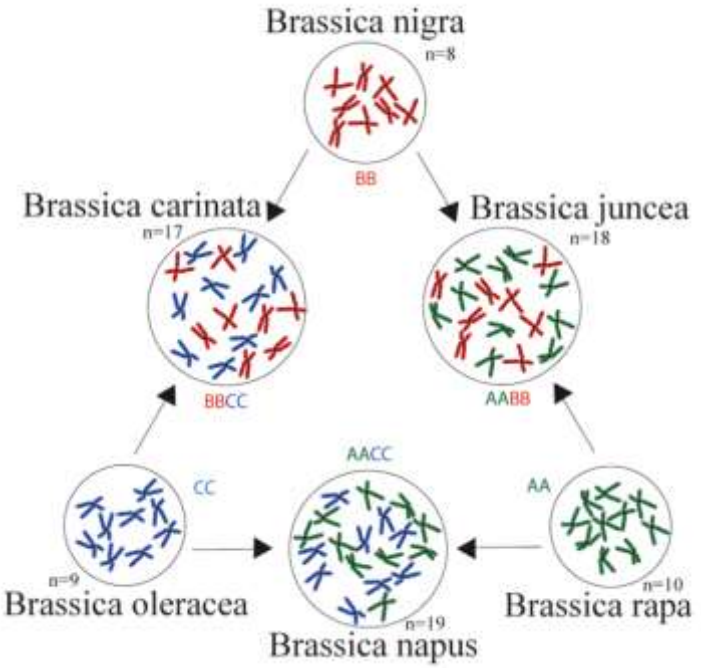


*Figure 25 : **Triangle de U**. Il est composé des 3 espèces diploïdes, Brassica rapa (AA), Brassica nigra (BB), et Brassica oleracea (CC); et de 3 espèces hybrides allotétraploïdes, Brassica juncea (AABB), Brassica napus (AACC), et Brassica carinata (BBCC). Les flèches montrent les relations entre les espèces diploïdes et tétraploïdes. La couleur des chromosomes est associée à une espèce diploïde*[114].

De leur côté, les bananiers sont des plantes monocotylédones, appartenant à l'ordre des *Zingiberales* et à la famille des *Musaceae*, cultivées essentiellement dans les pays tropicaux et subtropicaux (Figure 26)[115]. Leurs fruits sont à la base de l'alimentation de plus de 400 millions de personnes et sont exportés massivement vers les pays développés. Les bananes et les plantains sont considérés comme la quatrième culture la plus importante au monde après le riz, le blé et le maïs.

Quatre groupes génétiques ont été identifiés comme étant impliqués dans l'origine des cultivars actuels, principalement par hybridation inter(sub)spécifique et avec différents degrés de contribution : *Musa acuminata* incluant diverses sous-espèces

(génome A, 2n=2x=22) (ex : Cavendish AAA), *Musa balbisiana* (génome B, 2n=2x=22), *Musa Schizocarpa* (génome S, 2n=2x=22) et les espèces de la section *Australimusa* (génome T, 2n=2x=20). L'hybridation interspécifique est le croisement entre des groupes ou des taxa génétiquement différents. L'hybridation inter(sub)spécifique se produit entre des membres de sous-espèces différentes. Les variétés modernes de bananes et de bananes plantains sont polyploïdes et ont une structure génétique complexe. Comme la plupart des bananes cultivées sont parthénocarpiques (généralement sans graines), elles doivent être propagées végétativement.

Figure 26: ***Exemple d'accessions de Musa.*** *A; Musa coccinea, B: M. velutina, C: M. laterita, D: M. beccarii, E: M. textiles, F: M. acuminata, G: M. balbisiana, H: banane à dessert (AAA), I: plantain (AAB), J: un tétraploïde hybride (AAAB) and K: banane à cuisiner (ABB) avec deux inflorescences*[115].

Deux événements sont apparus au cours de la domestication

du bananier : la transition des diploïdes sauvages aux diploïdes comestibles et l'émergence des triploïdes à partir des diploïdes comestibles. Des résultats récents suggèrent que l'origine des cultivars comestibles est plus complexe que prévu, impliquant de multiples étapes d'hybridation, résultant en génomes mosaïques inter(sub)spécifiques[116].

9.2.2 Introduction

L'étude de ces génomes complexes nécessite l'obtention de génomes de référence de très haute qualité. La figure 27a présente le pipeline utilisé dans l'article ci-après, que j'ai coécrit. Un assemblage réalisé à partir de la combinaison de longues lectures avec une carte optique générée grâce à l'enzyme DLE-1 est ici pour la première fois décrite. Cette nouvelle enzyme a pour avantage de ne pas induire de cassures dans les fragments d'ADN et arbore une fréquence de marquage plus homogène. Les cartes optiques ainsi obtenues atteignent l'échelle des chromosomes ou des bras de chromosomes. Cette méthode a permis d'obtenir les génomes de référence de trois espèces jusque là non séquencées : *Brassica rapa ssp. trilocularis* (génotype Z1 - sarson jaune), *Brassica oleracea* ssp. *italica* (génotype HDEM - brocoli) et *Musa Schizocarpa* (ITC926). Les génomes de trois parents (*B. rapa* Chiifu, *B. oleracea* To1000 and *Musa acuminata* Pahang-HD) avaient déjà été réalisés à partir de séquençage courte lecture. Ces références sont fragmentées et contiennent un fort taux de bases indéterminées (20-30%). La taille cumulée des assemblages est également loin de la taille estimée. Il manque environ 273Mb sur les 529Mb estimées pour *B.rapa* Chiifu.

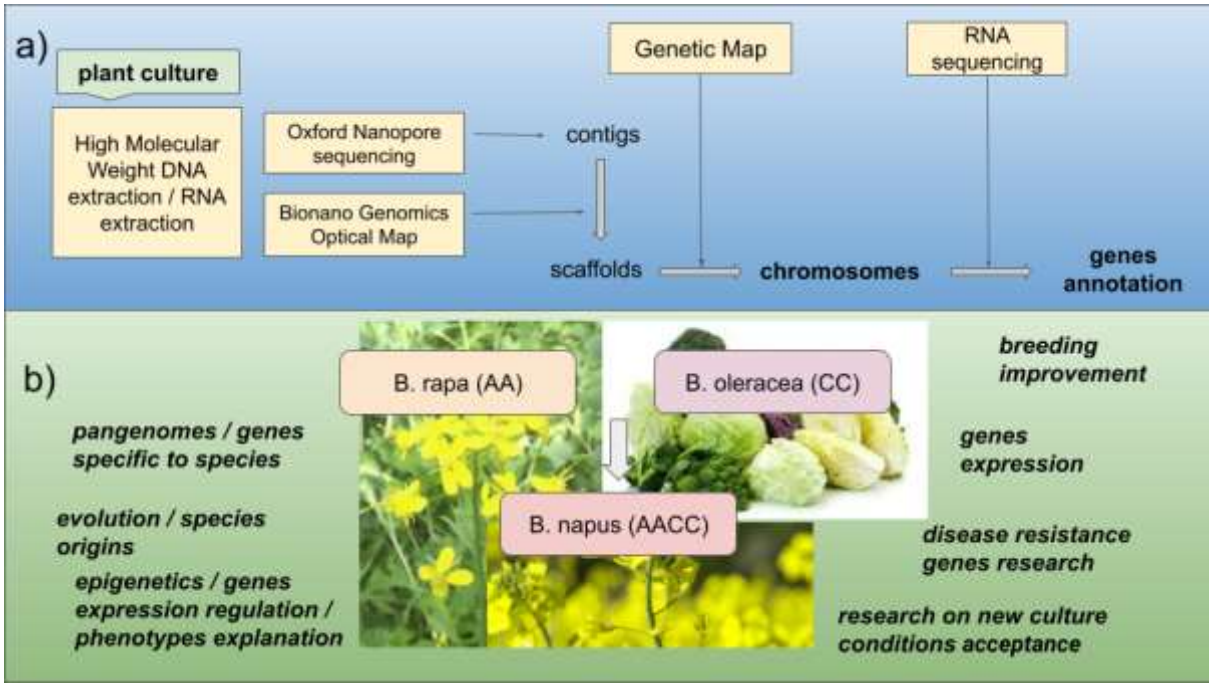


Figure 27 : ***Obtention de génomes de référence pour les plantes Brassica.*** *a) Pipeline utilisé pour générer les génomes de référence. b) Exemples d'applications accessibles grâce aux génomes de référence.*

9.2.3 Article : "Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps" Belser, C., Istace, B., Denis, E. et al. Nature Plants. 2018

J'ai inséré ici la dernière version acceptée de l'article. Celui-ci est également accessible sur le site de *Nature Plants* https://www.nature.com/articles/s41477-018-0289-4

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

Caroline Belser[1,§], Benjamin Istace[1,§], Erwan Denis[1,§], Marion Dubarry[1,§], 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,*]

[1] Commissariat à l'Energie Atomique (CEA), Institut de Biologie François-Jacob, Genoscope, F-91057 Evry, France
[2] CIRAD, UMR AGAP, F-34398 Montpellier, France
[3] AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France
[4] IGEPP, INRA, Agrocampus Ouest, Université Rennes 1, BP35327, 35653 Le Rheu
[5] Université Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France
[6] Commissariat à l'Energie Atomique (CEA), Institut de Biologie François-Jacob, Genoscope, CNRS UMR 8030, Université d'Evry, France

[§] These authors contributed equally to this work
* Corresponding author

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 significantly 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. 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 > 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[1] for dicotyledons and rice[2] 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 the Illumina sequencing technology, more than 200 plant genomes have now been sequenced, but most have poor contiguity (Figure 1) and are composed of thousands of scaffolds. Generally, the gene space is relatively complete and correctly assembled, but regions rich in transposable elements (TEs) are more fragmented or even under-represented. In addition, the dynamics of TEs is largely unknown and this knowledge gap is mostly due to the difficulties in assembling repeated elements from genome sequences obtained using short-reads 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[3-6]. 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 > 5 Mb: rice[2,7], arabidopsis[1], 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 (WGT) event and through parallel selection of paralogous genes[12]. This WGT 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 (*B. rapa* Chiifu[13,14], *B. oleracea* To1000[15] and *Musa acuminata* Pahang-HD[16,17]) have already been sequenced using short-read strategies but the resulting assemblies are fragmented (contig N50 < 50 Kb) and contain a high proportion of unknown or missing bases (22–30%, Supplementary Figure 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 the MinION, 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 38x and 79x Nanopore long reads containing a significant proportion of reads longer than 50 Kb, representing between 4.4x and 8.2x coverage (Supplementary Table 2). The resulting long-read assemblies showed high contiguity (less than 1000 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 and 450 times higher, while the scaffold N50s are in general lower but the published assemblies were built using genetic maps (Supplementary Figure 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, while 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 (Figure 2 and Supplementary Table 12).

A 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 Figure 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. TEs and, more generally, TE-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 LINE, LTR-retrotransposon and DNA transposon families and the average sizes of the detected TEs were higher (see Methods and Supplementary Figure 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) when compared with the reference genomes (Supplementary Table 15). The gene content was mostly the same between the short- and long-read assemblies, but the long-read assemblies improved the completeness of the TE catalogue as well as the genomic context of these TE-rich regions (Figure 3). Generally, genes inserted in TE-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 TE-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) due to the difficulty of sequencing long DNA fragments. We compared the recent *Vigna angularis*[19], *Vitis vinifera*[3], *Citrus maxima*[20], *Arabidopsis thaliana*, *Fragaria vesca* and *R. chinensis*[11] PACBIO data with our three ONT datasets and observed a higher proportion of long-reads (>50 Kb) in the Nanopore data (Supplementary Figure 3 and Supplementary Table 17). Moreover, the PACBIO coverage was higher (between 125x and 283x) 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

dataset that had the smallest coverage (~36x) and the longest reads (reads N50 of 31 Kb), showing the higher impact of read length over coverage and confirming the possibility of producing high-quality assemblies with 30x long reads coverage[6]. 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.

The 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 $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 flowcell that produced 17.6 Gb of data with a comparable read N50 size (26 kb vs 24 kb with the MinION, Supplementary Table 2). We assembled this PromethION dataset 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 $16.3 k to $6.5 k 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 the accessions regardless of the morphotype, except for the Chinese cabbage accessions of *B. rapa* (Supplementary Figure 13 and Table 18). And 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.

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 percent of 99.2% and 98.9%, Supplementary Figure 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 Figures 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 appear 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[21]) and three tandemly duplicated copies of FLC1, as suggested in a previous study[22]. In comparison we found 4 FLC genes in *B. rapa* Z1, as expected (Supplementary Figure 17). Furthermore, we investigated the S-locus, a 30–150 kb region that is strongly enriched in TEs[23,24], which causes major difficulties when attempting to assemble it in its entirety using short reads[25]. For example, the S-locus in the *Raphanus sativus* genome assembly is spread over several contigs[26], while that of *Brassica nigra*[27] is spread over 2.2 Mb, which suggests 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 (Supplementary Figure 18 and methods).

Whole genome comparison of *M. schizocarpa* with *M. acuminata*

revealed high variability in the centromeric regions (Supplementary Figure 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 and 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 generally difficult to assemble correctly. We compared the proportion of undetermined nucleotides for three orthologous R-genes clusters on both the *M. acuminata* and *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 (median identity percent of 98.0% between orthologs) and synteny (5 translocation events, due to assembly errors in the first M. acuminata assembly[17]) levels (see methods and Supplementary Figures 7 and 14).

We demonstrated that combining three technologies (Oxford Nanopore, BioNano Genomics and Illumina) can lead to high-quality and relatively low-cost genome assemblies when using the PromethION device (around $6k 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 TEs. We annotated the three assemblies and observed similar gene content in the gene-rich regions, but a more complete catalogue of TEs was produced and the S-loci of the two *Brassicas* could be entirely annotated thanks to the high quality of the assemblies in these regions. Further improvements are still needed to enable extraction of HMW DNA for all plants and systematic errors in the nanopore long reads mean Illumina sequences are still required to polish assemblies. Today, optical maps[8] or

chromosome conformation[28-31] capture is still mandatory to propose chromosome-scale assemblies for large plant genomes. Even though extraction of HMW 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.

Figure legends and Table

Figure 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 (red), Illumina (olive), Oxford Nanopore (green), Pacific Biosciences (blue) and Sanger (pink).

Figure 2. Circular representation of anchored scaffolds of *Brassica oleracea* HDEM, *Brassica rapa* Z1 and *Musa schizocarpa* genome assemblies. **a.** Density in Copia elements. **b.** Density in Gypsy elements. **c.** Density in TEs **d.** Gene density. **e.** Centromeric repeats positions. **f.** Telomeric repeats positions.

Figure 3. Base annotation of the three ONT genomes and the corresponding current references *(Brassica rapa*, *Brassica oleracea* and *Musa* species). Genomic regions were classified into several categories: repeats (olive), introns (violet), exons (pink), gaps (salmon) and unannotated regions (brown). Repeated elements were divided into five classes: DNA transposons (dark green), LINEs (light green), Copia (light blue), Gypsy (blue) and other known TEs (purple).

Table 1. Statistics of the genome assemblies. Statistics of HDEM, Z1 and *Musa schizocarpa* assemblies and gene content compared with existing reference genomes (To1000, Chiifu and *Musa acuminata*). If unspecified, the unit is base pair.

	Brassica oleracea		***Brassica rapa***		***Musa sp.***		
	To1000	HDEM	Chiifu	Z1	*Musa acuminata*	*Musa schizocarpa*	*Musa schizocarpa*
Reference	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 (Mbp)	630	630	529	529	523	587	587
# scaffolds (≥2Kb)	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)
Max size	64,984,695	48,260,371	54,546,898	38,870,275	46,622,217	52,742,985	52,101,276
# 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%)
# contigs (≥500bp)	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)
Max 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
#exons per gene pluri (avg:med)	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

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

Brassica rapa ssp. *trilocularis* Z1 and *Brassica oleracea* ssp. *botrytis italica* HDEM seeds were sown in Fertiss blocks (Fertil, France). 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.

Musa 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 (min 15°C, max 50°C), in 1 l pots until they reached a height of 50 cm and had five fully expanded leaves. Harvesting was performed after 8 months/1 year of culture.

DNA extraction

Musa schizocarpa DNA extraction

High molecular weight plant DNA extraction was performed using a modified mixed alkyl trimethyl ammonium bromide (MATAB) procedure[32,33]. A total of 2 g of 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 RNAse A. Crude extracts were maintained for 20 min at 74°C, extracted with an equal volume of chloroform-isoamylalcohol (CIAA, 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 of sterile water. Extract quality was evaluated using Pulse Field Gel Electrophoresis (PFGE) for size estimation, and spectrophotometry ($OD_{260/280}$ and $OD_{260/230}$ ratio) for purity estimation. DNA samples with a fragment size above 50 kb, $OD_{260/280}$ ratio close to 2 and $OD_{260/230}$ ratio above 1.5 were kept.

Brassica rapa and *Brassica 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 GmbH & Co. KG) 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 of saturated phenol (25:24:1) was added to the samples and the tubes were placed on a rotator at 40 rpm for 10 min to get a fine emulsion. The samples were centrifuged for 24 min at 4500g (Acc3/Dec3) and the aqueous phases were poured into new 50 ml tubes containing phase-lock gel. Subsequently, 10 ml of chloroform-octanol (24:1) was added to each sample and the tubes were placed on a rotator at 40 rpm for 10 min to get a fine emulsion. The samples were centrifuged for 24 min at 4500g (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 TE 10/1 buffer. The extract quality was evaluated using Field Inverted Gel Electrophoresis (FIGE) 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 100–1500-bp size range using a Covaris E220 sonicator (Covaris, Woburn, MA, USA). The fragments were end-repaired and 3′-adenylated, and Illumina adapters were added using the Kapa Hyper Prep Kit (KapaBiosystems, Wilmington, MA, USA). The ligation products were purified with AMPure XP beads (Beckmann Coulter Genomics, Danvers, MA, USA). The libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems), and the library profiles were assessed using a DNA High Sensitivity LabChip kit on an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The libraries were sequenced on an Illumina HiSeq2500 instrument (Illumina, San Diego, CA, USA) 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., Liege, Belgium) was first size-selected using a BluePippin (Sage Science, Beverly, MA, USA). The selected DNA fragments were end-repaired and 3′-adenylated with the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). The DNA was then purified with AMPure XP beads (Beckmann Coulter, Brea, CA, USA) and ligated with sequencing adapters provided by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK) using 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.

Nanopore Sequencing (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, Ipswich, MA, USA) 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 Oxford Nanopore Technologies using Concentrated T4 DNA Ligase 2M U/ml (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 8.0, 10 mM EDTA, pH 8.0, 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% TritonX-100 (NIBT) and 7.5% 2-mercaptoethanol (NIBTM). The nuclei suspensions were filtered through cheese cloth and Mira cloth and centrifuged at 1500g 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 57g for 2 min at 4°C. The supernatants were kept and centrifuged at 1500g 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 of NIBT and a final time with 50 ml of NIB. The pellets were suspended in residual NIB (approximately 200 µl), transferred to a 1.5 ml tube and centrifuged at 1500g 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 60g 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 400g for 15min at 4°C. The pellet was suspended in residual buffer and adjusted to 35 ml. The nuclei suspension was centrifuged once more at 400g for 15 min at 4°C. The pellet was suspended in residual buffer and adjusted to 35 ml. The nuclei suspension was centrifuged at 200g and the supernatant was kept and centrifuged at 400g for 10 min at 4°C. The nuclei were suspended in HB+ buffer (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 NLRS labelling (BspQI) protocols were performed for *B. rapa* Z1, *B. oleracea* HDEM and *M. schizocarpa* according to Bionano with 600, 300 and 191.2 ng of DNA, respectively. For the DLS labelling (DLE-1), *M. schizocarpa* DNA was concentrated by evaporation at room temperature. All DLS labelling was performed with 750 ng of 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[34]. The last step identified and discarded read pairs that mapped to the phage phiX genome, using SOAP[35] 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[36] for each dataset to detect potential contamination.

Long-read genome assemblies

We used the Ra[37], SMARTdenovo[38] 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[39] software (Supplementary Tables 3–5), as it has been proven that down-sampling the read coverage can be beneficial for the assembly phase[6]. We also tried to use Canu[40] but could not get to the final assembly stage due 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 the reads and had contig N50s of respectively 7.3 and 3.8 Mb. In contrast, the best *M. schizocarpa* assembly (N50 of 4.0 Mb) was obtained using a 30x 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[41] software and then three additional times using Illumina reads as input to the Pilon[42] 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 Direct Label and Stain (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 9-10). The nanopore contigs were then organized using the two Bionano maps (DLE and *Bsp*QI) 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

"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 60K array and a genetic map was constructed using the CarthaGène software[43]. A total of 6,528 markers were genetically mapped, totaling 817.3 cM. The sequence contexts of all SNP markers that were genetically mapped were blasted against our *B. oleracea* HDEM assembly to validate the quality of our assembly and 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 cM 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 Figure 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 Figure 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[44]. The two flanking contigs were joined if one alignment (score > 3000) was detected. This procedure resolved several negative gaps and improved the contig N50 (Supplementary Table 11).

Transposable element annotation

Transposable elements (TEs) and, more generally, TE-rich regions are under-represented in short-read assemblies. To investigate this aspect we performed TE detection for our three genomes and the three available reference assemblies. TEs were annotated using RepeatMasker[45] (with default parameters, taxon viridiplantae for *M. schizocarpa* and eudicotyledons for *Brassica*) and TE libraries. The TE database generated in [46] was used to annotate the *B. oleracea* and *B. rapa* genomes. The TE database for *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 15-16)

Gene prediction

Gene prediction was done using proteomes from homologous species. For *B. rapa*, we used the following three proteomes: *B. rapa* (UP000011750), *B. 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), *O. sativa* (UP000059680) and *P. dactylifera* (UP000228380).

Low complexity in protein sequences was masked with the SEG algorithm. Low complexity in genomic sequences was masked using the DustMasker algorithm[47]. Tandem repeats were masked using Tandem Repeat Finder[48]. The TEs 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[44] (default parameters) was used to quickly localize corresponding putative genes of the proteins on the genome. The best match and matches with a score ≥ 90% (70% for *A. thaliana* proteins) of the best match score were retained. Second, the alignments were refined using Genewise[49] (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[50] to predict gene structures. This tool can find CDSs 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 (ORFs) 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*[21] 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 any previous annotation.

Using this pipeline, 46,721, 61,279 and 32,809 genes models were predicted for *B. rapa*, *B. oleracea* and *M. schizocarpa*, respectively. We assessed the completeness of the annotation using BUSCO[51] (embryophyta dataset) 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 (AED) 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 organised 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. 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[52] or from scientific publications. This information is available in Supplementary File 2.

Comparison of ONT and PACBIO datasets

Six PACBIO datasets were downloaded from EBI-ENA and metrics were computed from the whole dataset with the following coverage: 127x (*Vigna angularis*, PRJDB3778), 231x (*Vitis vinifera*, PRJNA316730), 224x (*Rosa chinensis*, PRJNA413292), 125x (*Citrus maxima*, PRJNA318855), 168x (*Fragaria vesca*, PRJNA383733) and 283x (*Arabidopsis thaliana*, PRJNA237120). Likewise, metrics were computed from the whole ONT datasets: 79x (*B. rapa*), 51x (*M. schizocarpa*) and 36x (*B. oleracea*). All 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[53] package and dot[54], an interactive dot plot viewer, to generate genome–genome alignment dot plots (Supplementary Figure 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 18-19 and Figure 13). We retrieved the 65.7M and 188.0M 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 respectively 2,735 and 3,508 regions (larger than 1Kb and with a coverage of at least 10X) representing 5.26 Mb and 7.61 Mb and spread over all the chromosomes. We localized and annotated these regions as coding exon, intron or transposable element using bedtools (Supplementary Figures 14-15 and Table 20), and observed that nearly 20% of the bases (representing 1.14 Mb and 1.54 Mb for HDEM and Z1) were annotated as genic regions (intron + exon).

We computed Best Reciprocal Hits (BRH) 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 10 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 respectively 25, 951 and 658 genes. Synteny visualisations were performed using the MCScan tool, obtained from "https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)" (Supplementary Figures 7-9).

Estimation of sequencing costs

We computed the sequencing costs for each genome using the following public prices: $500 and $2,200 for each MinION and PromethION flowcell, $2,146 for one-third of a HiSeq2500 flowcell (2×250 bp), $175 for nanopore long-read library preparation, $51 for Illumina PCR-Free library preparation, $1,500 for a single BioNano Genomics chip and $474 for the library preparation for two optical maps (BspQI and DLE-1). Given these costs, we estimated sequencing costs of $13,621, $12,271 and $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 $6,546.

Analysis of Flowering Locus C (FLC) genes

The FLC genes were searched using blastp and a database of FLC genes composed of predicted genes from a previous study[22] and the pan-genome of *B. oleracea*[21]. 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 3 FLC1, 1 disrupted FLC2, 1 FLC3 and 1 FLC5 genes for *B. oleracea* HDEM and 1 FLC1, 1 FLC2, 1 FLC3 and 1 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 analyzing phylogenetic relationships between molecular sequences (Supplementary Figure 17).

Identification and analysis of the self-incompatibility

locus in the two *Brassica* genomes

Self-incompatibility (SI) emerged as an evolutionary strategy to foster genetic diversity and exchange in plant species. In Brassicaceae, SI is controlled by a single multiallelic locus (named the S-locus). We first identified the self-incompatibility alleles within each of the *Brassica* genomes with an unpublished S-locus (self-incompatibility 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* paralog, 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 a custom R-script, 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 Figure 19).

Analysis of assembly completeness of disease resistance-like gene (R-gene) clusters

Three orthologous disease resistance-like genes clusters identified in DH-Pahang assembly version 1[16] were searched in DH-Pahang version 2[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.

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.

Additional files

All supporting data are included as a single additional file that contains Tables 1–21 and Figures 1–19 (Supplementary File 1). Detailed information about the 105 plant genome assemblies is available as a separate additional excel file (Supplementary File 2).

References

1 Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).

2 The map-based sequence of the rice genome. *Nature* **436**, 793-800, doi:10.1038/nature03895 (2005).

3 Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050-+, doi:10.1038/Nmeth.4035 (2016).
4 Jiao, W. B. & Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. *Current opinion in plant biology* **36**, 64-70, doi:10.1016/j.pbi.2017.02.002 (2017).
5 Michael, T. P. *et al.* High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature communications* **9**, 541, doi:10.1038/s41467-018-03016-2 (2018).
6 Schmidt, M. H. *et al.* De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant cell* **29**, 2336-2348, doi:10.1105/tpc.17.00521 (2017).
7 Du, H. *et al.* Sequencing and de novo assembly of a near complete indica rice genome. *Nature communications* **8**, 15324, doi:10.1038/ncomms15324 (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, doi:10.1093/gigascience/gix124 (2018).
9 Dassanayake, M. *et al.* The genome of the extremophile crucifer Thellungiella parvula. *Nature genetics* **43**, 913-918, doi:10.1038/ng.889 (2011).
10 Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**, 763-768, doi:10.1038/nature08747 (2010).
11 Raymond, O. *et al.* The Rosa genome provides new insights into the domestication of modern roses. *Nature genetics*, doi:10.1038/s41588-018-0110-3 (2018).
12 Cheng, F. *et al.* Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. *Nature genetics* **48**, 1218-1224, doi:10.1038/ng.3634 (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, doi:10.1016/j.molp.2016.11.008 (2017).
14 Wang, X. W. *et al.* The genome of the mesopolyploid crop species Brassica rapa. *Nature genetics* **43**, 1035-U1157, doi:10.1038/ng.919 (2011).

15 Parkin, I. A. *et al.* Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome biology* **15**, R77, doi:10.1186/gb-2014-15-6-r77 (2014).
16 D'Hont, A. *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213-+, doi:10.1038/nature11241 (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, doi:10.1186/s12864-016-2579-4 (2016).
18 Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature biotechnology* **30**, 771-776, doi:10.1038/nbt.2303 (2012).
19 Sakai, H. *et al.* The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Scientific reports* **5**, doi:Artn 1678010.1038/Srep16780 (2015).
20 Wang, X. *et al.* Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nature genetics* **49**, 765-+, doi:10.1038/ng.3839 (2017).
21 Golicz, A. A. *et al.* The pangenome of an agronomically important crop plant Brassica oleracea. *Nature communications* **7**, 13390, doi:10.1038/ncomms13390 (2016).
22 Schranz, M. E. *et al.* Characterization and effects of the replicated flowering time gene FLC in Brassica rapa. *Genetics* **162**, 1457-1468 (2002).
23 Goubet, P. M. *et al.* Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. *PLoS genetics* **8**, e1002495, doi:10.1371/journal.pgen.1002495 (2012).
24 Shiba, H. *et al.* Genomic organization of the S-locus region of Brassica. *Bioscience, biotechnology, and biochemistry* **67**, 622-626, doi:10.1271/bbb.67.622 (2003).
25 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, doi:10.1534/g3.117.300467 (2018).
26 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, doi:10.1007/s10681-016-1681-7 (2016).
27 Yang, J. H. *et al.* The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nature genetics* **48**, 1225-1232, doi:10.1038/ng.3657 (2016).
28 Jarvis, D. E. *et al.* The genome of Chenopodium quinoa. *Nature* **542**, 307-312, doi:10.1038/nature21370 (2017).
29 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 research* **27**, 778-786, doi:10.1101/gr.213652.116 (2017).
30 Reyes-Chin-Wo, S. *et al.* Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature communications* **8**, 14953, doi:10.1038/ncomms14953 (2017).
31 Teh, B. T. *et al.* The draft genome of tropical fruit durian (Durio zibethinus). *Nature genetics* **49**, 1633-1641, doi:10.1038/ng.3972 (2017).
32 Gawel, N. J. J., R.L. A modified CTAB DNA extraction procedure for Musa and Ipomoea. *Plant Molecular Biology Reporter* **9**, 262-266, doi:https://doi.org/10.1007/BF02672076 (1991).
33 Risterucci A.M. , G. L., N'Goran J.A.K., Pieretti L., Flament M.H., Lanaud C. A high-density linkage map of Theobroma cacao L. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **101**, 948-955 (2000).
34 Engelen, S., Aury JM. *Fastxtend*, <http://www.genoscope.cns.fr/fastxtend/> (2015).
35 Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714, doi:10.1093/bioinformatics/btn025 (2008).
36 Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research* **26**, 1721-1729, doi:10.1101/gr.210641.116 (2016).
37 Ra assembler v. git commit 65bedfe.
38 SMARTdenovo v. git commit 3d9c22e.
39 FitLong v. git commit 8d81024.
40 Koren, S. *et al.* Canu: scalable and accurate long-read assembly

via adaptive k-mer weighting and repeat separation. *Genome research* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).
41 Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**, 737-746, doi:10.1101/gr.214270.116 (2017).
42 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).
43 de Givry, S., Bouchez, M., Chabrier, P., Milan, D. & Schiex, T. CARHTA GENE: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* **21**, 1703-1704, doi:10.1093/bioinformatics/bti222 (2005).
44 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664, doi:10.1101/gr.229202 (2002).
45 RepeatMasker Open-4.0, http://www.repeatmasker.org (2013).
46 Chalhoub, B. *et al.* Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345**, 950-953, doi:10.1126/science.1253435 (2014).
47 Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology* **13**, 1028-1040, doi:10.1089/cmb.2006.13.1028 (2006).
48 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573-580 (1999).
49 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome research* **14**, 988-995, doi:10.1101/gr.1865504 (2004).
50 Dubarry M., N. B., Rukwavu T., Farhat S., Da Silva C., Seeleuthner Y., Lebeurrier M, Aury JM. Gmove a tool for eukaryotic gene predictions using various evidences (poster). *F1000 research*, doi:10.7490/f1000research.1111735.1 (2016).
51 Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution*, doi:10.1093/molbev/msx319 (2017).
52 website, K. *https://www.kew.org/*, <https://www.kew.org/> (
53 Marcais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* **14**, e1005944, doi:10.1371/journal.pcbi.1005944 (2018).

54 Dot.
55 Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research* **36**, W465-469, doi:10.1093/nar/gkn180 (2008).

Acknowledgments

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). The authors are grateful to Oxford Nanopore Technologies Ltd. for early access to the MinION device through the MAP, and we thank their staff for technical help. Work by X. Vekemans and M. Genete 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

CF, GD, FCB, ED and CC extracted the DNA. CC and AL optimized and performed the sequencing. ED, WB and VB generated the optical maps. PD, RD and MMD generated the genetic map for the *Brassica oleracea* HDEM accession. BI, CB and JMA performed the genome assemblies. GM performed the anchoring of the *Musa schizocarpa* scaffolds. CF, JM and MRG performed the anchoring of the *Brassica oleracea* scaffolds. MD and JMA performed the anchoring of the *Brassica rapa* scaffolds. MD and BN performed the gene prediction for the genome assemblies. BI, CB, MD, FD, JMA and SE performed the bioinformatic analyses. XV and MG performed the S-locus annotation of the two Brassicaceae genomes. BI, CB, MD and JMA wrote the article. ADH, AMC, PW and JMA supervised the study.

Competing interests

The authors declare that they have no competing interests. BI, SE, CC, PW, and JMA are part of the MinION Access Programme (MAP) and JMA received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences.

9.2.4 Conclusion

La combinaison de trois technologies (ONT, Bionano Genomics et Illumina) permet de produire des assemblages de génome de haute qualité à un coût relativement faible. Les trois génomes de référence produits ont une meilleure continuité que les génomes déjà réalisés sur des espèces parentes et atteignent une taille proche de la taille estimée. L'annotation montre un contenu en gènes proche de celui déjà déterminé dans les versions réalisées à partir de courtes lectures mais une réelle amélioration de la caractérisation du contenu en éléments transposables et en éléments répétés. Par exemple, les loci S des deux *Brassica* ont pu être entièrement annotés. Le séquençage longues lectures a ainsi permis de mieux caractériser la structure de ces trois génomes, ajoutant des données dans des régions non assemblées et ouvrant des nouvelles perspectives d'analyse.

Cet article valorise tous les développements méthodologiques que j'ai réalisés pour la mise en place de la technologie Bionano Genomics au Genoscope. Il fut le premier à mettre en avant l'utilisation de cartes réalisées à partir de l'enzyme DLE-1 que nous avions testé en avant première. Cette nouvelle enzyme a lancé le succès de cette technologie pour la reconstruction des chromosomes à partir des assemblages longues lectures.

Au moment de la publication de cet article, des améliorations étaient encore nécessaires, notamment dans les méthodes d'extraction d'ADN de haut poids moléculaire. En effet, la taille des fragments d'ADN est un élément clé pour obtenir les lectures les plus longues possibles et aider à l'élucidation des régions complexes. Le taux d'erreur encore élevé des longues lectures produites lors du séquençage ONT rend obligatoire l'utilisation de séquences courtes lectures de type Illumina pour corriger les assemblages. De même, les cartes optiques ou la capture de la conformation des chromosomes sont essentielles pour proposer des assemblages à l'échelle du chromosome pour les grands génomes de plantes.

J'ai participé par la suite à la réalisation du génome de référence de *Brassica napus* Darmor-bzh version 10, colza ancestralement issu du croisement de *Brassica rapa* et *Brassica oleracea*, venant compléter les

génomes des deux progéniteurs. Cette réalisation a également fait l'objet d'une publication que j'ai coécrite, parue dans la revue *GigaScience*[113]. Les trois génomes sont une richesse pour les études sur les différentes variétés de colza cultivées et sur la compréhension de la dynamique de ces génomes (Figure 27b). Par exemple, dans une étude à laquelle j'ai également participé, Boideau *et al*. utilisent les génomes de référence pour montrer que la méthylation de l'ADN ainsi que les variants structuraux sont des facteurs importants qui façonnent la recombinaison méiotique chez *B. napus*[117]. Ces deux facteurs peuvent avoir des impacts positifs ou négatifs considérables sur la sélection végétale. Il propose d'assembler et de comparer les génomes des variétés donneuses et receveuses avant de lancer un programme de sélection de longue durée.

9.3 AMELIORATION DES ASSEMBLAGES

Des développements technologiques et méthodologiques sont nécessaires pour tendre vers une qualité toujours accrue. Utiliser les technologies de pointe demande des adaptations afin de les rendre applicables à toutes sortes de génomes. Les outils fournis par les sociétés commercialisant ces technologies montrent parfois certaines faiblesses pour une utilisation sur des génomes très complexes. Il faut alors développer nos propres outils pour pallier ces biais observés.

9.3.1 Développement d'un outil dédié à la correction du scaffolding par carte optique

9.3.1.1 Introduction

La société Bionano Genomics commercialise l'instrument Saphyr permettant de générer des cartes optiques. Des outils bioinformatiques (suite de logiciels appelée "bionano solve") sont fournis pour réaliser l'assemblage des molécules générées en carte optique ainsi que le scaffolding des assemblages, obtenus à partir de longues lectures, grâce à la carte optique.

Or, j'ai pu observer des artéfacts de scaffolding dans différents cas de figure conduisant à la production de gaps supplémentaires et à de la

fausse duplication. Ils sont le résultat d'un fonctionnement très conservatif des outils. Seules les erreurs d'assemblage sont traitées lors du processus de scaffolding. Les contigs considérés comme chimériques par comparaison avec la carte optique sont cassés. Par contre, il peut arriver que les logiciels d'assemblage ne rassemblent pas des contigs même si leurs extrémités sont similaires et auraient donc pu être fusionnées pour générer un contig plus long. Les outils de scaffolding de Bionano Genomics préfèrent conserver les deux contigs en plaçant un trou de 13 bases indéterminées créant ainsi un gap et une fausse duplication (Figure 28). Dans le cas des génomes hétérozygotes, des petits contigs correspondant au deuxième haplotype sont générés. Ils peuvent être également ajoutés lors du processus de scaffolding induisant la présence des deux haplotypes les uns à la suite de l'autre et générant également une fausse duplication. Enfin, un contig pouvant être inclus dans un plus grand contig est parfois placé à la suite de celui-ci créant une erreur de scaffolding. J'avais réalisé des corrections de scaffolding de façon manuelle en repérant les différents cas de figure grâce à l'interface web Bionano Access. Elle permet de visualiser les alignements entre la carte optique et les contigs qui forment les hybrides scaffolds. Le développement d'un outil, BiSCoT (Bionano Scaffolding Correction Tool) était nécessaire pour venir corriger spécifiquement ces artéfacts de façon automatisée. Il prend en charge les fichiers produits par les outils Bionano Genomics, détecte les régions problématiques et fournit un

fichier d'assemblage corrigé. Il est présenté dans l'article suivant.

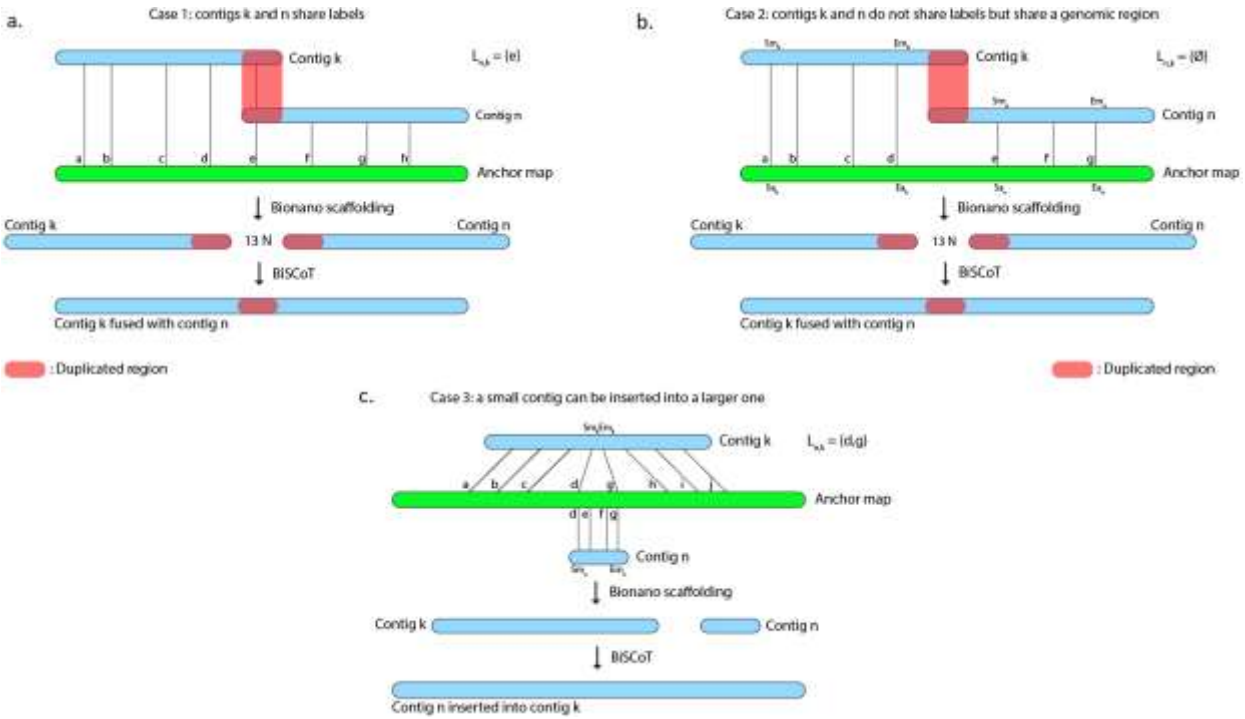


Figure 28: ***Cas d'artefacts corrigés par BiSCoT.*** *a. et b. gestion des cas de fausse duplication. c. gestion de l'insertion d'un contig à l'intérieur d'un autre contig*

9.3.1.2 Article : "BiSCoT: improving large eukaryotic genome assemblies with optical maps." Istace B, Belser C, Aury JM. PeerJ. 2020

BiSCoT: improving large eukaryotic genome assemblies with optical maps

Benjamin Istace, Caroline Belser and Jean-Marc Aury

Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

ABSTRACT

Motivation. Long read sequencing and Bionano Genomics optical maps are two techniques that, when used together, make it possible to reconstruct entire chromosome or chromosome arms structure. However, the existing tools are often too conservative and organization of contigs into scaffolds is not always optimal.

Results. We developed BiSCoT (Bionano SCaffolding COrrection Tool), a tool that post-processes files generated during a Bionano scaffolding in order to produce an assembly of greater contiguity and quality. BiSCoT was tested on a human genome and four publicly available plant genomes sequenced with Nanopore long reads and improved significantly the contiguity and quality of the assemblies. BiSCoT generates a fasta file of the assembly as well as an AGP file which describes the new organization of the input assembly.

Availability. BiSCoT and improved assemblies are freely available on GitHub at http://www.genoscope.cns.fr/biscot and Pypi at https://pypi.org/project/biscot/.

Subjects Bioinformatics, Genomics

Keywords Genome assembly, Bioinformatics, Tool, Scaffolding, Optical maps, Bionano, Long reads, Nanopore, PacBio

INTRODUCTION

Assembling large and repetitive genomes, such as plant genomes, is a challenging field in bioinformatics. The appearance of short reads technologies several years ago improved considerably the number of genomes publicly available. However, a high proportion of them are still fragmented and few represent the chromosome organization of the genome. Recently, long reads sequencing techniques, like Oxford Nanopore Technologies and Pacific Biosciences, were introduced to improve the contiguity of assemblies, by sequencing DNA molecules that can range from a few kilobases to more than a megabase in size (*Istace et al., 2017*; *Schmidt et al., 2017*; *Kim et al., 2019*; *Shafin et al., 2019*). Nevertheless and even if the assemblies were greatly improved, the chromosome-level organization of the sequenced genome cannot be deciphered in a majority of cases. In 2017, Bionano Genomics launched its Saphyr system which was able to generate optical maps of a genome, by using the distribution of enzymatic labelling sites. These maps were used to orient and order contigs into scaffolds but the real improvement came in 2018, when Bionano Genomics introduced their Direct Label and Stain (DLS) technology that was able to produce genome maps at the chromosome-level with a N50 several times higher than previously (*Belser et al., 2018*; *Formenti et al., 2018*; *Hu et al., 2019*).

Submitted 19 December 2019
Accepted 21 September 2020
Published 5 November 2020

Corresponding author
Benjamin Istace,
bistace@genoscope.cns.fr

Academic editor
Alexander Schliep

Additional Information and Declarations can be found on page 6

DOI **10.7717/peerj.10150**

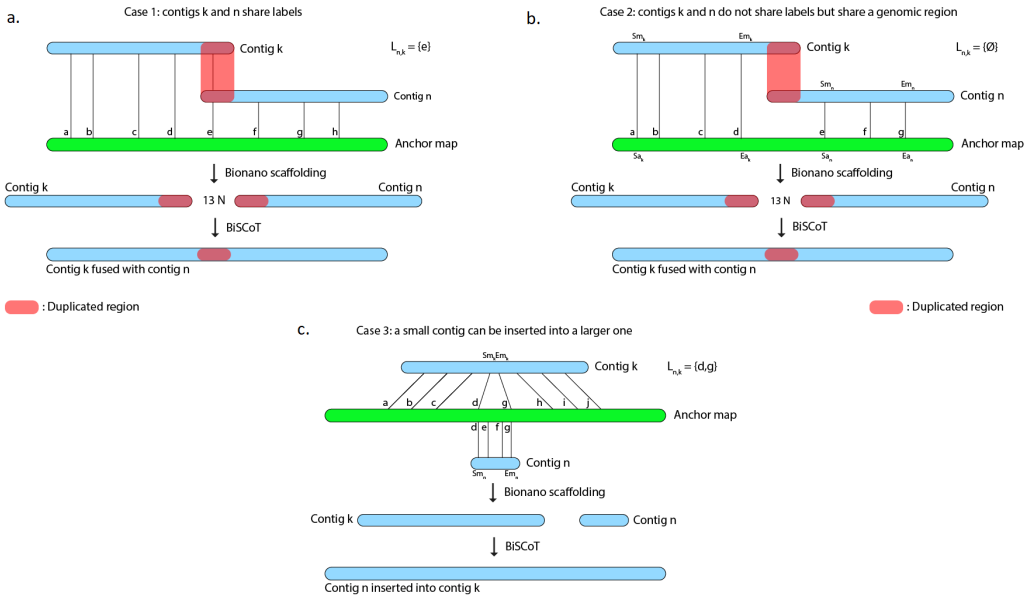


Figure 1 The Bionano scaffolding tool does not merge contigs even if they share labels. Instead, it inserts 13 N's gap between contigs, thus artificially duplicating the shared region. (A) BiSCoT merges contigs that share enzymatic labelling sites. (B) If contigs do not share labels but share a genomic region, BiSCoT attempts to merge them by aligning the borders of the contigs. (C) The Bionano scaffolding tool does not handle cases where contigs can be inserted into others. BiSCoT attempts to merge the inserted map with the one containing it if they share labels.

Full-size DOI: 10.7717/peerj.10150/fig-1

However, scaffolds generated with the tool provided by Bionano Genomics do not reach optimal contiguity. Indeed, when two contigs C_1 and C_2 are found to share labels, one could expect that the tool would merge the two sequences at the shared site. Instead, the software chooses a conservative approach and outputs the sequence of C_1 followed by a 13-Ns gap and then the C_2 sequence, thus duplicating the region that is shared by the two contigs (Fig. 1A and 1B) and in numerous cases, these duplicated regions could reach several kilobases. As an example, on the human genome we used to evaluate BiSCoT (see 'Results'), we could detect 515 of those regions, affecting 16 genes and corresponding to around 24.5 Mb of duplicated sequences, the longest being 237 kb in size. These duplicated regions affect the contiguity and have to be corrected as they can be problematic for downstream analyses, like copy number variation studies. They originate from overlaps that are not fused in the input assembly and usually correspond to allelic duplications. In addition, contigs can sometimes be inserted into other contigs, these cases are not handled by the Bionano scaffolding tool that discards the inserted contigs (Fig. 1C).

We developed BiSCoT, a python script that examinates data generated during a previous Bionano scaffolding and merges contigs separated by a 13-Ns gap if needed. BiSCoT also re-evaluates gap sizes and searches for an alignment between two contigs if the gap size is inferior to 1,000 nucleotides. BiSCoT is therefore not a traditional scaffolder since it can only be used to improve an existing scaffolding, based on an optical map.

METHODS

Mandatory files loading

During the scaffolding, the Bionano scaffolder generates a visual representation of the hybrid scaffolds that is called an 'anchor'. It also generates one 'key' file, which describes the mapping between map identifiers and contig names, several CMAP files, which contain the position of enzymatic labelling sites on contig maps and on the anchor, and a XMAP file, that describes the alignment between a contig map and an anchor. BiSCoT first loads the contigs into memory based on the key file. Then, the anchor CMAP file and contig CMAP files are loaded into memory. Finally, the XMAP file is parsed and loaded.

Scaffolding

Alignments of contigs onto anchors contained in the XMAP file are first sorted by their starting position on the anchor. Then, alignments on one anchor are parsed by pairs of adjacent contigs, i.e alignment of contig C_k is examined at the same time as contig C_n, with C_k aligned before C_n on the anchor. Aligned anchor labels are extracted from these alignments and a list of shared labels $L_{n,k}$ is built. For the following cases, we suppose C_k and C_n to be aligned on the forward strand (Fig. 1).

Case 1: contig maps share at least one anchor label

The last label l from $L_{n,k}$ is extracted and the position P_l of l on both contigs C_k and C_n is recovered from the CMAP files. In the resulting scaffold, the sequence of C_k will be included up to the P_l position and the sequence of C_n will be included from the P_l position. In this case, the gap is removed, both contigs C_k and C_n are fused and BiSCoT generates a single contig instead of two contigs initially separated by a gap in the input assembly.

Case 2: contig maps do not share anchor labels

Let $Size_k$ be the size of the contig C_k, Sm_k and Em_k the start and end of an alignment on a contig map and Sa_k and Ea_k the corresponding coordinates on the anchor. The number n of bases between the last aligned label of C_k and the first aligned label of C_n is then:

$$n = Sa_n - Ea_k \tag{1}$$

We then have to subtract the part d_k of C_k after the last aligned label of C_k and the part d_n of C_n before the first aligned label of C_n:

$$d_k = Size_k - Em_k \tag{2}$$

$$d_n = Sm_n \tag{3}$$

Finally, we can compute the gap size g with:

$$g = n - d_k - d_n \tag{4}$$

If $g \leq 1000$, a BLAT (*Kent, 2002*) alignment of the last 30 kb of C_k is launched against the first 30kb of C_n. If an alignment is found and if its score is higher than 5,000, C_k and C_n are merged at the starting position of the alignment and, as in case1, BiSCoT generates a single contig instead of two contigs initially separated by a gap in the input assembly. Otherwise, a number g of Ns is inserted between C_k and C_n.

Case 3: insertion of small contigs

Let Sm_k and Em_k the start and end of an alignment on a contig map. If $[Sm_n, Em_n] \subset [Sm_k, Em_k]$, then the left-most shared label identifier l_l and right-most shared label identifier l_r are extracted. If C_n has more of its labels mapped in this region than C_k, the sequence of C_n will be inserted between l_l and l_r in the scaffolds. Otherwise, the sequence of C_k remains unchanged and C_n will be included as a singleton sequence in the scaffolds file.

Finally, if an Illumina polishing step was done before or after Bionano scaffolding, we recommend doing one additional round of polishing using Illumina reads after BiSCoT has been applied. Indeed, short reads tend to be aligned only against one copy of the duplicated regions, leaving the other copy unpolished.

RESULTS AND DISCUSSIONS

Validation on simulated data

In order to simulate a genome assembly, we downloaded the chromosome 1 of the GRCh38.p12 human reference genome and fragmented it to create contigs. We generated 120 contigs with an N50 size of 2.4 Mb and a cumulative size of 231 Mb. Contigs were generated with either overlaps or gaps between them. We introduced 50 gaps with a mean length of 50 kb, the smallest being 3.4kbp long and the largest 99.6 kb long, and 50 overlaps with a mean size of 44kb, the smallest being 278b long and the largest 98.6 kb long. We also generated five contigs, with an N50 of 254 kb, that were subsequences of larger contigs, to simulate contained contigs.

Then, we used these contigs and Bionano DLE and BspQI optical maps available on the Bionano Genomics website as input to the Bionano scaffolder. We gave the results of this scaffolding to BiSCoT and aligned all assemblies to the chromosome 1 reference using Quast (*Gurevich et al., 2013*, v5.0.2).

BiSCoT was able to resolve 39 overlaps out of the 50 we introduced (Table S1), 31 using shared labels and 8 using a Blat alignment. The 11 remaining overlaps could not be resolved due to contigs not sharing enough labels or the overlap being too small to produce an alignment of sufficient confidence. BiSCoT was also able to integrate all contained contigs back to their original place in the assembly. Furthermore, BiSCoT did not close any of the real gaps introduced during the assembly generation.

Regarding assembly metrics (Table S2), The N50 decreased by 1.4% in scaffolds and increased by 22% in contigs. The number of Ns in scaffolds decreased from 20.7Mb to 20.4Mb. Moreover, the number of misassemblies decreased by 68% after applying BiSCoT and the duplication ratio estimated by Quast decreased from 1.026 in Bionano scaffolds to 1.021 in BiSCoT scaffolds.

In order to estimate the accuracy of gap sizes, we compared the gap sizes we introduced in the input assembly to the ones that were estimated using optical maps (Fig. S1). We found that estimated gap sizes were very close to the reality, with a mean scaled absolute error of 0.8%.

Table 1 Metrics of the NA12878 scaffolds and contigs before or after BiSCoT treatment. Bold formatting indicates the best scoring assembly among contigs.

	Nanopore contigs	Bionano		BiSCoT	
		Contigs	Scaffolds	Contigs	Scaffolds
Cumulative size	2,818,937,673	2,818,997,568	2,878,230,106	2,810,480,725	2,868,077,379
N50	11,821,944	10,566,783	86,858,024	**12,894,141**	86,833,728
L50	67	71	14	**64**	14
N90	2,143,851	1,863,173	26,054,782	**2,321,940**	26,037,000
L90	280	301	36	**254**	36
auN[a]	15,164,719	14,547,428	82,760,251	**15,977,835**	82,474,548
# Ns	0	0	59,232,538	0	57,596,654
NGA50	5,794,944	5,729,014	10,816,842	**6,360,576**	11,713,900
NGA75	1,511,206	1,495,174	2,701,541	**1,596,102**	2,938,187
# misassemblies	1,356	1,299	1,602	**1,278**	1,515
Complete BUSCOs	**235 (92.2%)**	234 (91.8%)	231 (90.6%)	**235 (92.2%)**	231 (90.6%)
Duplicated BUSCOs	5 (2.0%)	**4 (1.6%)**	4 (1.6%)	**4 (1.6%)**	4 (1.6%)
Missing BUSCOs	11 (4.3%)	**10 (3.9%)**	13 (5.1%)	**10 (3.9%)**	13 (5.1%)

Notes.
[a]auN is a new metric to measure assembly contiguity *Li (2020)*.

Validation on real data

We downloaded genome assemblies for which a DLE optical map was available: the NA12878 human genome (*Jain et al., 2018*), *Brassica oleracea* HDEM (PRJEB26621, *Belser et al., 2018*), *Brassica rapa* Z1 (PRJEB26620, *Belser et al., 2018*), *Musa schizocarpa* (PRJEB26661, *Belser et al., 2018*) and *Sorghum bicolor* Tx430 (PRJNA472170, *Deschamps et al., 2018*). The QUAST and BUSCO (*Simão et al. (2015)*, v4.0.5) tools were used respectively to evaluate the number of misassemblies to the GRCh38.p12 human reference genome and the number of conserved genes among eukaryotes. In all cases, we first used the Bionano workflow to scaffold the draft assembly and launched BiSCoT using the files generated by the Bionano tools (Table 1, Tables S3–S6). The output of the Bionano workflow and BiSCoT are scaffolds, but we generated a contig file for each assembly by splitting each scaffold at every position with at least one N.

Concerning the NA12878 genome, we could detect 515 overlapping regions with a mean size of 47kb and representing in total 24.5 Mb of duplicated sequences. Among these 515 regions, 499 were corrected by BiSCoT using either shared labels (113 regions) or a BLAT alignment (386 regions) when no shared labels were found.

Globally, the contig NX and NGAX metrics increased drastically: the contigs NGA50 of NA12878 increased by around 10%, going from 5.8 Mb to 6.3 Mb. The scaffolds NGAX metrics also increased: the scaffolds NGA50 increased from 10.8 Mb in Bionano scaffolds to 11.7Mb in BiSCoT scaffolds. Moreover, the number of Ns decreased marginally and the number of complete eukaryotic genes stayed the same in scaffolds. More importantly, when aligning the assemblies against the reference genome, we could detect a decrease in the number of mis-assemblies going from 1,602 in Bionano scaffolds to 1,515 in BiSCoT scaffolds. The same kind of results were observed in the four plant genomes with a slight

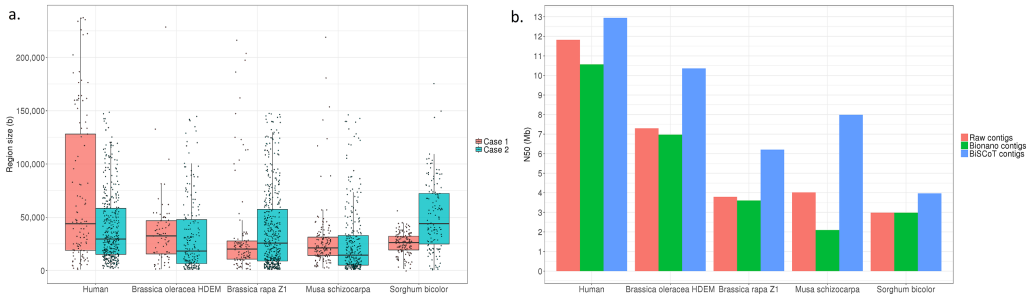


Figure 2 **(A) Distribution of the sizes of overlapping regions in the raw assemblies. Detection was done using either Bionano labels (Case 1) or a BLAT alignment (Case 2). (B) N50 contigs of raw assemblies and assemblies before or after BiSCoT treatment.**

Full-size DOI: 10.7717/peerj.10150/fig-2

decrease in scaffolds NX metrics and number of Ns but an increase in contigs NX metrics (Fig. 2 and Tables S2–S5).

SUMMARY

Thanks to the advent of long reads and optical maps technologies, it is now possible to obtain high-quality chromosome-scale assemblies. However, the official Bionano scaffolding tool does not always perform optimally when joining two contigs. Indeed, it does not merge two sequences when they share a genomic region, creating artificial gaps in the assembly. We developed BiSCoT, a tool that corrects these problematic regions in a prior Bionano scaffolding and showed that it increased significantly contiguity metrics of the resulting assembly, while preserving its quality.

ACKNOWLEDGEMENTS

The authors are grateful to the Bionano Genomics staff for technical help and would also like to thank the Whole Human Genome Sequencing Project for providing access to the Nanopore human genome assembly.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

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). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The Genoscope, the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA).
France Génomique: ANR-10-INBS-09-08.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Benjamin Istace conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Caroline Belser conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Jean-Marc Aury conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Data used and code are available at GitHub:
http://www.genoscope.cns.fr/biscot.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10150#supplemental-information.

REFERENCES

Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, Genete M, Berrabah W, Chèvre A-M, Delourme R, Deniot G, Denoeud F, Duffé P, Engelen S, Lemainque A, Manzanares-Dauleux M, Martin G, Morice J, Noel B, Vekemans X, D'Hont A, Rousseau-Gueutin M, Barbe V, Cruaud C, Wincker P, Aury J-M. 2018. Chromosome-scale assemblies of plant genomes using Nanopore long reads and optical maps. *Nature Plants* **4(11)**:879–887 DOI 10.1038/s41477-018-0289-4.

Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H. 2018. A chromosome-scale assembly of the Sorghum genome using Nanopore sequencing and optical mapping. *Nature Communications* **9(1)**:4844 DOI 10.1038/s41467-018-07271-1.

Formenti G, Chiara M, Poveda L, Francoijs K-J, Bonisoli-Alquati A, Canova L, Gianfranceschi L, Horner DS, Saino N. 2018. SMRT long reads and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (Hirundo rustica rustica). *GigaScience* **8(1)**: giy142 DOI 10.1093/gigascience/giy142.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29(8)**:1072–1075 DOI 10.1093/bioinformatics/btt086.

Hu L, Xu Z, Wang M, Fan R, Yuan D, Wu B, Wu H, Qin X, Yan L, Tan L, Sim S, Li W, Saski CA, Daniell H, Wendel JF, Lindsey K, Zhang X, Hao C, Jin S. 2019. The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nature Communications* **10(1)**:4702 DOI 10.1038/s41467-019-12607-6.

Istace B, Friedrich A, D'Agata L, Faye S, Payen E, Beluche O, Caradec C, Davidas S, Cruaud C, Liti G, Lemainque A, Engelen S, Wincker P, Schacherer J, Aury J-M. 2017. De novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *GigaScience* **6(2)**: giw018 DOI 10.1093/gigascience/giw018.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore sequencing and assembly of a Human genome with ultra-long reads. *Nature Biotechnology* **36**:338 DOI 10.1038/nbt.4060.

Kent W. 2002. BLAT–the BLAST-like alignment tool. *Genome Research* **12**:656–664 DOI 10.1101/gr.229202.

Kim H-S, Jeon S, Kim C, Kim YK, Cho YS, Kim J, Blazyte A, Manica A, Lee S, Bhak J. 2019. Chromosome-scale assembly comparison of the Korean Reference Genome KOREF from PromethION and PacBio with Hi-C mapping information. *GigaScience* **8(12)**: giz125 DOI 10.1093/gigascience/giz125.

Li H. 2020. auN: a new metric to measure assembly contiguity. *Available at https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity*.

Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Maß J, Pfaff C, Schurr U, Chetelat R, Maumus F, Aury J-M, Koren S, Fernie AR, Zamir D, Bolger AM, Usadel B. 2017. De novo assembly of a new solanum pennellii accession using nanopore sequencing. *The Plant Cell* **29(10)**:2336–2348 DOI 10.1105/tpc.17.00521.

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, Sedlazeck FJ, Marschall T, Mayes S, Costa V, Zook JM, Liu KJ, Kilburn D, Sorensen M, Munson KM, Vollger MR, Eichler EE, Salama S, Haussler D, Green RE, Akeson M, Phillippy A, Miga KH, Carnevali P, Jain M, Paten B. 2019. Efficient de novo assembly of eleven human genomes using PromethION sequencing and a novel Nanopore toolkit. *bioRxiv* DOI 10.1101/715722.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31(19)**:3210–3212 DOI 10.1093/bioinformatics/btv351.

9.3.1.3 Conclusion

L'utilisation de BisCoT, développé en python, permet donc de joindre les extrémités de contigs (artéfactuellement séparés par 13N) présentant des séquences communes et de supprimer les contigs dupliqués. Il permet également d'inclure des contigs dans de plus grands contigs si nécessaire. BiSCoT s'appuie sur la comparaison des sites de marquage entre la carte optique et les contigs digérés *in silico*. L'assemblage ainsi corrigé comporte moins de gaps et de régions faussement redondantes. Par conséquent, sa continuité augmente légèrement et la taille cumulée de l'assemblage diminue proportionnellement. BiSCoT est un outil spécifique des scaffoldings réalisés à partir des données de carte optique et ne peut être utilisé sur d'autres types de données. Nous l'utilisons en routine sur tous nos projets de production de génomes de référence. Il est disponible sur le github du Genoscope à l'adresse suivante : https://github.com/institut-de-genomique/biscot.

J'ai été invitée par la société Bionano Genomics à présenter BiSCoT lors d'une conférence d'utilisateurs au Sanger Institute. Avec notre accord, BiSCoT a été depuis partiellement inséré dans les outils délivrés par la société. En effet, la partie nécessitant un alignement d'une séquence contre l'autre n'étant pas possible avec leurs outils, seule la comparaison de positions de sites de coupure a été incluse. Les résultats sont donc un peu moins performants qu'avec BisCoT.

9.3.2 Tests méthodologiques pour générer les assemblages à l'échelle des chromosomes

9.3.2.1 Introduction

Sélectionner les stratégies les plus adaptées pour obtenir le génome d'un organisme d'intérêt demande de maîtriser les différents protocoles et outils. Il convient également de connaître les faiblesses et avantages de chaque stratégie et de savoir comment les combiner.

Dans le cadre des grands projets de production d'assemblage de génomes, il nous est apparu nécessaire de réaliser un retour d'expérience sur les procédures que nous avons mises en place autour des technologies de séquençage longues lectures (Oxford nanopore), de la génération des cartes optiques (Bionano Genomics) et du séquençage Hi-C.

En effet, le Genoscope est membre du projet européen ERGA et du programme ATLASea qui ont pour vocation de séquencer la diversité eucaryote européenne en donnant la priorité aux espèces en danger. Il est important d'utiliser les technologies de pointe et de standardiser nos procédures afin de les appliquer au plus grand nombre d'espèces.

J'ai donc coécrit un article portant sur la génération d'une version 2 du génome de référence de *Brassica rapa* cv. Z1. J'y ai comparé les résultats obtenus en utilisant les données de cartes optiques d'une part et de deux technologies Hi-C d'autre part: l'Omni-C, commercialisé par Dovetail Genomics et le Pore-C, développé pour réaliser un séquençage de banque Hi-C sur nanopore. J'ai ensuite présenté la possibilité de combiner ces deux technologies long range.

9.3.2.2 Article : « Sequencing and Chromosome-Scale Assembly of Plant Genomes, Brassica rapa as a Use Case » Istace B, Belser C, et al. Biology 2021

biology

MDPI

Article

Sequencing and Chromosome-Scale Assembly of Plant Genomes, *Brassica rapa* as a Use Case

Benjamin Istace [1,†], Caroline Belser [1,†], Cyril Falentin [2], Karine Labadie [3], Franz Boideau [2], Gwenaëlle Deniot [2], Loeiz Maillet [2], Corinne Cruaud [3], Laurie Bertrand [1], Anne-Marie Chèvre [2], Patrick Wincker [1], Mathieu Rousseau-Gueutin [2] and Jean-Marc Aury [1,*]

1 Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France; bistace@genoscope.cns.fr (B.I.); cbelser@genoscope.cns.fr (C.B.); lbertrand@genoscope.cns.fr (L.B.); pwincker@genoscope.cns.fr (P.W.)
2 IGEPP, INRAE, Institut Agro, Université de Rennes, Domaine de la Motte, 35653 Le Rheu, France; cyril.falentin@inrae.fr (C.F.); franz.boideau@inrae.fr (F.B.); gwenaelle.deniot@inrae.fr (G.D.); loeiz.maillet@inrae.fr (L.M.); anne-marie.chevre@inrae.fr (A.-M.C.); mathieu.rousseau-gueutin@inrae.fr (M.R.-G.)
3 Genoscope, Institut François Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France; klabadie@genoscope.cns.fr (K.L.); cruaud@genoscope.cns.fr (C.C.)
* Correspondence: jmaury@genoscope.cns.fr
† These authors contributed equally.

Simple Summary: Reconstructing plant genomes is a difficult task due to their often large sizes, unusual ploidy, and large numbers of repeated elements. However, the field of sequencing is changing very rapidly, with new and improved methods released every year. The ultimate goal of this study is to provide readers with insights into techniques that currently exist for obtaining high-quality and chromosome-scale assemblies of plant genomes. In this work, we presented the advanced techniques already existing in the field and illustrated their application to reconstruct the genome of the yellow sarson, *Brassica rapa* cv. Z1.

Abstract: With the rise of long-read sequencers and long-range technologies, delivering high-quality plant genome assemblies is no longer reserved to large consortia. Not only sequencing techniques, but also computer algorithms have reached a point where the reconstruction of assemblies at the chromosome scale is now feasible at the laboratory scale. Current technologies, in particular long-range technologies, are numerous, and selecting the most promising one for the genome of interest is crucial to obtain optimal results. In this study, we resequenced the genome of the yellow sarson, *Brassica rapa* cv. Z1, using the Oxford Nanopore PromethION sequencer and assembled the sequenced data using current assemblers. To reconstruct complete chromosomes, we used and compared three long-range scaffolding techniques, optical mapping, Omni-C, and Pore-C sequencing libraries, commercialized by Bionano Genomics, Dovetail Genomics, and Oxford Nanopore Technologies, respectively, or a combination of the three, in order to evaluate the capability of each technology.

Keywords: genome; assembly; scaffolding; chromosome-scale; nanopore; optical map; bionano; omni-C; pore-C; plants

check for updates

Citation: Istace, B.; Belser, C.; Falentin, C.; Labadie, K.; Boideau, F.; Deniot, G.; Maillet, L.; Cruaud, C.; Bertrand, L.; Chèvre, A.-M.; et al. Sequencing and Chromosome-Scale Assembly of Plant Genomes, *Brassica rapa* as a Use Case. *Biology* **2021**, *10*, 732. https://doi.org/10.3390/biology10080732

Academic Editors: Pierre Devaux and Pierre Sourdille

Received: 12 July 2021
Accepted: 28 July 2021
Published: 30 July 2021

1. Introduction

Assembling plant genomes has always been one of the most complex tasks in bioinformatics applied to genomics. Indeed, they often contain many repeated elements, such as satellites or transposable elements. This leads to an increase in the size of the genome, which can then reach tens of gigabases, for example the loblolly pine genome, which is 22 Gb in size [1]. Moreover, the difficulty is further increased due to the high levels of heterozygosity and highly variable ploidy [2]. All these characteristics make reconstruction

of such genomes almost impossible without the help of a large consortium. However, the appearance of Illumina [3] technology fifteen years ago, combined with significant sequencing depth and the advent of assemblers using De Bruijn graphs, paved the way for low-cost genome assemblies. However, the resulting assemblies remained highly fragmented, and most repeats were not resolved.

Recently, the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies' (ONT) single-molecule sequencing technologies have been commercialized, which offer the opportunity to sequence fragments of several tens of kilobases, thus facilitating the assembly of complex genomes. However, this increase in the size of the sequenced fragments has a cost in terms of the read's quality. Indeed, Nanopore and PacBio raw reads show an error rate of about 7% and 10%, respectively [4,5]. Due to errors remaining in long reads assemblies, a significant proportion of predicted genes may contain frameshifts [6], thus reducing the size of the predicted proteins, which in turn could cause problems in downstream analyses. To circumvent this issue, a number of polishing algorithms have surfaced. First, polishing algorithms were using the same type of data as was used to perform the assembly, as is the case in Nanopolish, for Nanopore data, or Quiver, for PacBio data. However, errors such as indels were still present in genome assemblies, and researchers quickly realized that pairing long reads data to generate the assembly and high-quality reads, such as the ones produced by Illumina sequencers, led to the least number of errors. This was, as an example, implemented in Pilon [7], Racon [8], or Hapo-G [9]. These gradual improvements in base quality and read length allowed researchers to generate high-quality plant genomes, reaching for the simplest (haploid genomes) chromosome scale without the need for long-range technologies [10,11].

Although the base quality of long reads has improved over the years, assembling and phasing heterozygous genomes still remains a difficult task. To try to solve this problem, PacBio drastically improved the quality of its generated sequencing reads by using a technique called Circular Consensus Sequencing (CCS) to generate high fidelity (HiFi) reads, thus breaking the 1% error rate barrier [12]. While standard PacBio reads are generated by a single pass of an enzyme around a circularized template, HiFi reads are produced by multiple passes of the enzyme through the same circularized sequences. This led to improvements in the base quality of the assemblies and made it possible to assemble and phase large genomes [13]. In a similar manner, ONT recently revealed their new Q20+ sequencing kit, which uses a new enzyme and optimized run conditions to lower the raw read error rate to 1%. It should be noted, however, that, at the moment of writing, we could not test the Q20+ kit and, therefore, could not validate the results shown by Oxford Nanopore.

However, this increase in read quality is not sufficient for all plant genomes. Indeed, highly heterozygous genomes or duplicated genomes suffer from regions that are often collapsed in genome assemblies. As an example, the recently sequenced tetraploid genome of potato [14] featured large regions (of the Mb order) that are identical between haplotypes, making the assembly impossible even with reads of several kilobases. To overcome this problem, a new technique called "gamete binning" has been recently developed. This technique, used to assemble the diploid apricot genome [15] and the autotetraploid potato genome [14], relies on the single cell sequencing of many gametes. This, in turn, makes it possible to reliably assign sequencing reads to a particular haplotype, and assemble long reads separately from the same haplotype. This simplifies the genome assembly process by reducing the complexity of the dataset, and enables a haplotype-by-haplotype assembly approach.

Although these technologies significantly simplify genome assembly, the resulting contigs often do not reflect the chromosomal organization of the original genome. Optical mapping is one of the long-range technologies now commonly used, in particular since the development of the Direct Label and Stain (DLS) protocol by Bionano Genomics (BNG). This protocol is designed to fluorescently label and repair high molecular weight (HMW) DNA at a specific location composed of six nucleotides (5′CTTAAG3′) using nicking

endonucleases. In contrast to the previous version (Nick Label Repair and Stain, NLRS), this labelling preserves the double stranded DNA and avoids fragmentation of long DNA molecules [16]. These labelled molecules are charged into a flow cell and migrate into nanochannels in order to stay linear. The fluorescence signal is scanned, and the images are converted into molecule files (containing the position of the markers on each molecule) with a provided software. The optical map is generated by assembling the individual molecules into larger ones which can represent chromosomes. An optical card does not contain sequences, only labeling positions and distances between these positions, which can be seen as a barcode. The older NLRS protocol is used less and less as optical maps, generated thanks to this protocol, are generally less contiguous due to the labelling process that tends to induce breaks into DNA molecules. Indeed, the enzyme recognizes a double stranded sequence of seven nucleotides (5′GCTCTTC3′) and operates as a restriction enzyme. If several restriction sites are very close to each other, it could be a possible break point during the labelling step [17].

These two optical maps can be combined in the scaffolding process, allowing users to add the benefit of the two labelling methods. Sequences from a given assembly are digested in silico, and the estimated position of the labels on the contigs are compared to those of the optical maps. The scaffolding process detects assembly errors in the contigs, breaks them in order to be consistent with the optical maps, and finally orients and orders the contigs to produce chromosome or chromosome arm scale sequences.

The chromosomal conformation capture technique can also be used to obtain high-continuity assemblies and give information on chromosomal regions adjacent in the nucleus [18]. Indeed, chromosomes fold into topologically associating domains (TADs) [19], and then sequences form loops and other folds. The number of contacts decreases as a function of genomic distance. The chromatin is fixed in order to preserve its three-dimensional (3D) organization. An enzymatic digestion and a proximity ligation step produce a chimeric fragment containing two portions of DNA, distant on the chromosomal sequence, but close in the nucleus space. The relative abundance of each ligation product is related to the probability that those DNA sequences interact in the 3D space. After library preparation and short read sequencing, the paired reads represent long distance linking information. The mapping of the Illumina paired reads on the assembly highlights the long-distance links between two contigs or scaffolds. Library preparation kits are commercialized by several companies, for example Dovetail Genomics who propose a kit named Omni-C. It has the advantage of using an endonuclease for the digestion step instead of restriction enzymes, as in other kits, which avoids biases in digestion. It is interesting to note that optical maps and Omni-C provide different types of information, as the former contains linear information while the latter provides information about the spatial organization of the DNA molecules. However, algorithms dedicated to the scaffolding of genome assemblies do not take advantage of this kind of information, as they hypothesize that adjacent regions share a higher number of contacts than more distant regions.

The Pore-C library preparation is quite similar to the Hi-C library preparation in that it uses restriction enzymes, but the DNA fragments can be a concatenation of several chimeric junctions. After sequencing, it is possible, and expected, that Nanopore reads contain the multiple interacting sites. Indeed, as Nanopore reads are long, they span entire amplicons. To determine from which part of the genome the Nanopore reads originates, an in silico digestion of the reference genome is performed, and each segment of the read is assigned to a subsection of the reference. Although relatively young [20], the Pore-C technology seems to be interesting and competitive with traditional Hi-C for the scaffolding of complex genomes [21].

Traditionally, the anchoring of sequences along the chromosomes was performed using genetic maps [22]. These are obtained by genotyping a segregating population using genetic markers, such as SNPs (Single Nucleotide Polymorphism). The polymorph SNPS are then arranged on linkage groups according to the recombination frequency between markers. Thereafter, the localization of genetically mapped markers on scaffolds allows the

validation, orientation, and anchoring of the scaffolds onto pseudomolecules [23]. However, some regions are still difficult to anchor due to a low density of genetic markers or the absence of recombination between markers, especially in (peri)centromeric regions [24] that are particularly rich in repetitive sequences [25], and thus require specific markers for anchoring [26].

Although sequenced with long reads in a previous study [27], the *Brassica rapa* (cv. Z1, AA, 2*n* = 20, double haploid line, genome size of 450 Mb) genome still contains a large number of unknown bases (33 Mb, representing 8.2% of the assembly). *Brassica* genus includes many important crops that are cultivated worldwide, notably for their oil production, or as vegetables. In addition, these species are one of the best models to study the importance of polyploidy in plant evolution, diversification, and adaptation, due to the occurrence of both ancient and recent polyploidization events [28]. Moreover, the *B.rapa* chromosomes [29] contain large centromeric regions that are notoriously difficult to assemble (at least, more complicated than the C genome, which contains, however, more transposable elements) [30], motivating us to resequence *B. rapa* cv.Z1 with the Oxford Nanopore technology and obtain the best assembly possible using current assemblers and long-range techniques.

2. Materials and Methods

2.1. DNA Extraction for Nanopore Sequencing

High-quality and high-molecular-weight (HMW) DNA was extracted in order to generate long reads using ONT. For this purpose, DNA was isolated from one gram of plant leaves previously placed in the dark following the protocol provided by Oxford Nanopore Technologies (Oxford, UK), "High molecular weight gDNA extraction from plant leaves" downloaded from the ONT Community in March, 2019. This protocol involves a conventional CTAB extraction followed by a purification using the commercial Qiagen Genomic tip (QIAGEN, Germantown, MD, USA), and is described in detail in Belser et al. [11]. HMW gDNA quality was checked on a 2200 TapeStation automated electrophoresis system (Agilent, Santa Clara, CA, USA) and the length of the DNA molecules was estimated to be over 60 Kb.

2.2. Nanopore Sequencing

Two libraries were prepared simultaneously according to the following protocol and using the Oxford Nanopore SQK-LSK109 kit. Genomic DNA fragments (2 µg) were repaired and 3′-adenylated with the NEBNext FFPE DNA Repair Mix and the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). Sequencing adapters provided by ONT (Oxford Nanopore Technologies Ltd., Oxford, UK) were then ligated using the NEBNext Quick Ligation Module (NEB). After purification with AMPure XP beads (Beckmann Coulter, Brea, CA, USA), the two libraries have been pooled into one. One third of the library was mixed with the Sequencing Buffer (ONT) and the Loading Bead (ONT) and loaded on a PromethION (Oxford Nanopore Technologies, Oxford, UK) R9.4.1 flow cell. A second third of the library was then loaded onto the flow cell after a Nuclease Flush using the Flow Cell Wash Kit EXP-WSH003 (ONT) according to the Oxford Nanopore protocol. The Nuclease Flush treatment consists of the use of a nuclease, which digests nucleic acids loaded on the flow cell. It allows the recovery of active pores, and increases the final yield of the run. ONT reads were basecalled using Guppy version 4.0.1 (Oxford Nanopore Technologies, Oxford, UK).

2.3. Nanopore Genome Assembly and Polishing

Three sets of reads were generated (Supplementary Table S1), in order to test assemblers with different read coverages and datasets. The first was composed of all reads, and for the second we selected a 30× coverage of the longest reads. The last was composed of 30× of the highest-scoring Filtlong [31] reads. Then, we launched Smartdenovo [32] (git commit 8488de9), Wtdbg2/Redbean [33] (git commit b77c565), and Flye [34] (version 2.8.3)

on all sets. In addition, Necat [35] (git commit d377878) was launched with the complete readset, as it corrects reads given as input and applies its own downsampling algorithm. We launched Smartdenovo with "-k 17", as advised by the developers in case of larger genomes, and "-c 1" to generate a consensus sequence. Redbean was launched with "-xont -X5000 -g 500m" and Flye with "-g 500m". NECAT was launched with a genome size of 500 Mb, and other parameters were left as default.

The assembly produced by Necat was retained, as it was the closest to the expected size of the genome and the most contiguous (Supplementary Tables S2–S5). The Necat assembly was polished once by using Racon (version 1.4.13) [8] and Medaka [36] (version 1.2.0) with Nanopore reads and twice with 250 bp-long paired-end Illumina reads (PRJEB26620) by using Hapo-G (version 1.0) [9]. Racon was launched with the following parameters: "-m 8 -x -6 -g -8 -w 500 -u", as advised by the Racon developers, and Medaka was launched with the "-m r941_prom_high_g360" parameter, in order to comply with the version of the basecaller that we used. Finally, Hapo-G was launched with the "-u" parameter.

2.4. Optical Mapping and Hybrid Scaffolding

Two optical maps were generated using previously generated molecules [27] and the assembly pipeline developed by Bionano Genomics (BNG) with the following two options: "add pre-assembly" and "non haplotype without extend and split". The two-enzyme hybrid scaffolding pipeline (bionano solve and tools Version: 1.6.1) was used to generate the hybrid scaffolds (Table 1). BisCoT [37] was used to correct artifactual duplications (negative gaps) introduced during the scaffolding process. A final step of polishing was performed using Hapo-G and Illumina [27] sequencing data.

Table 1. Metrics of scaffolds generated by using only one scaffolding technology, compared to input contigs. Statistics were generated using sequences of more than 30 kb in size.

	Input Contigs (Necat Assembly)	**Bionano**	**Omni-C**	**Pore-C**
Cumulative size	443,649,441	443,951,349	439,638,897	443,677,941
Number of sequences	299	236	590	253
N50 (L50)	10,461,875 (12)	17,017,634 (8)	25,523,596 (7)	20,151,380 (9)
N90 (L90)	857,267 (58)	3,409,175 (30)	221,999 (98)	1,472,408 (41)
auN	14,202,687	20,478,883	22,995,704	16,823,684
Max. size	45,115,632	44,069,534	42,018,994	32,180,808
Number of Ns (%)	0 (0%)	2,914,945 (0.66%)	20,900 (0.00%)	28,500 (0.01%)
Complete busco genes (%)	1604 (99.4%)	1604 (99.4%)	1604 (99.4%)	1604 (99.4%)
Merqury score	36.4423	37.1176	36.4875	36.4872

2.5. Omni-C Library Preparation and Illumina Sequencing

The Dovetail Omni-C library was prepared using the Dovetail Hi-C preparation kit (Dovetail Genomics, Scotts Valley, CA, USA), according to the manufacturer's protocol (manual version 1.0 for non-mammalian samples), using young frozen leaves previously placed in the dark. Briefly, after sample crosslinking, chromatin was digested using a sequence-independent endonuclease. Proximity ligation (which creates chimeric molecules) was performed using a biotin-labeled bridge between the ends of the digested DNA. After reversal crosslinking, the DNA was purified and followed by library generation (omitting the fragmentation step). Finally, the biotinylated chimeric molecules were captured and amplified before sequencing on the Novaseq 6000 instrument (Illumina, San Diego, CA, USA) using 150 base-length read chemistry in paired-end mode.

2.6. Scaffolding Using the Omni-C Library

Scaffolding was realized thanks to the 3D de novo assembly (3D-DNA [38]) pipeline (version 180419). Hi-C raw reads were aligned against the assembly (-s none option) using Juicer. The resulting merged_nodups.txt file and the assembly were given to the run-asm-

pipeline.sh script with the options "–editor-repeat-coverage 5 –splitter-coarse-stringency 30 –editor-coarse-resolution 100,000". Contact maps were visualized through the Juicebox tool [39] (version 1.11.08) and edited to adjust the construction of scaffolds or break misjoins. After edition, the new.assembly file was downloaded from the Juicebox interface. The file is filtered and converted into a fasta file thanks to the juicebox_assembly_converter.py script [40].

2.7. Pore-C Library Preparation and Nanopore Sequencing

The RE-Pore-C library was carried out as described in the Oxford Nanopore Technologies RE-Pore-C protocol for plant samples (30 July 2020 version), with the exception of tissue fixation, which was performed following the protocol of Chang Liu [41]. Crosslinked plant nuclei were isolated prior to in situ restriction digestion with NlaIII (New England Biolabs, Ipswich, MA, USA). After overnight incubation, the restriction enzyme was heat-denatured; crosslinked DNA clusters were ligated in proximity, followed by protein degradation and de-crosslinking, releasing the chimeric Pore-C dsDNA polymers. Libraries were constructed using the ONT Ligation Sequencing Kit (SQK-LSK109), following the library preparation recommendations. Sequencing was carried out on a R.9.4.1 PromethION flowcell. Nuclease washes were used to maximize output.

2.8. Scaffolding Using the Pore-C Library

Prior to scaffolding, we removed all nanopore Pore-C reads that were larger than 100 kb in size, as it drastically increases running times and can be problematic for the pipeline to handle, as stated in a github issue [42]. Then, we used the Pore-C Snakemake [43] pipeline (git commit 6b2f762) developed by ONT to generate the necessary files for scaffolding with the Salsa2 [44] scaffolder. Both Salsa2 and the Pore-C Snakemake pipeline were launched with default parameters.

2.9. Super-Scaffolding

In addition to testing the scaffolding of the nanopore assembly with one long-range technology, we combined several techniques to test if combining scaffolding methods would lead to better results. In particular, we super-scaffolded the BNG scaffolds with the Pore-C and Omni-C libraries and vice versa. Tools and parameters were the same as already described in previous sections. The 3D-DNA additional option -r 0 was used for scaffolding the BNG scaffolds with the Omni-C library.

2.10. Validation of the Assemblies

2.10.1. Comparison to Reference Genomes

We used the *B.rapa* cv. Chiifu v3 [45] and the *B.napus* cv. Darmor-BZH v10 [30] A-subgenome to generate dotplots and compare our different assemblies to established reference genomes. To do so, we used minimap2 with the "-x asm20" parameter to generate alignments of the assemblies against each reference. Then, we used D-Genies [46] in order to visualize the alignments and enabled the "Sort contigs" and "Hide noise" options.

2.10.2. Quality Assessment with Merqury

We used merqury (version 1.3, git commit 6b5405e) to obtain a quality score for each of our assemblies. First, we used the bundled best_k.sh script with a tolerable collision rate of 0.0001 and a genome size of 500 Mb to find the best size of kmer for our genome, which gave us a kmer size of 21. Then, we used meryl (version 1.3, git commit 3400615) to compute the 21-mer counts with Illumina reads via the meryl count command with default parameters. Finally, we used merqury to compare the kmers of each assembly to the kmers extracted from the Illumina reads.

2.10.3. Gene-Completeness Estimation with Busco

In order to estimate the gene completeness of the genome assemblies, we launched Busco version 5.1.2 with the embryophyta datasets (odb10). All other options were left as default.

2.11. Construction of a B. rapa Genetic Map and Anchoring

To construct a *B. rapa* genetic map, we created a F2 population (149 plants) deriving from an initial cross between the doubled haploid *B. rapa ssp trilocularis* cv. Z1 and the inbred line *B. rapa ssp pekinensis* cv. Chiifu-401-42. DNA from the parental lines, the F1 hybrid and the 149 plants were extracted using the sbeadex plant kit (LGC Genomics, Teddington Middlesex, UK) on the oKtopure robot at the GENTYANE platform (INRAE, Clermont-Ferrand, France), and thereafter genotyped using the *Brassica* 19 K Illumina infinium SNP array (TraitGenetics, Gatersleben, Germany). From these data, a total of 4030 markers were found polymorph between the parental lines, and were used to create a genetic map (985.9 cM in total) using CarthaGene [47] software version 1.2.3, with a LOD score of 4 and a maximal genetic distance of 0.21 cM. For all these genetically mapped markers, we then blasted their sequence contexts against the *B. rapa* cv. Z1 scaffolds obtained using either additional Bionano data only or both Bionano and Omni-C data, totaling 3867 and 3865 physically anchored markers, respectively. This step allowed us to help with the ordering and orientation of scaffolds, as well as to compare the quality of both assemblies.

2.12. Putative Position of Peri-Centromere and Sub-Telomere Regions

The putative position of (peri)centromeres was inferred by blasting the centromere-specific repeat sequences CentBr1 and CentBr2 (CW978699 and CW978837, respectively [48]) against this novel assembly, as well as sequences that are found in the pericentromeric heterochromatin blocks of *Brassica* chromosomes: the centromere-specific Ty1/copia-like retrotransposon of *Brassica* (CRB, AC166739), the pericentromeric Ty3/Gypsy-like retrotransposon of *B.rapa* (PCRBr, ACC166740), a 238-bp degenerate tandem repeat (TR238, AC166740), and a 805-bp tandem repeat (TR805, AC166739 [49]) (Lim et al. 2007). The putative positions of subtelomeres were inferred by blasting the *B. rapa* subtelomeric satellite repeats (pBrSTRa/b, EU294384 and EU294385 [50]) against this novel assembly.

2.13. Gene Prediction

Gene prediction was performed using several proteomes: 8 from other genotypes of *B. napus* [51] (Westar, Zs11, QuintaA, Zheyou73, N02127, GanganF73, Tapidor3, and Shengli3), *Arabidopsis thaliana* (UP000006548), and the 2021 annotation of Darmor-bzh [30]. Regions of low complexity in genomic sequences were masked with the DustMasker algorithms [52] (version 1.0.0 from the blast 2.10.0 package). Proteomes were then aligned on the genome in a 2-step strategy. First, BLAT [53] (version 36 with default parameter) was used to quickly localize corresponding putative regions of these proteins on the genome. The best match, and the matches with a score $\geq$90% of the best match score, have been retained. Second, alignments were refined using Genewise [54] (version 2.2.0 default parameters), which is more accurate for detecting intron/exon boundaries. Alignments were kept if >75% of the length of the protein was aligned on the genome.

All the protein alignments were combined using Gmove [55], which is an easy-to-use predictor with no need for a pre-calibration step. Briefly, putative exons and introns extracted from alignments were used to build a graph, where nodes and edges represent exons and introns, respectively. Gmove extracts all paths from the graph, and searches open reading frames that are consistent with the protein evidence. Finally, we decided to exclude single-exon genes composed of >80% of untranslated regions. Following this pipeline, we predicted 56,073 genes with 4.39 exons per gene on average.

3. Results

3.1. Nanopore Sequencing and Long Reads Assembly

A single PromethION R9.4.1 flowcell produced 93 Gb of data with an N50 of 26.9 kb (Table S1), representing a genome coverage of approximately 186×, with 38× being composed of reads longer than 50 Kb.

We subsampled data as previously described (30× longest, 30× highest scoring Filtlong reads, and all reads). Flye and Redbean produced their most contiguous assembly with a subset of reads, showing that experimenting with different coverages may be beneficial for the assembly. Necat with the entire readset led to the most contiguous assembly, with a cumulative size of 442 Mb, a contig N50 of 10.4 Mb, a merqury quality score of 27.5, and 1567 (97.1%) complete and single-copy embryophyta BUSCO genes (Table S5). After polishing with Nanopore and Illumina reads (Table S6), the merqury quality score rose to 36.4 and the number of complete embryophyta BUSCO genes increased to 1604 (99.4%). By aligning the Necat assembly to the *B. rapa* cv. Chiifu v3 and the *B. napus* Darmor-bzh v10 reference genomes (Figure S1), we could detect the presence of a contig, showing a translocation between the A07 and A03 chromosomes. We concluded that this contig was chimeric as it was detected as such and cut later by the Hi-C and BNG software. However, this chimeric junction was left as is, and used to see if scaffolding algorithms would be able to correct it.

3.2. Long Range Genome Assembly

3.2.1. Hybrid Scaffolding

The produced DLE and BspQI maps achieved a N50 of 13.5 Mb and 1.9 Mb and a cumulative size of 466 Mb and 434 Mb, respectively. The combination of the two optical maps with the nanopore contigs lead to an assembly that reached a N50 of 16.8 Mb with a cumulative size of 446.4 Mb (Table S7). Very few gaps were introduced underlining the great concordance between the length of the optical maps and the nanopore assembly. Finally, as expected, BisCoT software increased the N50 and decreased the cumulative size. Indeed, this was expected, as BisCoT is designed to remove artifactual tandem duplications (negative gaps) and the possible redundancies in the assembly. A final round of polishing gave an assembly with a N50 of 17 Mb and a cumulative size of 443.9 Mb (Table S7). The comparisons with the *Chiifu* genome and Darmor-bzh A-subgenome showed neither a chimeric scaffold nor a misorganization (Figure S2).

3.2.2. Hi-C Scaffolding

The sequencing of the Omni-C library produced 108 M paired reads (32 Gb) that were aligned on the nanopore polished contigs using the Juicer pipeline. The mapping information was used by 3D-DNA [38] to orient and order sequences of the input assembly. A contact map was generated and visualized through the Juicebox [39] interface. We edited the contact map and merged some scaffolds (two larger scaffolds were constructed by merging two scaffolds for each) and broke one misjoin (Figure S9). We obtained an assembly of 439 Mb with a N50 of 25.5 Mb (Table 1). Comparisons with the *Chiifu* genome and *B. napus* A-subgenome showed neither a misjoin nor a misorganization (Figure S3).

3.2.3. Pore-C Scaffolding

The sequencing of the Pore-C library (Table S8) produced 15.55 Gb of data with an N50 of 3.99 Kb. As advised by the developers of the Pore-C Snakemake pipeline, we removed reads that were longer than 100 Kb and obtained a resulting dataset composed of 5.8 million reads with an N50 of 3.97 Kb, for a total cumulative size of 15.50 Gb. This readset was used to perform all scaffoldings with the Pore-C technology presented in this study.

We applied the Pore-C Snakemake pipeline released by Oxford Nanopore, as well as Salsa2 to scaffold the Necat assembly (Table 1), and obtained 253 scaffolds for a total cumulative size of 443 Mb, with a scaffold N50 of 20.1 Mb. After aligning scaffolds onto

the reference genomes, and inspecting alignments and the resulting dotplots (Figure S4), we could not find any evidence indicating the presence of chimeric sequences.

3.3. Combination of Several Long-Range Techniques

Additionally, we investigated whether the combination of long-range technologies could add the benefits of each (Figure 1). For this purpose, we scaffolded the Omni-C and Pore-C assemblies with the optical maps. Hi-C scaffolding sometimes introduces misjoins (chimera or chromosome arm inversion) that could be corrected by the comparison with optical maps. Surprisingly, the scaffolding of the Omni-C assembly (Omni-C + BNG assembly) decreased the contiguity and added more than 2% of undetermined bases. It produced an assembly of 445 Mb with a N50 of 20.3 Mb (Figure S7). On the other hand, the Pore-C + BNG assembly produced an assembly of 450 Mb and a N50 of 25.9 Mb (Figure S8). The N50 slightly increased, but scaffolds did not achieve the chromosome scale. When comparing final assemblies to reference genomes, no misjoins were detected, with the exception of a part of the A01 chromosome being duplicated in the Pore-C + BNG assembly. Conversely, we scaffolded the BNG scaffolds with the Hi-C or the Pore-C libraries (Table 2). As the BNG scaffolds were close to chromosome scale, we expected that, with the long distance contacts of the Hi-C, we could organize and orient the BNG scaffolds into complete chromosomes. We obtained a BNG + Omni-C assembly of 440 Mb with a N50 of 33.3 Mb after editing the contact map (Figure S10). The size of the 10 largest scaffolds were compatible with chromosome length, and dotplots showed good consistency with the *B. napus* A-subgenome and the *Chiifu* genome (Figure S5). Likewise, we obtained a BNG + Pore-C assembly of nearly 444 Mb with a N50 of 17 Mb (Figure S6). Again, the alignments with reference genomes did not show any inconsistencies, but no scaffold achieved the chromosome scale. As metrics were similar to those obtained with the BNG scaffolds, we put aside the BNG + Pore-C assembly.

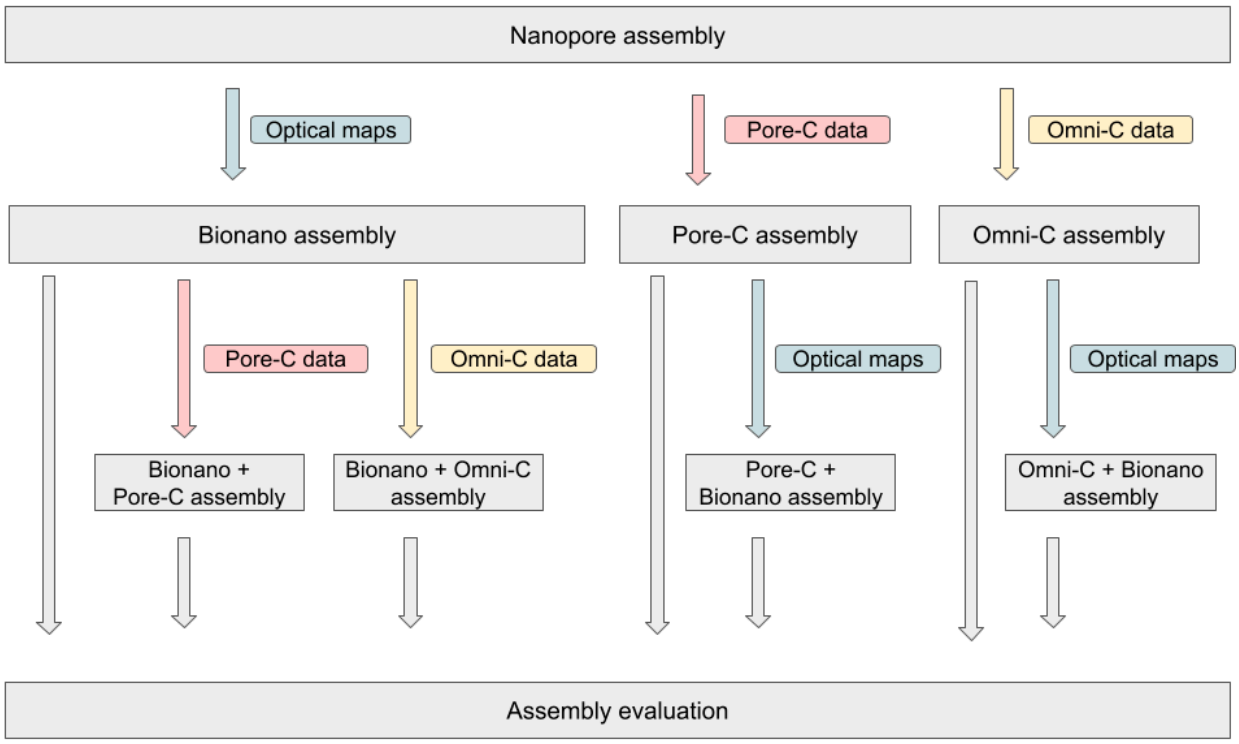


Figure 1. Global view of the different scaffolding experiments. Different scaffolding techniques were applied to the polished Nanopore contigs. Single-technology strategies were first tested. These include Bionano, Omni-C, and Pore-C scaffolding. In a second phase, scaffolds obtained previously were scaffolded again using a different method. In particular, Bionano scaffolds were super-scaffolded using either Pore-C (BNG + Pore-C) or Omni-C (BNG + Omni-C). Pore-C and Omni-C scaffolds were super-scaffolded using Bionano optical maps (respectively, Pore-C + BNG and Omni-C + BNG).

Table 2. Metrics of scaffolds generated by using a combination of different scaffolding technologies, compared to input contigs. Statistics were generated using sequences of more than 30 kb in size.

	Input Contigs (Necat Assembly)	**Bionano + Omni-C**	**Bionano + Pore-C**	**Omni-C + Bionano**	**Pore-C + Bionano**
Cumulative size	443,649,441	440,038,627	443,961,849	445,844,245	450,760,401
Number of sequences	299	511	222	401	215
N50 (L50)	10,461,875 (12)	33,316,896 (5)	17,017,634 (8)	20,321,816 (7)	25,915,290 (7)
N90 (L90)	857,267 (58)	275,999 (61)	3,409,175 (29)	1,763,661 (32)	3,720,451 (26)
auN	14,202,687	34,864,156	20,976,778	22,582,099	22,825,029
Max. size	45,115,632	64,589,792	44,069,534	51,305,606	43,928,997
Number of Ns (%)	0 (0%)	2,930,045 (0.67%)	2,925,445 (0.66%)	10,143,940 (2.28%)	3,522,960 (0.78%)
Complete buscos genes (%)	1604 (99.4%)	1604 (99.4%)	1604 (99.4%)	1604 (99.4%)	1604 (99.4%)
Merqury score	36.4423	37.1179	37.1176	37.0247	37.0566

3.4. Anchoring

Based on the previous results, the anchoring with the genetic map was performed on two assemblies: The BNG scaffolds and the BNG + Omni-C scaffolds. The comparison of the position of the markers allows us to verify the structure of the scaffolds, to assign and orientate scaffolds onto chromosomes. We were able to anchor 384 Mb of the BNG scaffolds and 369 Mb of the BNG + Omni-C scaffolds on the chromosomes (compared to 357 Mb of anchored sequences in the V1 version). The putative chromosomes are composed of two to five BNG scaffolds or one to two BNG + Omni-C scaffolds. No misjoins were detected in the BNG scaffolds, but one inversion was detected in the BNG + Omni-C scaffolds. To construct the final version, we decided to keep the 36 anchored BNG scaffolds that contained more sequences. However, we used the Omni-C data to add two scaffolds in chromosome A03 and one scaffold in chromosome A09 which were difficult to anchor with the genetic map. In the end, we succeeded to anchor 386.05 Mb (86.96 % of the whole genome assembly). The final *B. rapa* cv. Z1 version 2 is composed of 210 scaffolds with a cumulative size of around 444 Mb and a N50 of 39,2 Mb (L50 = 5) (Table 3 and Figure 2). Among the ten chromosomes, seven chromosomes contained telomeric repeats (TTTAGGG motif) at both ends, and the remaining three had telomeric repeats at one end. In the same way, subtelomeric repeats have been found in each chromosome (except for chromosome A03) (Figure 2). Compared to the *B. rapa* cv. Z1 version 1, the cumulative size and the contig N50 have greatly increased (6.6 Mb and 10.2 Mb for versions 1 and 2, respectively) showing that the new assembly covered a higher proportion of the estimated genome length (Table 3). In the same way, the rate of undetermined bases dropped from 8.22 to 0.66%. The final comparison of the genetic and physical positions revealed that only 179 out of the 3867 markers (4.63%) were discordant, most often due to an inaccurate position on the genetic map (of a few cM).

Table 3. Metrics of the final version of *Brassica rapa* cv Z1 Version 2.

	Input Contigs (Necat Assembly)	**Bionano**	***Brassica rapa* cv Z1 V2**	***Brassica rapa* cv Z1 V1 [27]**
Cumulative size	443,649,441	443,951,349	443,953,949	401,164,957
Number of sequences	299	236	210	237
N50 (L50)	10,461,875 (12)	17,017,634 (8)	39,217,720 (5)	34,481,996 (5)
N90 (L90)	857,267 (58)	3,409,175 (30)	4,034,065 (13)	2,865,407 (12)
auN	14,202,687	20,478,883	38,695,451	36,201,043
Max. size	45,115,632	44,069,534	68,194,707	57,670,803
Number of Ns (%)	0 (0%)	2,914,945 (0.66%)	2,917,545 (0.66%)	32,966,574 (8.22%)
Number of contigs	299	301	295	297

Table 3. *Cont.*

	Input Contigs (Necat Assembly)	**Bionano**	***Brassica rapa* cv Z1 V2**	***Brassica rapa* cv Z1 V1 [27]**
Contigs N50 (L50)	10,461,875 (12)	10,256,333 (14)	10,256,333 (14)	6,651,009 (14)
Complete busco genes (%)	1604 (99.4%)	1604 (99.4%)	1604 (99.4%)	1594 (98.7%)
Merqury score	36.4423	37.1176	37.119	28.4862
Number of genes	-	-	56,073	46,721
Number of exons/gene	-	-	4.39	4.72
Complete busco genes (%)	-	-	1573 (97.5%)	1553 (96.2%)
Duplicated busco genes (%)	-	-	226 (14.0%)	216 (13.4%)
Fragmented busco genes (%)	-	-	16 (1.0%)	20 (1.2%)
Missing busco genes (%)	-	-	25 (1.5%)	41(2.6%)

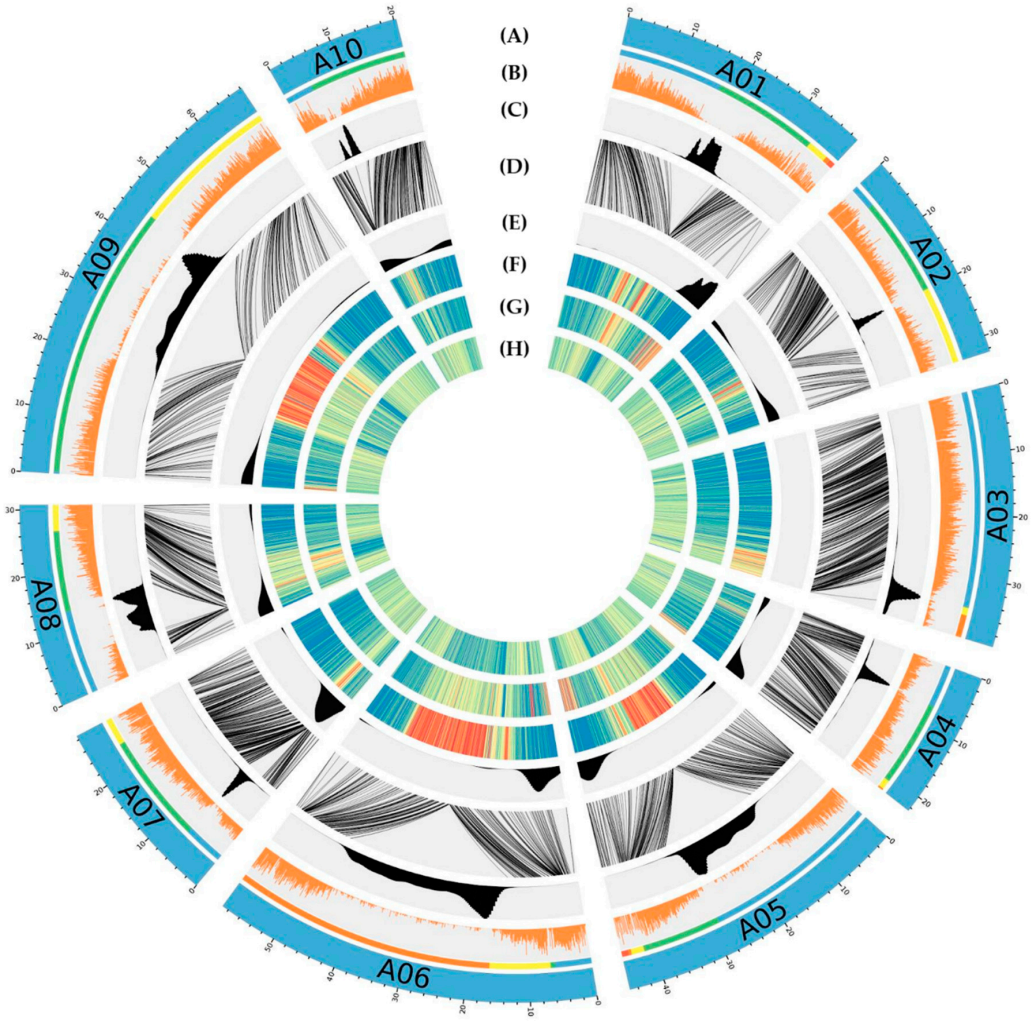


Figure 2. Circular representation of the 10 chromosomes obtained in this novel *B. rapa* cv Z1 assembly V2. (**A**) Scaffolds used to generate each pseudomolecule; (**B**) gene density; (**C**) putative position of (peri)centromeres; (**D**) relationships between the physical and genetic position of the SNP markers used to create the *B. rapa* genetic maps and order the scaffolds onto pseudomolecules; (**E**) density of subtelomeric satellite repeats; (**F**) density of Gypsy elements; (**G**) density of Copia elements; and (**H**) density of DNA retrotransposons.

4. Discussion

In this study, we used the genome of *B. rapa* cv. Z1 to feature how plant genomes can now be generated. We chose the Oxford Nanopore PromethION technology and also compared current long-range scaffolding techniques, namely Bionano optical mapping and the Hi-C and Pore-C chromatin conformation capture techniques, in order to obtain chromosome-scale assembly.

We sequenced the genome of *B. rapa* cv. Z1 on a single R9.4.1 PromethION flowcell and showed that sequencing a medium-sized genome is now affordable to individual laboratories as the total cost of sequencing, including consumables, was about USD 1110. The great amount of data we obtained, added to the size of the reads, made it possible for us to assemble the genome using Necat into large contigs, with an N50 of 10.4 Mb. However, after aligning this assembly to the *B. rapa* cv. Chiifu or the *B. napus* cv. Darmor-bzh A-subgenome, we detected a chimeric contig, showing a translocation from the A07 to the A03 chromosome. We polished the contigs using a combination of Nanopore data and Illumina data, and reached a quality score of 36.4, representing an error rate of 0.02%. Remarkably, the quality score was already high, even when polishing only with Nanopore reads (31.04), showing that the Nanopore sequencing technology has made sizable progress since its beginnings [56].

The polished contigs were then scaffolded using optical maps, or two chromatin conformation capture techniques, Omni-C and Pore-C. Regarding the continuity of the resulting scaffolds, the ones obtained with Omni-C had the largest N50 (25 Mb), but the BNG and Pore-C scaffolds were very close to it, with an almost identical L50. However, a scaffold N90 of 3.4 Mb seemed to indicate that scaffolding with optical maps makes it possible to anchor a higher proportion of small contigs. In comparison, the N90 of the Pore-C and Omni-C assemblies were 1.47 Mb and 222 kb, respectively. Scaffolding with optical maps also introduced 2.9 Mb of unknown bases (Ns), as it is the only technique tested here that can estimate gap sizes, with other techniques only inserting an arbitrary amount in each gap, giving the false impression that the assembly is complete. Finally, when scaffolds were compared to reference genomes, we were pleased to see that every technique was able to successfully correct the chimeric region that was initially present in Nanopore contigs.

As we wanted to see if combining long-range technologies would lead to a better assembly, we scaffolded polished Nanopore contigs by combining two long-range techniques. We were able to reach chromosome-scale, first by using optical maps, and then Omni-C. Indeed, we obtained a single scaffold per chromosome assembly (N50 of 33.3 Mb and L50 of 5), with 2.9 Mb (0.67%) of undetermined bases.

In contrast, we noticed that Omni-C or Pore-C scaffolds were fragmented when integrated with optical maps (Figures S11 and S12). Indeed, the sequences are broken when conflicts are detected between the optical maps and the assembly. As we know that, in our case, the Hi-C scaffolding did not produce chimeras (Figures S3 and S4), we assume that the conflicts may result from a poor estimate of the gap size during the Hi-C scaffolding process. In this case, the scaffolds are chained with fixed gap sizes (100 nucleotides) even if some small contigs can take place in these gaps, with the consequence of generating conflicts with the optical maps. This is as BNG software cannot handle underestimated gap sizes. Therefore, we do not recommend chaining Hi-C scaffolding and then optical maps in this order. Conversely, Hi-C libraries generally contain more distant information than optical maps, and the integration of Hi-C data on the BNG assembly did not suffer from incompatibility issues. Likewise, this combination can also be used to validate the first scaffolding, as BNG scaffolds can already reach the chromosome scale.

The use of the genetic map allowed us to obtain complete chromosomes from the BNG assembly, and to validate the BNG + Omni-C scaffolds. These assemblies were chosen for their global metrics: BNG scaffolds had good N50 and N90, BNG + Omni-C scaffolds already reached chromosome scale. The quantity of anchored bases is lower with the BNG + Omni-C assembly, showing that the Omni-C scaffolding had fragmented scaffolds during

the scaffolding process. Highly repeated regions are more likely difficult to organize with Hi-C data, and the very large centromeres in Z1 were problematic. As an advantage of the optical map, complex regions as centromere or rDNA genes clusters are generally well covered, thanks to the very long DNA molecules (>200 Kb long), allowing the organization of long-read assemblies.

In our opinion, optical mapping is the recommended long-range technology that leads to complete assemblies, mostly in complex regions such as centromeres, but requires additional tools which are not provided by BNG. Indeed, bioinformatic tools dedicated to optical maps are rare, which makes the use of this data more complicated than Hi-C short-reads. Additionally, it could be difficult to extract high molecular weight DNA in particular for plant genomes. Obtaining very long DNA molecules without any contaminant as polyphenol compounds that can inhibit the enzymatic reactions is critical. In our own hands, the DLE-1 enzyme was more sensitive to contaminants than the previous enzyme BspQI. The alternative, in the case of unsuccessful optical map preparation, is to switch to Hi-C technology, for which sequencing libraries and bioinformatic analysis are easier. Indeed, Hi-C can produce chromosome-scale assemblies, even if the anchoring of complex regions, such as centromeres, is less effective, and results should be reviewed with attention.

5. Conclusions

The assembly of genomes, and in particular plant genomes, is a challenging field which is undergoing major technological and software evolutions. It generally requires the combination of several technologies, with which it is important to be familiar with, to obtain high-quality results. In this study, we share our experience of reconstructing the genome of *Brassica rapa* (cv. Z1) using long-read sequencing and long-range scaffolding techniques. In our opinion, Bionano Genomics optical mapping is a good choice for organizing and validating long-read assemblies. It anchors the greatest amount of nucleotides, especially in complex regions, such as those with a high proportion of repetitive elements. In addition, it allows estimation of the size of the gaps, unlike scaffolding methods based on Hi-C (or Pore-C) data. However, Hi-C technology is the most popular scaffolding technique, and represents a powerful long-range technology with an active community regularly developing tools and methods. Indeed, the requirement in terms of input material and its sequencing on widely distributed Illumina sequencers makes it an commonly chosen technology. Using only one or carefully combining two methods, we have organized the *Brassica rapa* long-read contigs into a chromosome-scale assembly and obtained the most contiguous genome assembly for this species to date.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/biology10080732/s1, Table S1: Metrics of Nanopore datasets, Table S2: Metrics of raw Smartdenovo Nanopore assemblies, Table S3: Metrics of raw Flye Nanopore assemblies, Table S4: Metrics of raw Wtdbg2 Nanopore assemblies, Table S5: Metrics of the raw Necat Nanopore assembly, Table S6: Metrics of the Necat assembly after each polishing step, Table S7: Hybrid scaffolding results, Table S8: Metrics of the Pore-C raw PromethION run with (left column) or without (right column) reads of more than 100 kb in size, Figure S1: Dotplots of polished Necat contigs aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right), Figure S2: Dotplots of Bionano scaffolds aligned to Brassica rapa Chiifu (left) or Brassica napus Darmor-BZH A genome (right), Figure S3: Dotplots of Omni-C scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right), Figure S4: Dotplots of Pore-C scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right), Figure S5: Dotplots of Bionano + Omni-C scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right), Figure S6. Dotplots of Bionano + Pore-C scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right), Figure S7: Dotplots of Omni-C + Bionano scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right), Figure S8: Dotplots of Pore-C + Bionano scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right), Figure S9: Contact Map generated with the Omni-C library (mapping on the nanopore contigs), Figure S10: Contact Map generated with the Omni-C library (mapping on the bionano scaffolds), Figure S11:

Conflict example detected by the optical maps on the Omni-C scaffolds, and Figure S12: Conflict example detected by the optical maps on the Pore-C scaffold.

Author Contributions: B.I.: Methodology, Software, Validation, and Writing—Original Draft; C.B.: Methodology, Software, Validation, and Writing—Original Draft; C.F.: Resources, Software, Validation, and Formal analysis; K.L.: Resources; F.B.: Software and Formal analysis; G.D.: Resources; L.M.: Software and Formal analysis; C.C.: Resources; L.B.: Resources; A.-M.C.: Supervision and Writing—Review and Editing; P.W.: Supervision; M.R.-G.: Supervision and Writing—Review and Editing; and J.-M.A.: Conceptualization, Supervision, and Writing—Review and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: 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), the "Région Bretagne" as well as INRAE 'Plant Breeding' department that funded the Ph.D. scholarship of Franz Boideau.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The genome assembly and gene predictions are freely available at http://www.genoscope.cns.fr/plants (accessed on 30 July 2021). The Illumina, and the bionano data of the *B. rapa* cv Z1 version 1 are available in the European Nucleotide Archive under the following projects: PRJEB26620. The PromethION sequencing data, the Omni-C Illumina sequencing data, the Pore-C nanopore sequencing data and the new optical maps are available in the European Nucleotide Archive under the following projects: PRJEB46167.

Acknowledgments: We thank the UMR INRA 1095 'GENTYANE platform' (Clermont-Ferrand, France, http://gentyane.clermont.inra.fr/) (accessed on 30 July 2021) for the DNA extraction of segregating population and the platform TraitGenetics (Gatersleben, Germany, http://www.traitgenetics.com/) (accessed on 30 July 2021) for genotyping using the *Brassica* 19K Illumina infinium SNP array.

Conflicts of Interest: J.-M.A. received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences. J.-M.A. and C.B. received accommodation expenses to speak during Bionano Genomics user meetings. The authors declare that they have no other competing interests.

References

1. Zimin, A.V.; Stevens, K.A.; Crepeau, M.W.; Puiu, D.; Wegrzyn, J.L.; Yorke, J.A.; Langley, C.H.; Neale, D.B.; Salzberg, S.L. An Improved Assembly of the Loblolly Pine Mega-Genome Using Long-Read Single-Molecule Sequencing. *Gigascience* **2017**, *6*, 1–4. [PubMed]
2. Claros, M.G.; Bautista, R.; Guerrero-Fernández, D.; Benzerki, H.; Seoane, P.; Fernández-Pozo, N. Why Assembling Plant Genome Sequences Is so Challenging. *Biology* **2012**, *1*, 439–459. [CrossRef]
3. Bentley, D.R.; Balasubramanian, S.; Swerdlow, H.P.; Smith, G.P.; Milton, J.; Brown, C.G.; Hall, K.P.; Evers, D.J.; Barnes, C.L.; Bignell, H.R.; et al. Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry. *Nature* **2008**, *456*, 53–59. [CrossRef] [PubMed]
4. Zhang, H.; Jain, C.; Aluru, S. A Comprehensive Evaluation of Long Read Error Correction Methods. *BMC Genom.* **2020**, *21*, 889. [CrossRef] [PubMed]
5. Sahlin, K.; Medvedev, P. Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis. *Nat. Commun.* **2021**, *12*, 2. [CrossRef] [PubMed]
6. Watson, M.; Warr, A. Errors in Long-Read Assemblies Can Critically Affect Protein Prediction. *Nat. Biotechnol.* **2019**, *37*, 124–126. [CrossRef]
7. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K.; et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* **2014**, *9*, e112963. [CrossRef] [PubMed]
8. Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads. *Genome Res.* **2017**, *27*, 737–746. [CrossRef]
9. Aury, J.-M.; Istace, B. Hapo-G, Haplotype-Aware Polishing of Genome Assemblies with Accurate Reads. *NAR Genom. Bioinform.* **2021**, *3*, lqab034. [CrossRef] [PubMed]
10. Driguez, P.; Bougouffa, S.; Carty, K.; Putra, A.; Jabbari, K.; Reddy, M.; Soppe, R.; Cheung, N.; Fukasawa, Y.; Ermini, L. LeafGo: Leaf to Genome, a Quick Workflow to Produce High-Quality De Novo Genomes with Third Generation Sequencing Technology. *bioRxiv* **2021**. [CrossRef]
11. Belser, C.; Baurens, F.-C.; Noel, B.; Martin, G.; Cruaud, C.; Istace, B.; Yahiaoui, N.; Labadie, K.; Hřibová, E.; Doležel, J.; et al. Telomere-to-Telomere Gapless Chromosomes of Banana Using Nanopore Sequencing. *bioRxiv* **2021**. [CrossRef]

12. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **2009**, *323*, 133–138. [CrossRef]
13. Hon, T.; Mars, K.; Young, G.; Tsai, Y.-C.; Karalius, J.W.; Landolin, J.M.; Maurer, N.; Kudrna, D.; Hardigan, M.A.; Steiner, C.C.; et al. Highly Accurate Long-Read HiFi Sequencing Data for Five Complex Genomes. *Sci. Data* **2020**, *7*, 399. [CrossRef]
14. Sun, H.; Jiao, W.-B.; Krause, K.; Campoy, J.A.; Goel, M.; Folz-Donahue, K.; Kukat, C.; Huettel, B.; Schneeberger, K. Chromosome-Scale and Haplotype-Resolved Genome Assembly of a Tetraploid Potato Cultivar. *bioRxiv* **2021**. [CrossRef]
15. Campoy, J.A.; Sun, H.; Goel, M.; Jiao, W.-B.; Folz-Donahue, K.; Wang, N.; Rubio, M.; Liu, C.; Kukat, C.; Ruiz, D.; et al. Gamete Binning: Chromosome-Level and Haplotype-Resolved Genome Assembly Enabled by High-Throughput Single-Cell Sequencing of Gamete Genomes. *Genome Biol.* **2020**, *21*, 306. [CrossRef] [PubMed]
16. Yuan, Y.; Chung, C.Y.-L.; Chan, T.-F. Advances in Optical Mapping for Genomic Research. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2051–2062. [CrossRef] [PubMed]
17. Bionano Genomics. Generating Accurate and Contiguous De Novo Genome Assemblies Using Hybrid Scaffolding. 2017. Available online: https://bionanogenomics.com/wp-content/uploads/2017/02/Bionano_HumanPAG_Hybrid-Scaffolding-White-Paper.pdf (accessed on 30 July 2021).
18. Belton, J.-M.; McCord, R.P.; Gibcus, J.H.; Naumova, N.; Zhan, Y.; Dekker, J. Hi-C: A Comprehensive Technique to Capture the Conformation of Genomes. *Methods* **2012**, *58*, 268–276. [CrossRef]
19. McCord, R.P.; Kaplan, N.; Giorgetti, L. Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. *Mol. Cell* **2020**, *77*, 688–708. [CrossRef]
20. Ulahannan, N.; Pendleton, M.; Deshpande, A.; Schwenk, S.; Behr, J.M.; Dai, X.; Tyer, C.; Rughani, P.; Kudman, S.; Adney, E.; et al. Nanopore Sequencing of DNA Concatemers Reveals Higher-Order Features of Chromatin Structure. *bioRxiv* **2019**. [CrossRef]
21. Choi, J.Y.; Dai, X.; Peng, J.Z.; Rughani, P.; Hickey, S.; Harrington, E.; Juul, S.; Ayroles, J.; Purugganan, M.; Stacy, E.A. Selection on Old Variants Drives Adaptive Radiation of Metrosideros across the Hawaiian Islands. *bioRxiv* **2020**. [CrossRef]
22. Fierst, J.L. Using Linkage Maps to Correct and Scaffold de Novo Genome Assemblies: Methods, Challenges, and Computational Tools. *Front. Genet.* **2015**, *6*, 220. [CrossRef]
23. Yu, A.; Li, F.; Xu, W.; Wang, Z.; Sun, C.; Han, B.; Wang, Y.; Wang, B.; Cheng, X.; Liu, A. Application of a High-Resolution Genetic Map for Chromosome-Scale Genome Assembly and Fine QTLs Mapping of Seed Size and Weight Traits in Castor Bean. *Sci. Rep.* **2019**, *9*, 11950. [CrossRef] [PubMed]
24. Li, S.; Yang, G.; Yang, S.; Just, J.; Yan, H.; Zhou, N.; Jian, H.; Wang, Q.; Chen, M.; Qiu, X.; et al. The Development of a High-Density Genetic Map Significantly Improves the Quality of Reference Genome Assemblies for Rose. *Sci. Rep.* **2019**, *9*, 5985. [CrossRef]
25. Zhang, W.; Cao, Y.; Wang, K.; Zhao, T.; Chen, J.; Pan, M.; Wang, Q.; Feng, S.; Guo, W.; Zhou, B.; et al. Identification of Centromeric Regions on the Linkage Map of Cotton Using Centromere-Related Repeats. *Genomics* **2014**, *104*, 587–593. [CrossRef] [PubMed]
26. Round, E.K.; Flowers, S.K.; Richards, E.J. Arabidopsis Thaliana Centromere Regions: Genetic Map Positions and Repetitive DNA Structure. *Genome Res.* **1997**, *7*, 1045–1053. [CrossRef]
27. Belser, C.; Istace, B.; Denis, E.; Dubarry, M.; Baurens, F.-C.; Falentin, C.; Genete, M.; Berrabah, W.; Chèvre, A.-M.; Delourme, R.; et al. Chromosome-Scale Assemblies of Plant Genomes Using Nanopore Long Reads and Optical Maps. *Nat. Plants* **2018**, *4*, 879–887. [CrossRef]
28. Wang, X.; Wang, H.; Wang, J.; Sun, R.; Wu, J.; Liu, S.; Bai, Y.; Mun, J.-H.; Bancroft, I.; Cheng, F.; et al. The Genome of the Mesopolyploid Crop Species Brassica Rapa. *Nat. Genet.* **2011**, *43*, 1035–1039. [CrossRef]
29. Nagaharu, U. Genome Analysis in Brassica with Special Reference to the Experimental Formation of B. Napus and Peculiar Mode of Fertilization. *J. Jpn. Bot.* **1935**, *7*, 389–452.
30. Rousseau-Gueutin, M.; Belser, C.; Da Silva, C.; Richard, G.; Istace, B.; Cruaud, C.; Falentin, C.; Boideau, F.; Boutte, J.; Delourme, R.; et al. Long-Read Assembly of the Brassica Napus Reference Genome Darmor-Bzh. *Gigascience* **2020**, *9*. [CrossRef]
31. Wick, R.; Github. Filtlong. Available online: https://github.com/rrwick/Filtlong (accessed on 30 July 2021).
32. Liu, H.; Wu, S.; Li, A.; Ruan, J. SMARTdenovo: A de Novo Assembler Using Long Noisy Reads. *Gigabyte* **2021**, *2021*, 1–9. [CrossRef]
33. Ruan, J.; Li, H. Fast and Accurate Long-Read Assembly with wtdbg2. *Nat. Methods* **2020**, *17*, 155–158. [CrossRef]
34. Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P.A. Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nat. Biotechnol.* **2019**, *37*, 540–546. [CrossRef]
35. Chen, Y.; Nie, F.; Xie, S.-Q.; Zheng, Y.-F.; Dai, Q.; Bray, T.; Wang, Y.-X.; Xing, J.-F.; Huang, Z.-J.; Wang, D.-P.; et al. Efficient Assembly of Nanopore Reads via Highly Accurate and Intact Error Correction. *Nat. Commun.* **2021**, *12*, 60. [CrossRef]
36. Github. Medaka. Available online: https://github.com/nanoporetech/medaka (accessed on 30 July 2021).
37. Istace, B.; Belser, C.; Aury, J.-M. BiSCoT: Improving Large Eukaryotic Genome Assemblies with Optical Maps. *PeerJ* **2020**, *8*, e10150. [CrossRef]
38. Dudchenko, O.; Batra, S.S.; Omer, A.D.; Nyquist, S.K.; Hoeger, M.; Durand, N.C.; Shamim, M.S.; Machol, I.; Lander, E.S.; Aiden, A.P.; et al. De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds. *Science* **2017**, *356*, 92–95. [CrossRef]
39. Github. Juicebox. Available online: https://github.com/aidenlab/Juicebox (accessed on 30 July 2021).
40. Github. Juicebox_scripts. Available online: https://github.com/phasegenomics/juicebox_scripts (accessed on 30 July 2021).

41. Liu, C. In Situ Hi-C Library Preparation for Plants to Study Their Three-Dimensional Chromatin Interactions on a Genome-Wide Scale. *Methods Mol. Biol.* **2017**, *1629*, 155–166. [PubMed]
42. Github. Pore-C-Snakemake. Available online: https://github.com/nanoporetech/Pore-C-Snakemake/issues/11 (accessed on 30 July 2021).
43. Github. Pore-C-Snakemake. Available online: https://github.com/nanoporetech/Pore-C-Snakemake (accessed on 30 July 2021).
44. Ghurye, J.; Pop, M.; Koren, S.; Bickhart, D.; Chin, C.-S. Scaffolding of Long Read Assemblies Using Long Range Contact Information. *BMC Genom.* **2017**, *18*, 527. [CrossRef]
45. Zhang, L.; Cai, X.; Wu, J.; Liu, M.; Grob, S.; Cheng, F.; Liang, J.; Cai, C.; Liu, Z.; Liu, B.; et al. Improved Brassica Rapa Reference Genome by Single-Molecule Sequencing and Chromosome Conformation Capture Technologies. *Hortic. Res.* **2018**, *5*, 50. [CrossRef] [PubMed]
46. Cabanettes, F.; Klopp, C. D-GENIES: Dot Plot Large Genomes in an Interactive, Efficient and Simple Way. *PeerJ* **2018**, *6*, e4958. [CrossRef]
47. De Givry, S.; Bouchez, M.; Chabrier, P.; Milan, D.; Schiez, T. Cathagene: Multipopulation Integrated Genetic and Radiated Hybrid Mapping. *Bioinformatics* **2004**, *14*, 2.
48. Lim, K.-B.; de Jong, H.; Yang, T.-J.; Park, J.-Y.; Kwon, S.-J.; Kim, J.S.; Lim, M.-H.; Kim, J.A.; Jin, M.; Jin, Y.-M.; et al. Characterization of rDNAs and Tandem Repeats in the Heterochromatin of Brassica Rapa. *Mol. Cells* **2005**, *19*, 436–444.
49. Lim, K.-B.; Yang, T.-J.; Hwang, Y.-J.; Kim, J.S.; Park, J.-Y.; Kwon, S.-J.; Kim, J.; Choi, B.-S.; Lim, M.-H.; Jin, M.; et al. Characterization of the Centromere and Peri-Centromere Retrotransposons in Brassica Rapa and Their Distribution in Related Brassica Species. *Plant. J.* **2007**, *49*, 173–183. [CrossRef]
50. Koo, D.-H.; Hong, C.P.; Batley, J.; Chung, Y.S.; Edwards, D.; Bang, J.-W.; Hur, Y.; Lim, Y.P. Rapid Divergence of Repetitive DNAs in Brassica Relatives. *Genomics* **2011**, *97*, 173–185. [CrossRef]
51. Song, J.-M.; Guan, Z.; Hu, J.; Guo, C.; Yang, Z.; Wang, S.; Liu, D.; Wang, B.; Lu, S.; Zhou, R.; et al. Eight High-Quality Genomes Reveal Pan-Genome Architecture and Ecotype Differentiation of Brassica Napus. *Nat. Plants* **2020**, *6*, 34–45. [CrossRef]
52. Morgulis, A.; Gertz, E.M.; Schäffer, A.A.; Agarwala, R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *J. Comput. Biol.* **2006**, *13*, 1028–1040. [CrossRef]
53. Kent, W.J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **2002**, *12*, 656–664. [CrossRef]
54. Birney, E.; Clamp, M.; Durbin, R. GeneWise and Genomewise. *Genome Res.* **2004**, *14*, 988–995. [CrossRef]
55. Dubarry, M.; Noel, B.; Rukwavu, T.; Aury, J.M. Gmove a Tool for Eukaryotic Gene Predictions Using Various Evidences. *F1000research Publ. Online* **2016**. [CrossRef]
56. Laver, T.; Harrison, J.; O'Neill, P.A.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D.J. Assessing the Performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **2015**, *3*, 1–8. [CrossRef]

9.3.2.3 Conclusion

L'assemblage des génomes, et en particulier des génomes de plantes, est un domaine qui connaît des évolutions technologiques et logicielles majeures. Il nécessite généralement la combinaison de plusieurs technologies, avec lesquelles il est important de se familiariser, pour obtenir des résultats de qualité. Dans cette étude, nous faisons un retour d'expérience autour de la génération du génome de *Brassica rapa* (cv. Z1) (version 2) en utilisant les techniques de séquençage long-read et de scaffolding avec des données dites "long range".

Généralement, la cartographie optique de Bionano Genomics est la technologie de choix pour organiser et valider les assemblages. Elle permet d'ancrer la plus grande quantité de nucléotides, notamment dans les régions complexes, comme celles qui présentent une forte proportion d'éléments répétés. Les centromères sont par exemple mieux reconstruits. De plus, elle permet d'estimer la taille des trous, contrairement aux méthodes de scaffolding basées sur des données Hi-C (ou Pore-C). Par contre, il peut être difficile d'obtenir la qualité d'ADN nécessaire à la préparation des marquages. Ceci rend cette technologie onéreuse car il est nécessaire de réaliser plusieurs tests d'extraction d'ADN (pour obtenir de l'ADN de très haut poids moléculaire) et de conditions de réaction de marquage (les enzymes étant très sensibles aux divers contaminants résiduels lors de l'extraction).

La technologie Hi-C est la technique de scaffolding la plus populaire, avec une communauté active développant régulièrement des outils et des méthodes. Deux nouveaux outils performants ont d'ailleurs été récemment publiés[118]. Les protocoles de préparation ne nécessitent pas d'extraire de l'ADN de très haut poids moléculaire mais nécessitent de disposer de tissus frais très rapidement congelés, ce qui rend plus accessible leur réalisation. Les séquenceurs Illumina sont largement distribués ce qui facilite également la mise en place de ces protocoles. Les limites de la technologie Hi-C sont la résolution des régions contenant de fortes répétitions et l'absence d'estimation des gaps.

En utilisant une seule ou en combinant soigneusement deux

méthodes, j'ai organisé les contigs de *Brassica rapa* cv. Z1 obtenus à partir d'un séquençage nanopore en scaffolds à l'échelle des chromosomes et obtenu l'assemblage du génome le plus continu pour cette espèce à ce jour.

9.4 VERS DES ASSEMBLAGES SIMPLIFIES

9.4.1 Introduction

A la suite de l'article paru dans *Nature Plants* en 2018, le génome de *Musa balbisiana* a été publié (*Nature Plants* 2019)[119]. Cet article présente le génome B de Musa assemblé à l'échelle des chromosomes grâce à des données PacBio et Hi-C.

Il est paru alors indispensable de fournir à la communauté un assemblage de qualité comparable pour *Musa acuminata* (génome A) en plus de celui de *Musa schizocarpa* (génome S). En effet, *Musa acuminata* avait été séquencé et assemblé une première fois en 2012 en utilisant des données de courtes lectures Illumina, du séquençage Sanger et 454 ainsi qu'une carte génétique trop peu dense[120]. En 2016, l'ajout de données de séquençage Mate Pair Illumina, d'une carte optique de faible résolution et d'une carte génétique plus dense avait permis d'améliorer la continuité de l'assemblage et de baisser le nombre de gaps[121]. 76% de la taille estimée du génome avait ainsi pu être ancrée. Le répertoire des gènes avait été également annoté.

Dans l'article présenté ci-après, j'ai effectué un nouvel assemblage pour la même variété : Pahang-HD, un haploïde doublé. L'utilisation de la carte optique et l'ajout d'une carte génétique m'ont permis d'obtenir un assemblage d'une très grande continuité et dont la séquence couvre les chromosomes d'un télomère à l'autre. L'évolution de la technologie Oxford Nanopore (conférant un taux d'erreur bien plus bas) et une extraction d'ADN avec de très grands fragments ont permis d'obtenir 5 des 11 chromosomes en un seul contig et l'ancrage d'environ 90% de la taille estimée du génome. Cette nouvelle version (V4) a permis la mise en évidence de clusters entiers de gènes de résistance et de la structure extrêmement complexe des centromères.

9.4.2 Article : "Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing" Belser, C., Baurens, FC., et al. Communication Biology. 2021

communications biology

ARTICLE

https://doi.org/10.1038/s42003-021-02559-3 OPEN

Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing

Caroline Belser[1,6], Franc-Christophe Baurens[2,3,6], Benjamin Noel[1], Guillaume Martin[2,3], Corinne Cruaud[4], Benjamin Istace[1], Nabila Yahiaoui[2,3], Karine Labadie[4], Eva Hřibová[5], Jaroslav Doležel[5], Arnaud Lemainque[4], Patrick Wincker[1], Angélique D'Hont[2,3] & Jean-Marc Aury[1✉]

Long-read technologies hold the promise to obtain more complete genome assemblies and to make them easier. Coupled with long-range technologies, they can reveal the architecture of complex regions, like centromeres or rDNA clusters. These technologies also make it possible to know the complete organization of chromosomes, which remained complicated before even when using genetic maps. However, generating a gapless and telomere-to-telomere assembly is still not trivial, and requires a combination of several technologies and the choice of suitable software. Here, we report a chromosome-scale assembly of a banana genome (*Musa acuminata*) generated using Oxford Nanopore long-reads. We generated a genome coverage of 177X from a single PromethION flowcell with near 17X with reads longer than 75 kbp. From the 11 chromosomes, 5 were entirely reconstructed in a single contig from telomere to telomere, revealing for the first time the content of complex regions like centromeres or clusters of paralogous genes.

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France. [2] CIRAD, UMR AGAP Institut, Montpellier, France. [3] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France. [4] Commissariat à l'Energie Atomique (CEA), Institut François Jacob, Genoscope, Evry, France. [5] Institute of Experimental Botany of the Czech Academy of Sciences, Centre of the Region Haná for Biotechnological and Agricultural Research, Olomouc, Czech Republic. [6] These authors contributed equally: Caroline Belser, Franc-Christophe Baurens. ✉email: jmaury@genoscope.cns.fr

Long-read technologies are now the standard for generating high-quality assemblies, especially for complex genomes such as plant genomes[1–4]. Although the impact of these technologies is undeniable, they still lack the maturity to reconstruct complete chromosome sequences from telomere to telomere. Generally, assemblies based on long-reads are complemented with long-range data, like optical maps or chromosomal conformation sequencing. Recently, the Telomere-to-Telomere (T2T) consortium proposed a telomere-to-telomere assembly of the X chromosome sequence of the human genome[5]. This high-quality assembly of the human genome was based on a combination of several existing technologies: nanopore sequencing from Oxford Nanopore Technology (ONT), single-molecule real-time (SMRT) sequencing provided by Pacific Biosciences (PACBIO), linked reads sequencing from 10X Genomics (10X) and optical mapping provided by Bionano Genomics (BNG). Even if the final assembly is very contiguous, there are still several gaps, and the complete X chromosome sequence was obtained by manual curation. This huge effort is not possible for all genome projects because it is far too expensive and time-consuming. It is clear that these multilayer assemblies reveal the architecture of complex regions as well as the complete organization of chromosomes, which remained complicated before. Long-range technologies make it possible to organize contigs based on long-reads but they are not able to fill the gaps between these contigs. Indeed, usually complex regions like centromeres or telomeres still contain many gaps, depending on their repetitive content.

We selected the banana genome, a medium-size genome in the plant lineage (~500 Mbp), and hypothesized that recent improvement of the ONT technology, coupled with dedicated DNA extraction protocol and efficient software enables the reconstruction of gapless and Telomere-to-Telomere chromosome sequences.

Banana species are monocotyledonous plants and part of the Zingiberales order and of the Musaceae family. Bananas are mostly cultivated in tropical and subtropical countries, and their fruits are the basis of the diet of several hundred million people and are massively exported to industrialized countries. Four genetic groups have been predicted to be involved in the origins of cultivars, mainly through inter(sub)specific hybridization and with different extents of contribution: *Musa acuminata* including various subspecies (A-genome), *Musa balbisiana* (B-genome), *Musa schizocarpa* (S-genome) and species of the *Australimusa* section (T-genome). Two events appeared during banana domestication: the transition from wild to edible diploids and the emergence of triploids from edible diploids[6–8]. Recent results suggest that edible cultivar origins are more complex than expected, involving multiple hybridization steps, resulting in inter(sub)specific mosaic genomes. They also revealed that additional genetic pools to the ones expected were involved, for which the wild contributors are still unidentified[6]. In addition, large structural variations in form of reciprocal translocations and a few inversions have been characterized in genetic pools involved in cultivar origins and found widespread in cultivated germplasm[6]. The complexity of these genomes underlies the importance of producing high-quality assemblies of banana genomes to decipher their evolutionary history and to support genetic studies.

In this context, two versions of the *Musa acuminata* 'DH-Pahang' genome have already been proposed[9,10]. The first draft version of the genome of this double haploid genotype (V1) was published in 2012 and based on 454, Sanger (fosmids and BAC-ends), and Illumina sequencing. Furthermore, scaffolds were organized using a sparse genetic map, resulting in the anchoring of 63% of the estimated genome size. The second version (V2) was published in 2016 and added Illumina long-insert sequences, a low-contiguity optical map as well as a more dense genetic map. Martin et al. proposed an assembly of the 11 chromosomes that included 76% of the estimated genome size. Herein we propose to generate a new version (named V4, the third version was an internal assembly not shared with the community) of the DH-Pahang assembly based on nanopore long-reads.

Results

Highly contiguous genome assembly of the banana genome. The efficiency of long-reads sequencing depends on the quality of the DNA extraction. Here, DNA was extracted following a plant-dedicated protocol provided by Oxford Nanopore Technologies ("High molecular weight gDNA extraction from plant leaves"). This protocol was particularly effective and allowed us to obtain long DNA fragments (> 50 kbp). Residual short fragments were filtered out using the Short Read Eliminator (SRE) XL kit (Circulomics, MD, USA). Almost 93 Gb of nanopore sequences were obtained with a single PromethION flowcell R9.4.1. The 5.2 M reads had a N50 of 31.6 kbp and the genome was covered at 17X with reads longer than 75 kbp. This high-quality set of long reads was assembled using several bioinformatics tools and the assembly obtained with NECAT[11] was retained. Indeed the NECAT assembly of this haploid cultivar was the most contiguous (contig N50) and had a larger cumulative size. This assembly was polished first using long reads with Racon[12] and Medaka[13] and then using Illumina short-reads with Hapo-G[14]. This assembly, based on nanopore long reads and without long-range information, was composed of 124 contigs (larger than 50 kbp) and had a cumulative size of 485 Mbp. Half of the assembly size was composed of contigs larger than 32 Mbp and only 16 contigs covered 90% of the total length (Table 1, contig N50 and contig L90). More importantly, the seven largest contigs had a size compatible with complete chromosomes (ranging from 47.7

Table 1 Comparison of *Musa acuminata* (DH-Pahang) genome assemblies.

	D'hont et al. [9] V1	Martin et al. [10] V2	This study V4
Number of contigs	29,437	19,312	124
Cumulative size (bp)	390,477,446	405,522,014	484,058,756
N50 (bp)	28,319	43,237	32,091,396
L50	3,428	2,363	7
N90 (bp)	5,108	9,026	6,704,534
L90	16,551	10,327	16
Longest contig (bp)	306,660	602,020	47,719,527
Number of chromosomes	11	11	11
Cumulative size (bp)	331,812,599	397,008,016	468,821,802
Cumulative size (ACGT only)	286,824,765	363,519,833	468,133,046
N50 (bp)	30,470,408	37,593,364	43,931,232
L50	5	5	5
N90 (bp)	25,514,024	29,070,452	34,826,100
L90	10	10	10
Longest (bp)	35,439,739 A08	44,889,171 A08	51,314,288 A08
% of *N*	13.56%	8.44%	0.68%
% of estimated genome	63.4%	75.9%	89.6%
% of estimated genome (ACGT only)	54.8%	69.5%	89.5%
Complete BUSCO ($N = 1,614$)	92.6%	98.5%	98.8%

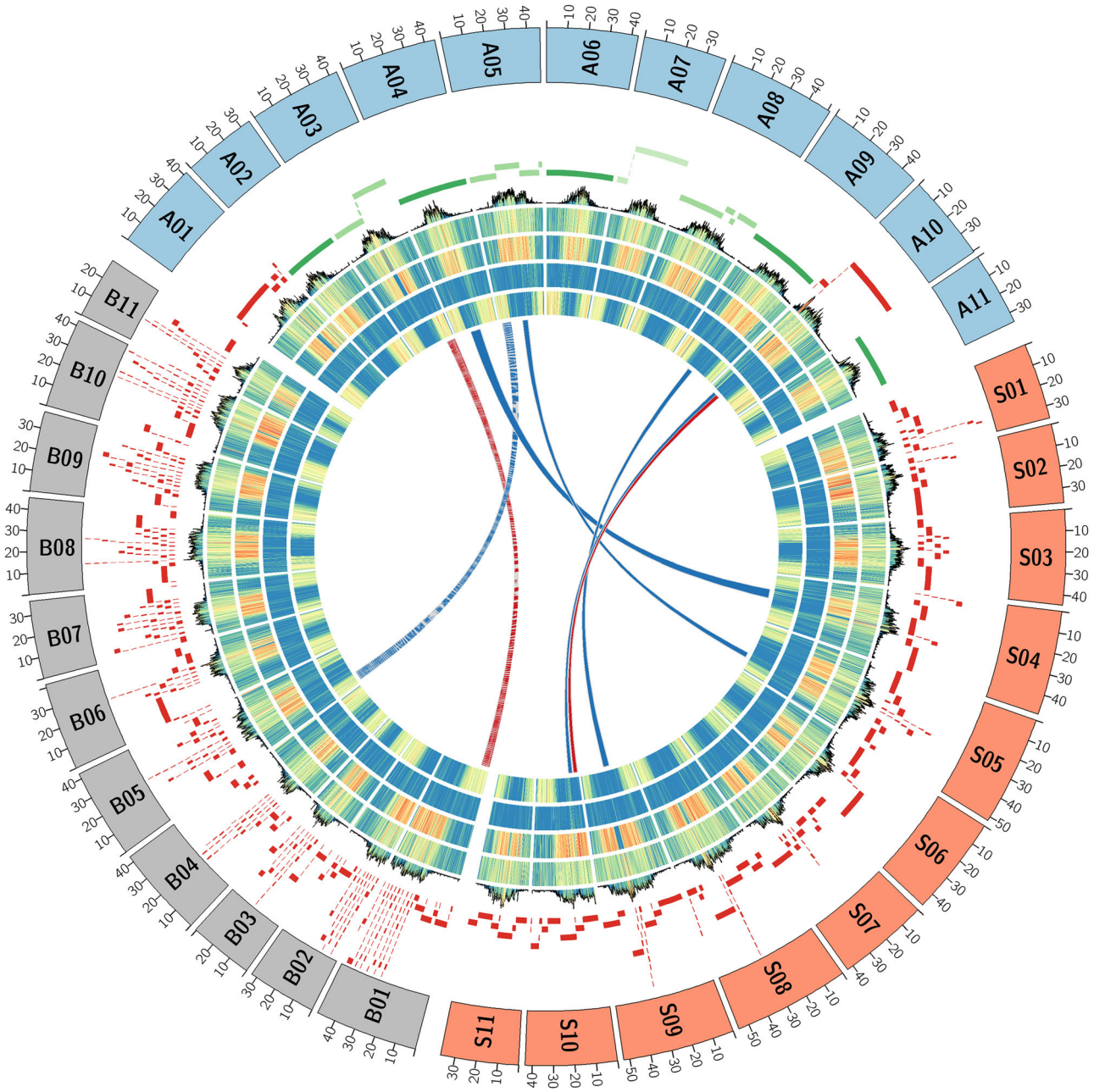


Fig. 1 Musa genomes architecture comparison. The tracks represent the following elements (from outer to inner): (1) schematic representation of *M. acuminata* (A), *M. balbisiana* (B) and *M. schizocarpa* (S) chromosome sequences, (2) contigs colored in green if the chromosome sequence is composed of 1−4 contigs, in red if the chromosome sequence is composed of more than 5 contigs. (3) Density of the centromeric repeats. (4) Density of the Gypsy elements. (5) Density of the Copia elements. (6) Density of the DNA transposons. (7) Density of genes. (8) Synteny relationships. The red lines show translocations between B01 and A03 and between S10 and A10. The blue lines show inversions between B05 and A05, S04 and A04, S05 and A05, S09 and A09.

to 32.1 Mbp). The anchoring of contigs was performed following the methodology described in Martin et al. [10]. As expected, the five largest contigs correspond to complete chromosome sequences and harbor telomeric repeats at both extremities (Fig. 1). The six remaining chromosome sequences were composed of a small number of contigs (between four and eight). Interestingly, the remaining gaps are mainly located in rDNA clusters: 5S for chromosomes 1, 3, and 8 and 45S for chromosome 10 or in other tandem and inverted repeats: chromosomes 1 and 5 (Fig. 2a). These rDNA clusters are composed of a large number of tandemly repeated genes and are generally very difficult to assemble. Even if these clusters still contain a few gaps, it is now possible to decipher the architecture of these large and complex regions. In addition, smaller contigs not anchored to the 11 chromosomes correspond to the chloroplastic and mitochondrial genomes (one and 45 contigs, respectively). A total of 37 contigs were filtered out because they were included in larger contigs and contained highly repeated sequences.

Validation of telomere-to-telomere chromosome sequences. A kmer analysis and a first alignment of the largest contigs with the previous version of the DH-Pahang assembly did not reveal chimeric contigs (Supplementary Figs. 1 and 2). In addition, all

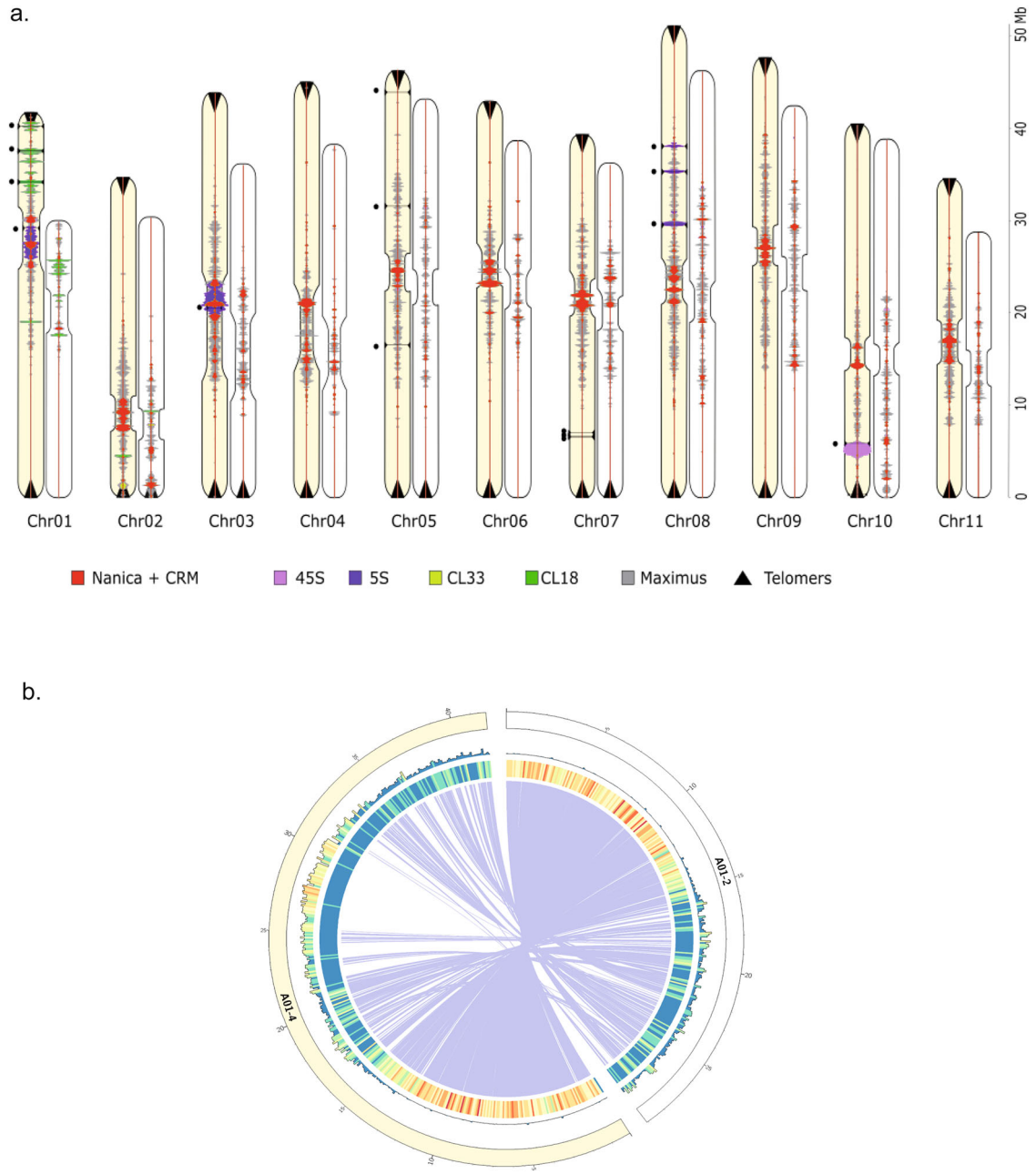


Fig. 2 Comparison of the V2 and V4 assemblies. a Localization and density of several repeated elements on chromosome sequences of the V4 (light orange) and V2 (white) assemblies (scale in Mbp on the right) with Nanica LINE and CRM chromovirus Gypsy retrotransposon (red), 5S rDNA (blue), 45S rDNA (violin), tandem repeat cluster CL18 (dark green), tandem repeat cluster CL33 (light green), Maximus Copia retrotransposon (gray) and telomeric sequences (black triangles). Horizontal black lines and black dots correspond to the 15 remaining gaps in the V4 assembly. **b** Comparison of the A01 chromosomes of the V2 and V4 assemblies. Tracks represent the following elements (from outer to inner): (1) density of the centromeric repeats. (2) Density of genes. (3) Synteny relationships between the V2 chromosome 1 and the V4 chromosome 1.

eleven chromosome sequences harbor plant-specific telomeric repeats (T3AG3) at both sides, underlining the complete assembly of chromosome ends.

However, we decided to validate the quality of our assembly using two Bionano optical maps that were generated using the Saphyr instrument commercialized by Bionano Genomics (BNG). High molecular weight DNA was extracted and labeled using two different labeling chemistries independently, the Direct Label chemistry (DLS) and the Nick-Label-Repair and Stain chemistry (NLRS) based on nicking endonucleases. Two optical maps were generated using DLS with the DLE-1 enzyme and NLRS with the BspQI enzyme. The resulting DLE-1 and BspQI optical maps were 469 and 474 Mbp lengths, respectively, and had a N50 of 35 and 16 Mbp, respectively. We used these two optical maps to first validate the contigs, and then order and orient them. As a result, only one contig of 380 kbp, composed of tandem repeated elements, was flagged as conflictual with the optical maps and split into two contigs (Supplementary Fig. 3). All other contigs were in accordance with the maps, which strongly validate the accuracy of the NECAT assembler. The 124 contigs were ordered in 96 scaffolds using the Bionano Solve workflow and the BiscoT[15] software (88 scaffolds correspond each to one contig). In the end, eight of the eleven chromosomes are represented by a single scaffold and the other four remain in two scaffolds. The whole-

genome assembly contains only fifteen gaps that are concentrated in large highly repetitive regions (Supplementary Table 1).

As Illumina paired-end were available for the DH-Pahang genome, an assessment of quality and completeness was performed using Merqury[16] (Supplementary Table 2). As expected the reported completeness is higher for the long-read assembly (95.7% compared to 98.1%). However, the consensus quality (QV) is lower for the nanopore assembly (38.8 vs 49.2). This lower value can be explained by the fact that firstly the nanopore assembly contains regions that are not present in the short-read assembly (such as repetitive regions that are more difficult to polish and can generally contain more errors in nanopore assemblies) and that secondly, the error rate of the nanopore technology is still too high, complicating the correction of the consensus using polishing algorithms. As a control, we calculated the QV score only on the regions shared between both assemblies, and observed a decrease in the difference in quality between the two assemblies (45.9 versus 50.1). This difference may be due to the even higher error rate in the consensus of nanopore assemblies compared to short-read assemblies.

Comparison of *Musa acuminata* assemblies. Unsurprisingly, compared to previous versions, the contiguity of our DH-Pahang assembly is greatly improved. The contig N50 goes from a few tens of kbp (28 and 43 kbp for V1 and V2 respectively) to a few tens of Mbp (32 Mbp). More importantly, the cumulative size is closer to the estimated genome size, suggesting that complex regions are better represented in this new release (Table 1). With a very small number of contigs, anchoring on the eleven chromosomes using the genetic map was easier especially in the centromeric regions, which are generally difficult to organize due to their lower density of genetic markers. The size of the DH-Pahang genome was estimated by flow cytometry at 523 Mbp[9]. The 11 chromosome sequences of our long-read assembly cover almost 90% of this estimated size, while the first two versions were largely incomplete (55 and 70% respectively, Table 1). As a consequence, the assembled size of each chromosome sequence has increased, between 7% for chromosome 10−43% for chromosome 1 (Fig. 2a and Supplementary Fig. 2). Chromosome sequences of the two versions were aligned and large genomic regions (>100 kbp) absent in the previous assembly were reported. The 11 chromosome sequences totalized 247 new regions that covered 141.4 Mbp, i.e., 29.2% of the assembly (Supplementary Fig. 4). The largest region, close to 6 Mbp, is localized on chromosome 1 (Fig. 2b). Unsurprisingly, these blocks are mainly composed of repeated elements (more than 85%), and localized in centromeric regions or rDNA clusters (Supplementary Fig. 5). We annotated 246 Mbp of the genome (52.6%) as transposable elements (TE), compared to 152 Mbp in V2, which illustrates the much better representation and completion of these repetitive elements in the V4 assembly (Supplementary Table 3 and Supplementary Fig. 6). Figure 2a shows the distribution of several tandem repeats and TE along the chromosomes, including the Maximus Copia retrotransposons, which are the most abundant TEs in the Pahang genome.

Architecture of centromeric regions and rDNA clusters. Earlier cytogenetic analysis showed that a long interspersed element (LINE), named Nanica, is present in the centromeric regions of banana[9,17]. A very few LINE sequences were present in the first release of the assembly despite being present in unassembled reads[9] and they had a scattered distribution on the pericentromeric and centromeric regions of the V2 assembly. In this new assembly, clusters of Nanica tandem repetitions are found grouped in the centromeric regions of all chromosomes (Fig. 2a and Supplementary Fig. 6). Several elements of chromovirus CRM clade, a lineage of Ty3/Gypsy retrotransposons, were also found restricted to these centromeric regions. Some members of this plant retroelement have been shown to have the ability to target their insertion almost exclusively to the functional centromeres[18,19]. The position of two other tandem repeats (CL18 and CL33) previously identified[17] could also be refined between V2 and V4 and the localization of the main clusters on chromosomes 1 and 2 are in accordance with cytogenetic karyotypes[17]. Regarding the 5S rDNA sequences, in the V2 assembly, they were present in a few numbers in chromosomes 5, 9, and 8 spanning 7,5 kbp of sequences (around 130 gene units) (Supplementary Table 4). In the new assembly, six major loci containing 5S rDNA gene clusters are present accounting for around 7,696 gene units. Three clusters are located on one arm of chromosome 8 representing in total around 2.2 Mbp and two large clusters of 5S rDNA repeat have been integrated to chromosome 1 and 3 centromeric regions, representing around 3.5 and 4 Mbp, respectively. These results are in accordance with previous cytogenetic results[17] and the position of the rDNA clusters have been clarified showing that they colocalize with the centromeric Nanica clusters of these chromosomes. Clusters of 5S rDNA are organized in canonical gene/spacer tandem repeat of different lengths due to the insertion in the spacer of various repeated elements such as Nanica or CRM sequences as observed in the 5S cluster of the centromeric regions of chromosomes 1 and 3 (Fig. 3A, B and c). Furthermore, a large cluster of 1.8 Mbp containing around 110 45S rDNA units, consisting in canonical gene/spacer tandem repeat, is localized on chromosome 10 between positions 4.4 and 6.2 Mbp (Fig. 2a).

Tandemly duplicated genes (TDGs). Gene duplication is an important evolutionary mechanism that contributes to the appearance of novel functions and to adaptation. The events leading to gene duplication have contributed to important plant agronomic traits, such as grain quality, fruit shape, and flowering time[20]. A special case of gene duplication relates to genomic/tandem duplication events, which generate, locally, repetitive regions in the genome. These TDGs are generally harder to capture in short-read assemblies, especially in the case of recent multi-copy clusters. We found 1,700 genes that have been annotated only in our long-read assembly. These genes are distributed over the different chromosomes, with chromosomes 1 and 10 having the greatest number of new genes, 13.3 and 14.3% respectively (Supplementary Table 5). Interestingly, a large proportion (38.3%) of these genes are TDGs included in a gene cluster and the proportion of TDGs in new genes is higher when compared to the whole gene catalog (38.3% versus 9.9%).

By focusing on TDGs and detecting gene clusters in the short and long-read assemblies, we found 31% more clusters in the V4 compared to V2 assembly (1,134 compared to 866 clusters). These blocks of TDGs contain respectively 3,649 and 1,134 genes. The largest in the long-read assembly contains 38 genes on chromosome 7 (between 31.8 and 32.8 Mbp) and was split into two smaller clusters of 11 and 9 genes in the V2. This TDG cluster is located in a region with several gaps in previous versions, and we found three regions (165, 111 and 110 kbp) between positions 31.8 and 32.3 Mbp of the chromosome 7 that are specific to the V4 assembly. These new regions allow the creation of a complete cluster of TDGs (Fig. 4a) which contain motifs of the terpene synthase family that are responsible for the synthesis of terpenoid compounds playing a role in plant flavor[21] and more generally in the interactions between the plant and its environment[22,23]. This family is known to contain TDGs and is expanded in several plant species[24].

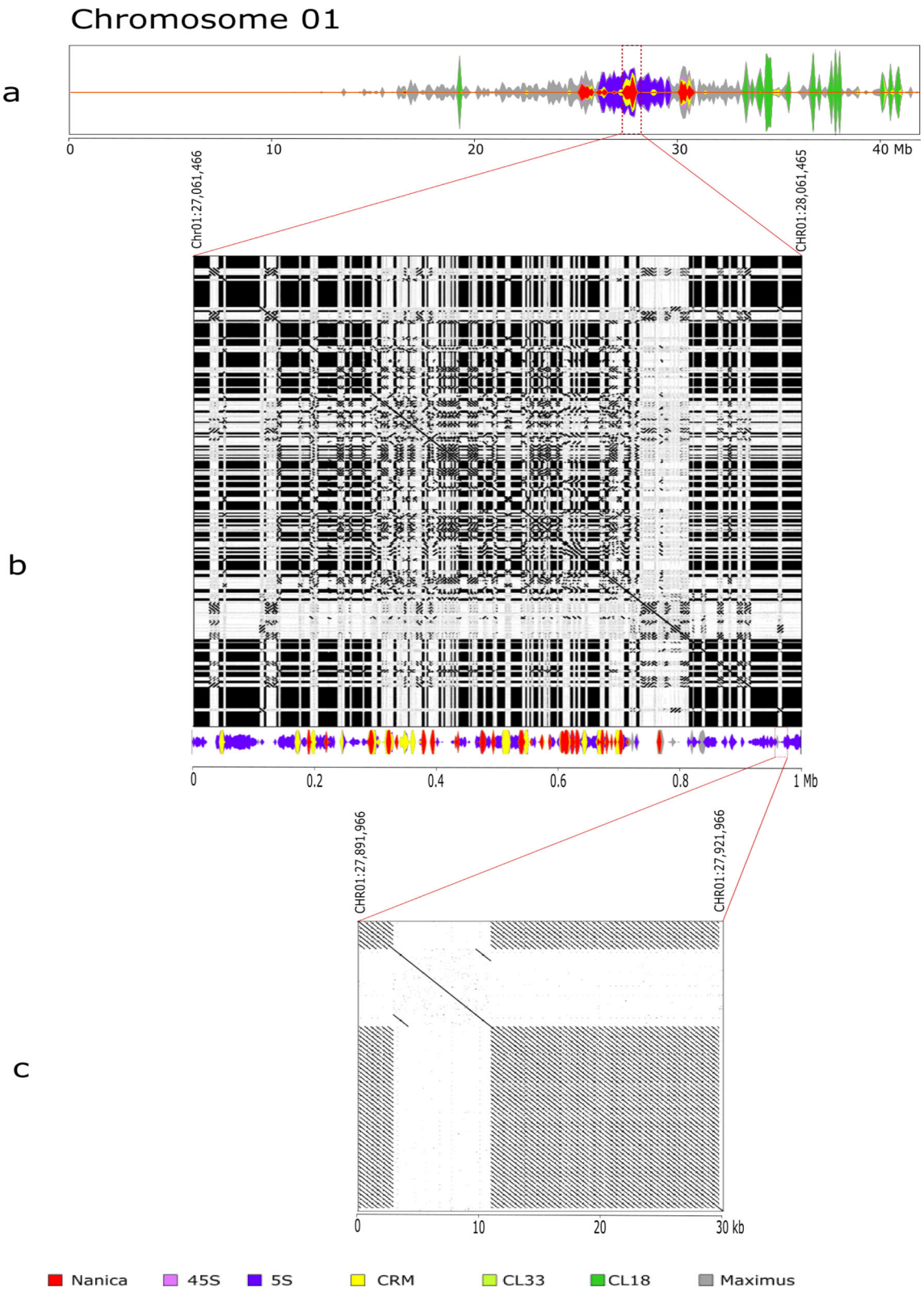


Fig. 3 Fine structure and density of main (peri)centromeric repeated sequences on chromosome 1. Nanica LINE (red), CRM chromovirus Gypsy retrotransposon (yellow), 5S rDNA (lilac), tandem repeat cluster CL18 (dark green), Maximus Copia retrotransposon (gray) are represented on: (**a**) the entire chromosome 1: (**b**) a zoom and a dot-plot alignment of a 1 Mbp segment in the centromeric region containing Nanica, 5S rDNA, and CMR repeats, (**c**) a zoom and a dot-plot alignment of a 30 kbp segment containing 5S rDNA repeats and a CMR.

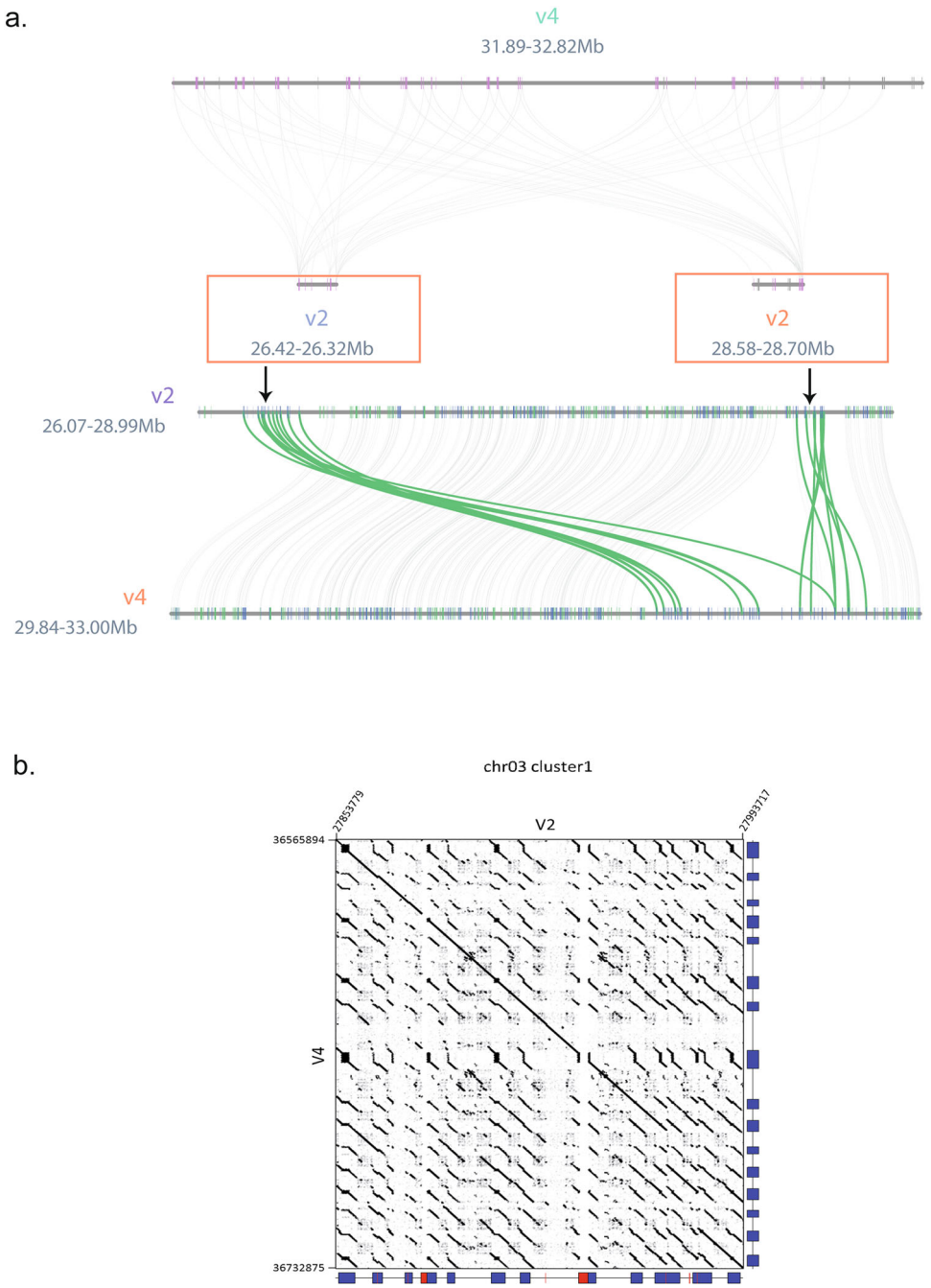


Fig. 4 Comparison of tandemly duplicated gene regions in the V2 and V4 assemblies. a Synteny visualization of gene clusters of the terpene synthase family. The cluster is located between 31.89 and 32.82 Mbp on the V4 (chromosome 7). Gene synteny relationships are colored in green, and genes are colored according to their orientation (blue if forward and green otherwise). **b** Comparison of the structure of a NLR cluster on chromosome 3. The predicted NLR loci for each version are represented by blue boxes on the *x*- and *y*-axis of the dot plots. Red boxes represent regions bearing undetermined nucleotides. Region coordinates are also indicated.

Resistance genes. Plant disease resistance genes encoding proteins with nucleotide-binding leucine-rich repeat (NLR) domains are often clustered in genomes, sometimes forming large, rapidly evolving clusters of highly homologous genes[25]. The NLR-annotator program[26] allows the identification of NLR loci i.e., genomic regions likely associated with an NLR gene (or pseudogene). A total of 128 NLR loci were detected in this assembly compared to 111 loci in the V2 assembly (Table 2, and Supplementary Data 1 and 2). Four major clusters of NLR loci were found in this assembly: two in chromosome 3, one in chromosome 7, one in chromosome 10 (Fig. 4b and Supplementary Fig. 8). They all have a larger size compared to V2, with sizes ranging from 132 up to 227 kbp and additional detected NLR loci. These clusters were improved in sequence quality with a complete absence of undetermined nucleotides within the corresponding genomic regions (Supplementary Table 6).

Comparison of A, B and S-genome assemblies. The B and S genomes have already been recently sequenced using a long-read strategy[27,28]. Taking into account this new version of the *M. acuminata* genome, three high-quality banana genomes are now available. The A and S genomes were sequenced using ONT while

Table 2 Comparison of gene prediction statistics.

Reference	Martin et al. [10] V2	This study V4
# Number of genes	35,276	36,979
#Exons per spliced gene (avg:med)	6.08: 5	6.05: 4
Gene sizes (avg:med)	4,542: 2,824	4,604.68: 2,758
CDS sizes (avg:med)	1,171: 981	1,180: 972
Complete BUSCO ($N = 1,614$)	98.5%	98.8%
NLR loci	111	128

Table 3 Comparison of A (*Musa acuminata*), B (*Musa balbisiana*) and S (*Musa schizocarpa*) genomes assemblies.

	Musa acuminata	*Musa balbisiana*	*Musa schizocarpa*
Number of contigs	124	3,787	379
Cumulative size	484,058,756	491,421,783	517,486,196
N50 (bp)	32,091,396	1,801,976	6,493,909
L50	7	59	24
N90 (bp)	6,704,534	56,360	1,047,001
L90	16	578	84
Longest contig (bp)	47,719,527	14,987,599	18,138,554
Number of chromosomes	11	11	11
Cumulative size	468,821,802	430,021,147	496,921,565
Cumulative size (ACGT only)	468,133,046	429,290,714	490,105,212
% Anchored sequences	96.8%	87.3%	94.7%
N50 (bp)	43,931,232	42,323,520	46,993,692
L50	5	5	5
N90 (bp)	34,826,100	30,518,812	36,762,080
L90	10	10	10
Longest (bp)	51,314,288	48,736,620	54,858,060
Number of gaps	15	683	166
Estimated genome size	523 Mb	520 Mb	587 Mb
% of estimated genome size	89.6%	82.6%	84.6%
Number of annotated genes	36,979	35,148	32,809
Complete BUSCO ($N = 1,614$)	98.8%	96.9%	97.6%

the B genome was sequenced using the PACBIO technology. Interestingly, the two ONT assemblies have a higher contiguity (contig N50 of 32 and 6.5 Mbp compared to 1.8 Mbp) suggesting the usage of longer reads, or the difficulty to extract and sequence long DNA fragments with the PACBIO device (Table 3). Indeed, the PACBIO library was size-selected in order to obtain fragments around 20 kbp[28], which was perhaps an optimal condition for PACBIO sequencing at this time. The B genome was assembled from reads with an N50 of 16.6 kbp whereas A and S genomes were assembled with reads having a N50 of 31.6 and 24.4 kbp respectively. As a consequence, in addition, we noticed that chromosome sequences of the A and S genomes contain fewer gaps. A difference between the PACBIO and ONT sequencing technologies is already mentioned[29]. The eleven A and S chromosome sequences contain 15 and 166 gaps respectively whereas B chromosome sequences contain 683 gaps and no chromosome sequence is gapless (Fig. 1). Centromeric regions, detected with centromeric repeats, are very fractionated in the case of the PACBIO-based assembly (from 24 contigs for the chromosome 7 to 111 contigs for the chromosome 1), underlying the importance of ultra-long reads to resolve these highly repetitive regions. However, sequencing technologies evolve rapidly and this comparison does not reflect the current capability of each technology.

Overall, the synteny conservation between the three genomes is high, we detected one inversion between the chromosome B05 of *Musa balbisiana* and the chromosome A05 of *Musa acuminata* and a translocation between the chromosome B01 and the chromosome A03 as already reported[28] (Fig. 1 and Supplementary Fig. 9). Four inversions between *Musa schizocarpa* and *Musa acuminata* were also detected on the chromosomes S10, S04, S05, and S09 and one translocation on chromosome A10 (Fig. 1 and Supplementary Fig. 10). The corresponding regions on chromosome A10 contain the 45S rDNAS gene cluster. The contigs organization in these five regions was manually validated in the two genome assemblies using optical maps.

Discussion

Long read sequencing technology emergence has paved the way for high-quality genome assemblies. The rapid evolution of the DNA extraction protocols now allows the community to sequence very long DNA sequences. Coupled with the evolution of the bioinformatic tools, the generation of high-quality assemblies has been greatly simplified. The latest improvements of the ONT technology, especially the base-calling efficiency, result in a decrease of the error rate. Several assembly tools were specially developed around long reads and are able to manage noisy reads and have their specificity, as well as a specific margin of progress. We think that it is still important to use the latest release of several assemblers and choose the most efficient for each genome assembly project.

In this study, we combined recent development from DNA extraction, sequencing, and genome assembly and showed that plant chromosome sequences can now be assembled in a single contig, gapless, and from telomere to telomere, at least to a certain extent. We chose the genome of *Musa acuminata*, the first monocotyledonous species sequenced outside Poales, because its reference genome, even of low-quality, has been widely used and constitutes an important resource for the scientific community. To date, three *Musa* species: *Musa acuminata*, *Musa balbisiana,* and *Musa schizocarpa* are available at the chromosome-scale. These three species are of particular interest because they are involved in the origin of banana cultivars; their high-quality genome assemblies will thus be a valuable resource to explore the evolutionary history and biology of current banana cultivars. We reported that the PACBIO assembly of *M. balbisiana* is more fragmented which can be related to the input read size and underline the importance of long reads to resolve highly repetitive regions.

We generated a highly contiguous assembly of the eleven chromosome sequences of *Musa acuminata* of which five were obtained in a single contig. At the same time, optical maps were used to validate the nanopore assembly. It is important to mention that only one small contig, essentially composed of repetitive elements, was detected as a potential chimera underlining the high quality of the contigs produced by the NECAT assembler. However, it should be noted that homozygous material was used and this limited complexity may not be representative of the situation in other species. All eleven chromosome sequences, build with the help of a genetic map, contain telomeric repeats at both ends, which is an important element in asserting on the one hand that the reconstruction of the chromosome sequences is of good quality and on the other hand that the still missing part of the genome is contained in the remaining fifteen gaps, although 7 are of unknown length. One of the advantages of optical maps is that the size of the gaps can be estimated if the map is sufficiently contiguous (Supplementary Table 1 and Supplementary Fig. 11).

Comparison of the distribution of repeated sequences (tandem repeat and TE) between V2 and V4 showed that the integration of these elements that were typically difficult to assemble with past technologies are greatly improved in the new assembly and are now very congruent with cytogenetic karyotype. All centromeres are now clearly identified with large clusters of Nanica LINE tandem repeats and CMR TE, and in addition, for two of them large clusters of 5S rDNA tandem repeats. Such a case of recruitment of 5S rDNA gene array in centromere was also reported in one of the switchgrass chromosomes[30]. This high-resolution of centromeric regions opens new avenues to study how satellites repeats originate and evolve in the centromeric region and more generally to better understand the organization and functioning of centromeres that are essential chromosomal domains for kinetochore assembly and correct chromosome segregation[31,32]. In addition, chromosome reciprocal translocations were recently shown to have accompanied subspecies evolution in *Musa*[6], and some of them have their breakpoints in centromeric regions. Having access to the sequence of these centromeric regions will permit investigating the mechanism and sequences involved in the origin of these translocations. Finally, comparison with the *Musa acuminata* V2 assembly highlights a higher proportion of each class of transposable elements, and a large amount of additional sequences in the centromeric regions, like Nanica elements, or large retro-transposon derivatives.

It is often mistakenly thought that short-read assemblies are complete at the gene level. This hypothesis is mainly based on the results given by the BUSCO software, which only focuses on single-copy genes. Accordingly, the gene content completeness was already high, in previous *M. acuminata* assembly versions, according to the BUSCO score. However, here, using ultralong reads, we were able to assemble many additional copies of TDG clusters which contain important gene families like terpene synthases or disease resistance genes. Banana crops are currently particularly threatened by diseases including Black leaf streak disease that requires massive use of pesticide[33] and by a new strain of Fusarium wilt (Tropical Race 4) that is currently spreading around the world and for which no chemical control is possible[34]. This new assembly will facilitate the search for resistance genes to these devastating diseases[35].

Finally, we showed that gapless and telomere-to-telomere assembly of chromosome sequences is now possible thanks to long-read sequencing, at least in the case of homozygous genomes. The critical point remains the DNA extraction protocols that generally need adaptation for each species. These closed assemblies will allow new discoveries and will shed new light on these genomes in particular in complex repetitive regions such as centromeres, which have essential biological function but are so far poorly characterized.

Methods

Plant material. Double haploid *Musa acuminata* spp *malaccensis* (*DH-Pahang*) plant material was obtained from the CRB Plantes Tropicales Antilles CIRAD-INRA Guadeloupe under the collection number PT-BA-00461.

DNA extraction. For Illumina sequencing libraries, DNA was extracted using a modified mixed alkyl trimethyl ammonium bromide (MATAB) procedure[36]. 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 mgl^{-1} RNAse A. Crude extracts were maintained for 20 min at 74 °C, extracted with an equal volume of chloroform-isoamyl alcohol (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 kbp, a A260/A280 ratio close to 2 and a A260/A230 ratio above 1.5 were kept.

In order to generate long reads on the Oxford Nanopore Technologies devices, high-quality and high-molecular-weight DNA is needed. To that end, DNA was isolated following the protocol provided by Oxford Nanopore Technologies, "High molecular weight gDNA extraction from plant leaves" downloaded from the ONT Community in March, 2019 (CTAB-Genomic-tip). This protocol involves a conventional CTAB extraction followed by purification using the commercial Qiagen Genomic tip (QIAGEN, MD, USA), but size selection was performed using Short Read Eliminator XL (Circulomics, MD, USA) instead of AMPure XP beads. Briefly, 1.2 g of leaves were cryoground in liquid nitrogen. The fine powder was transferred to 20 mL of Carlson buffer (100 mM Tris-HCl pH 9.5, 2% CTAB, 1.4 M NaCl, 1% PEG 8000, 20 mM EDTA, 0.25% b-mercaptoethanol (v/v)) prewarmed to 65 °C. Then 40 µl of RNase A (100 mg/ml) was added before incubation at 65 °C for 1 h (with intermittent agitation). Proteins removal was performed by addition of one volume of chloroform and centrifugation at 5,500×*g* for 10 min at 4 °C. DNA was then precipitated with 0.7 V of isopropanol and centrifugation at 5,500×*g* for 30 min at 4 °C. The pellet was then purified using the Qiagen Genomic-tip 100/G, following the manufacturer's instruction: DNA pellet was first dissolved at 50 °C for 15 min in 9.5 mL of G2 buffer before loading onto the pre-equilibrated Genomic-tip column. Purified gDNA was finally precipitated with 0.7 volumes of isopropanol, washed with 2 ml of 70% ethanol, dried, and eluted in 100 µL of TE Buffer. DNA was quantified by a dsDNA-specific fluorometric quantitation method using Qubit dsDNA HS Assays (ThermoFisher Scientific, Waltham, MA). DNA quality was checked on a 2200 TapeStation automated electrophoresis system (Agilent, CA, USA) (Supplementary Fig. 12).

Generating optical maps requires high molecular weight (HMW) DNA. Here HMW DNA of *M. acuminata DH Pahang* was prepared according to Safář et al. [37] with several modifications. Briefly, 0.5 cm long segments of leaf midribs and young leaf tissues were fixed for 20 min at 4 °C in Tris buffer (10 mM Tris, 10 mM EDTA, 100 mM NaCl, pH 7.5) containing 2% formaldehyde. After three 5 min washes in Tris buffer, the segments were homogenized using chopping by a razor blade in petri dish containing 1 ml of ice-cold IB buffer (15 mM Tris, 10 mM EDTA, 130 mM KCl, 20 mM NaCl, 1 mM spermine, 1 mM spermidine and 0.1% Triton X-100, pH 9.4) and immediately before use, 33 µl of β-mercaptoethanol were added to 10 ml of IB buffer. Nuclei suspension was passed through a 50 µm nylon mesh and stained with DAPI at a final concentration of 2 µg/mL. Six batches of 900,000 G1-phase nuclei were sorted into 77 µl of IB buffer with β-mercaptoethanol in 1.5 ml polystyrene tubes using a FACSAria SORP flow cytometer and sorter (Becton Dickinson, San José, CA, United States) equipped with solid-state UV laser. One 20 µl agarose mini-plug was prepared from each batch of nuclei according to Šimková et al. [38]. Miniplugs were washed and solubilized using agarase enzyme (Thermo Fisher Scientific) to release high molecular weight (HMW) DNA. HMW DNA was further purified by drop dialysis and was then homogenized a few days prior to the quality control.

The concentration and purity of the extracted DNA were evaluated using a Qubit fluorometer (Thermo Fisher Scientific) and a Nanodrop spectrophotometer (Thermo Fisher Scientific). DNA integrity was checked by pulsed-field gel electrophoresis (Pippin Pulse, Sage Science). DNA molecules were detectable between 50 and 300 kbp in size.

Illumina PCR-free library preparation and sequencing. DNA (1.5 µg) was sonicated to a 100−1500 bp size range using a Covaris E220 sonicator (Covaris, Woburn, MA, USA). The fragments were end-repaired and 3′-adenylated. Illumina adapters were added using the Kapa Hyper Prep Kit (KapaBiosystems, Wilmington, MA, USA). The ligation products were purified with AMPure XP beads (Beckman Coulter Genomics, Danvers, MA, USA). The libraries were quantified by qPCR 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, Santa Clara, CA, USA). The libraries were sequenced on an Illumina HiSeq2500 instrument (Illumina, San Diego, CA, USA) using 250 base-length read chemistry in paired-end mode.

After the Illumina sequencing, an in-house quality control process was applied to the reads that passed the Illumina quality filters. The first step discards 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[39]. The last step identifies and discards read pairs that are mapped to the phage phiX genome, using SOAP aligner[40] and the Enterobacteria phage PhiX174 reference sequence (GenBank: NC_001422.1). This processing, described in Alberti et al. [41], resulted in high-quality data.

PromethION library preparation and sequencing. The library was prepared according to the following protocol, using the Oxford Nanopore SQK-LSK109 kit. Genomic DNA fragments (4 µg) were repaired and 3′-adenylated with the NEBNext FFPE DNA Repair Mix and the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). Sequencing adapters provided by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford,

UK) were then ligated using the NEBNext Quick Ligation Module (NEB). After purification with AMPure XP beads (Beckmann Coulter, Brea, CA, USA), half of the library was mixed with the sequencing buffer (ONT) and the loading bead (ONT) and loaded on a PromethION R9.4.1 flow cell. The second half of the library was loaded on the flow cell after a Nuclease Flush using the Flow Cell Wash Kit EXP-WSH003 (ONT) according to the Oxford Nanopore protocol. Reads were basecalled using Guppy version 4.0.1. The nanopore long reads were not cleaned and raw reads were used for genome assembly.

Optical mapping. The Direct Label and Stain (DLS) labeling (using the DLE-1 enzyme) and the Nick Label Repair and Stain (NLRS) labeling (using the BspQI enzyme) protocols were performed according to Bionano Genomics with 750 and 600 ng of DNA respectively. The Chip loadings were performed as recommended by Bionano Genomics.

Long reads-based genome assembly. We generated three samples of reads: all reads, 30X of the longest reads, and 30X of the filtlong[42] highest-score reads. We then applied four different assemblers, Smartdenovo[43], Redbean[44], Flye[45], and NECAT[11] on these three subsets of reads (Supplementary Table 7), with the exception of NECAT being only launched with all reads, as it applies a down-sampling algorithm in its pipeline. Smartdenovo was launched with -k 17, as advised by the developers in case of larger genomes and -c 1 to generate a consensus sequence. Redbean was launched with '-xont -X5000 -g450m' and Flye with '-g 450m'. NECAT was launched with a genome size of 450 Mbp and other parameters were left as default. After the assembly phase, we selected the best assembly (NECAT with all reads) based on the cumulative size and contiguity. The assembler output was polished one time using Racon[12] with Nanopore reads, then one time with Medaka[13] and Nanopore reads, and two times with Hapo-G[14] and Illumina PCR-free reads (Supplementary Table 8).

Assembly validation. The DLE-1 map was generated using the Direct Label and Stain (DLS) technology and the BspQI map using the Nick Label Repair and Stain (NLRS) technology. Genome map assemblies were performed using Bionano Solve Pipeline version 3.3 and Bionano Access version 1.3.0. We used the parameter "Add Pre-Assembly" which produced a rough assembly. This first result was used as a reference for a second assembly, using the parameters "non-haplotype without extend and split". We filtered out molecules smaller than 150 kbp and molecules with less than nine labeling sites (Supplementary Tables 9 and 10). The nanopore contigs were then validated using the two Bionano maps and organized with the scaffolding procedure provided by Bionano Genomics (Supplementary Fig. 13). Negative gap sizes were checked and corrected using the BiscoT software[15] to avoid artifactual genomic duplications (Supplementary Table 1). As recommended by the BiscoT authors, we performed a last iteration of Hapo-G[14] to polish merged regions (Supplementary Table 11). In order to obtain a quality score used to compare the different versions of the assembly, we downloaded and used merqury[16] version 1.3 (git commit 6b5405e). We first used the included best_k.sh script with a tolerable collision rate of 0.0001 and a genome size of 500 Mbp, which gave us an estimated best k-mer size of 21. Then, we used meryl[46] version 1.3 (git commit 3400615) to compute the reads k-mer counts via the meryl count command with default parameters. Merqury was then launched on two sets of sequences. The first one consists of the V1 and V4 assemblies, in order to compare them in their globality. As the V4 assembly is larger than V1, we aligned the V1 assembly to the V4 assembly using minimap2 and kept alignments that were larger than 50 kbp. Regions of both assemblies corresponding to these alignments were extracted and used as a second set that we used as input to merqury, to compare assemblies only in regions that are shared.

Chromosome sequences reconstruction. DH-Pahang sequences were anchored on chromosomes using segregating markers obtained from the selfing of the 'Pahang' accession PT-BA-00267, described in Martin et al. [10] (Supplementary Table 12 and Supplementary Fig. 14). Data are available on the Banana Genome Hub[47] in the download section under '*AF-Pahang marker matrix file*' and '*AF-Pahang marker sequence (FASTA)*' for coded segregating markers and marker sequence respectively. Sequences anchoring was performed following methodology described in Martin et al. [10]. The complete process was performed using scaffhunter tools[48] available at the South Green platform.

In addition, based on scaffold BLAST against *Musa acuminata* chloroplast sequence[49], *Musa acuminata* putative 12 mitochondrial scaffolds[10] and *Phoenix dactylifera* protein sequences[50], 1 putative chloroplastic (corresponding to one initial contig), and 45 putative mitochondrial scaffolds (corresponding to 45 initial contigs) were identified in the assembly. The 45 putative mitochondrial scaffolds were grouped into a sequence named *putative_mito_scaff*. The chloroplastic scaffold was discarded from the assembly as the chloroplast genome of DH-Pahang was already fully assembled and published[49].

The 37 remaining scaffolds (cumulative size of 5.2 Mbp) (corresponding each to one initial contig) showed a strong BLAST homology to larger scaffolds included in the chromosome sequences (36 with more than 95% of their length and 1 with more than 88% of its length). Investigation (dot-plot analysis using gepard v1.30[51] and BLAST against nr/nt of ncbi) of these scaffolds revealed a repetitive nature, most of them corresponding to rDNA sequences. Because of their strong homology to scaffolds included in the chromosome sequences these scaffolds were discarded from the assembly.

Gene prediction. Repeats in the genome assembly were masked using Tandem Repeat Finder[52] for tandem repeats and RepeatMasker[53] for simple repeats, as well as known repeats included in RepBase[54]. In addition, known *Musa* transposable elements (from D'Hont et al. [9]), were detected using RepeatMasker.

Gene prediction was done using proteomes from homologous species, *Musa acuminata* (UP000012960), *Oryza. sativa* (UP000059680), *Phoenix dactylifera* (UP000228380), *Musa schizocarpa* (www.genoscope.cns.fr/plants) and *Musa balbisiana* (banana-genome-hub.southgreen.fr).

The proteomes were aligned against the genome assembly in two steps. Firstly, BLAT[55] (default parameters) was used to quickly localize corresponding putative genes of the proteins on the genome. The best match and matches with a score ≥ 90% of the best match score were retained. Secondly, the alignments were refined using Genewise[56] (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.

To allow the detection of UTRs in the gene prediction step, we aligned not only the protein of *M. acuminata*, but also the virtual mRNAs of the *M. acuminata* predicted genes[57] on a masked version of the genome assembly. Then, transcript sequences of *M. acuminata* predicted genes were aligned by BLAT (default parameters) on the masked genome assembly. Only the alignments with an identity percent greater or equal to 90% were kept. For each transcript, the best match was selected based on the alignment score. Finally, alignments were recomputed in the previously identified genomic regions by Est2Genome[58] in order to define precisely intron boundaries. Alignments were kept if more than 80% of the length of the transcript was aligned to the genome with a minimal identity percent of 95%.

To proceed to the gene prediction, we integrated the protein homologies and transcript mapping using a combiner called Gmove[59]. This tool can find CDSs based on genome located evidence without any 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 (ORFs) consistent with the protein evidence. A selection step was applied to all candidate genes, essentially based on gene structure. Also, all gene predictions included in a genomic region tagged as transposable elements were removed from the gene set (Table 2). The completeness of the predicted genes was assessed with BUSCO[60] version 5 (embryophyta dataset odb10).

The search for NLR loci was performed using NLR-annotator tools[26] that scan specifically the 6 reading frames of the nucleotide sequence for the presence of 19 NLR-associated motifs and reconstruct a potential NLR locus which might correspond to a complete or partial gene and might also be a pseudogene.

Transposable element detection. Transposable elements were detected using RepeatMasker[53] associated with the TE Musa library[9] and CR sequences[18]. The same procedure was used to detect TEs in the *Musa acuminata* V2, V4, *Musa balbisiana* and *Musa schizocarpa* assemblies. The gff output file was converted into a bed file and the TE coverage was calculated using bedtools[61] coverage (version v2.29.2-17-ga9dc5335) on a 100 kbp window. Centromeric boundaries were defined using the density of daterra-Maximus, ITS-5S, ITS-18S, ITS-26S, Nanica, maca-Angela, caturra-Reina elements. TEs were grouped following Wicker et al. classification[62].

Genome assemblies comparisons. The synteny relationships between *Musa balbisiana*, *Musa schizocarpa* and *Musa acuminata* V4 were determined using Assemblytics[63]. First, genome sequences were aligned against each other using nucmer[64] version 3.23 (-maxmatch -l 100 -c 500) as recommended by the Assemblytics authors. Assemblytics was launched on the nucmer delta file (unique_length_required = 10000). Figures 1 and 2b were generated using the Circos software[65]. In the same way, each chromosome sequence of the *Musa acuminata* V4 was aligned against its relative chromosome sequence of *Musa acuminata* V2 using nucmer version 3.23 (-r −1 -l 10000) and dot plots were generated using the mummerplot command.

Detection of specific regions of the V4 assembly. New regions of the *Musa acuminata* V4 were determined using blast[66] (ncbi-tools/6.1.20120620) alignment between the chromosomes of each assembly version. Regions of the V4 assembly larger than 100 kbp without any alignment to the V2 assembly were considered new. In addition, the *Musa acuminata* V2 gene predictions were aligned (see Gene prediction section) on the V4 assembly, and the positions of the genes were compared with the V4 gene catalog using bedtools[61] (version bedtools-2.29.2) (-v option). Genes from the V4 assembly without any correspondence in the V2 were considered new.

Detection of tandemly duplicated genes. An all-against-all comparison of the *Musa acuminata* V4 proteins was performed using Diamond[67] (version 0.9.24). Mapping output was filtered according to the following parameters: an e-value

lower than 10e−20 and a coverage of the smallest protein greater than 80%. Genes were considered as tandemly duplicated if they were co-localized on the same chromosome and not distant from more than 10 genes to each other. Figure 4a was realized using the MCscan tool[68] with the two following commands: jcvi.compara.synteny (--iter=1) and jcvi.graphics.synteny (--glyphcolor = orientation).

Statistics and reproducibility. No statistical tests were used in this study, and to allow the reproducibility of our analysis and results, all the sequencing data are available in public databases and the scripts developed to generate the figures are available on Zenodo and on a Github repository, as described in the data and code availability sections.

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

Data availability

All the supporting data are included in three additional files which contain (a) Supplementary Tables 1−12 and Supplementary Figs. 1−14, (b) Supplementary Data 1 (position of NLR genes in the V4 assembly) and (c) Supplementary Data 2 (position of NLR genes in the V2 assembly). The genome assembly is freely available at http://www.genoscope.cns.fr/plants and http://banana-genome-hub.southgreen.fr. The ONT, Illumina, and Bionano Genomics data are available in the European Nucleotide Archive under the following projects PRJEB35002.

Code availability

All the code and data used to generate the figures are available on Zenodo[69] and on a Github repository https://github.com/institut-de-genomique/Pahang-associated-data

Received: 11 June 2021; Accepted: 13 August 2021;
Published online: 07 September 2021

References

1. Michael, T. P. & VanBuren, R. Building near-complete plant genomes. *Curr. Opin. Plant Biol.* **54**, 26–33 (2020).
2. Rousseau-Gueutin, M. et al. Long-read assembly of the Brassica napus reference genome Darmor-bzh. *GigaScience* **9**, giaa137 (2020).
3. Zhang, W. et al. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat. Commun.* **11**, 3719 (2020).
4. Schmidt, M. H.-W. et al. De novo assembly of a New *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* **29**, 2336–2348 (2017).
5. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
6. Martin, G. et al. Genome ancestry mosaics reveal multiple and cryptic contributors to cultivated banana. *Plant J.* **102**, 1008–1025 (2020).
7. Němečková, A. et al. Molecular and cytogenetic study of East African Highland Banana. *Front. Plant Sci.* **9**, 1371(2018).
8. Langhe, E. D., Vrydaghs, L., Maret, P., de, Perrier, X. & Denham, T. Why bananas matter: an introduction to the history of banana domestication. *Ethnobot. Res. Appl* **7**, 165–177 (2009).
9. D'Hont, A. et al. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213–217 (2012).
10. 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).
11. Chen, Y. et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **12**, 60 (2021).
12. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
13. *nanoporetech/medaka*. (Oxford Nanopore Technologies, 2021).
14. Aury, J.-M. & Istace, B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genom. Bioinform.* **3**, lqab034 (2021).
15. Istace, B., Belser, C. & Aury, J.-M. BiSCoT: improving large eukaryotic genome assemblies with optical maps. *PeerJ* **8**, e10150 (2020).
16. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
17. Čížková, J. et al. Molecular analysis and genomic organization of major DNA satellites in banana (Musa spp.). *PLoS One* **8**, e54808 (2013).
18. Tran, T. D. et al. Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus Genlisea.*Plant J. Cell Mol. Biol.* **84**, 1087–1099 (2015).
19. Neumann, P. et al. Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA* **2**, 4 (2011).
20. Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294–2316 (2016).
21. Del Terra, L. et al. Functional characterization of three Coffea arabica L. monoterpene synthases: Insights into the enzymatic machinery of coffee aroma. *Phytochemistry* **89**, 6–14 (2013).
22. Jiang, S.-Y., Jin, J., Sarojam, R. & Ramachandran, S. A comprehensive survey on the terpene synthase gene family provides new insight into its evolutionary patterns. *Genome Biol. Evol.* **11**, 2078–2098 (2019).
23. Falara, V. et al. The tomato terpene synthase gene family. *Plant Physiol.* **157**, 770–789 (2011).
24. Martin, D. M. et al. Functional annotation, genome organization and phylogeny of the grapevine (Vitis vinifera) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.* **10**, 226 (2010).
25. Wersch, Svan & Li, X. Stronger when together: clustering of plant NLR disease resistance genes. *Trends Plant Sci.* **24**, 688–699 (2019).
26. Steuernagel, B. et al. The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482 (2020).
27. Belser, C. et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **4**, 879–887 (2018).
28. Wang, Z. et al. Musa balbisiana genome reveals subgenome evolution and functional divergence. *Nat. Plants* **5**, 810–821 (2019).
29. Lang, D. et al. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience* **9**, giaa123 (2020).
30. Yang, X. et al. Amplification and adaptation of centromeric repeats in polyploid switchgrass species. *N. Phytol.* **218**, 1645–1657 (2018).
31. Miga, K. H. Centromere studies in the era of 'telomere-to-telomere' genomics. *Exp. Cell Res.* **394**, 112127 (2020).
32. Comai, L., Maheshwari, S. & Marimuthu, M. P. A. Plant centromeres. *Curr. Opin. Plant Biol.* **36**, 158–167 (2017).
33. Bellaire, L., de, L., de, Fouré, E., Abadie, C. & Carlier, J. Black leaf streak disease is challenging the banana industry. *Fruits* **65**, 327–342 (2010).
34. Kema, G. H. J. et al. Editorial: Fusarium wilt of banana, a recurring threat to global banana production. *Front. Plant Sci.* **11**, 628888 (2021).
35. Ahmad, F. et al. Genetic mapping of Fusarium wilt resistance in a wild banana Musa acuminata ssp. malaccensis accession. *Theor. Appl. Genet.* **133**, 3409–3418 (2020).
36. Gawel, N. J. & Jarret, R. L. A modified CTAB DNA extraction procedure forMusa andIpomoea. *Plant Mol. Biol. Rep.* **9**, 262–266 (1991).
37. Safár, J. et al. Creation of a BAC resource to study the structure and evolution of the banana (Musa balbisiana) genome. *Genome* **47**, 1182–1191 (2004).
38. Šimková, H., Číhalíková, J., Vrána, J., Lysák, M. A. & Doležel, J. Preparation of HMW DNA from plant nuclei and chromosomes isolated from root tips. *Biol. Plant.* **46**, 369–373 (2003).
39. Engelen S., Aury J. M. fastxtend https://www.genoscope.cns.fr/externe/fastxtend/.
40. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
41. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 170093 (2017).
42. rrwick/Filtlong. quality filtering tool for long reads https://github.com/rrwick/Filtlong.
43. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte* **2021**, 1–9 (2021).
44. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).
45. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
46. Miller, J. R. et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).
47. Droc, G. et al. The banana genome hub. *Database* **2013**, bat035 (2013).
48. *SouthGreenPlatform/scaffhunter*. (South Green Bioinformatics platform, 2019).
49. Martin, G., Baurens, F.-C., Cardi, C., Aury, J.-M. & D'Hont, A. The complete chloroplast genome of banana (Musa acuminata, Zingiberales): insight into plastid monocotyledon evolution. *PLoS One* **8**, e67350 (2013).
50. Fang, Y. et al. A complete sequence and transcriptomic analyses of date palm (Phoenix dactylifera L.) mitochondrial genome. *PLoS One* **7**, e37164 (2012).
51. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
52. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
53. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker http://repeatmasker.org/.
54. Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).

55. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
56. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988 (2004).
57. 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).
58. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci. CABIOS* **13**, 477–478 (1997).
59. Dubarry, M. et al. Gmove a tool for eukaryotic gene predictions using various evidences. *F1000Research* **5** (2016).
60. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
62. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
63. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
64. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
65. Krzywinski, M. I. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* https://doi.org/10.1101/gr.092759.109 (2009).
66. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
67. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
68. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
69. Belser, C. et al. Musa acuminata DH-Pahang genome assembly: associated data. Zenodo https://doi.org/10.5281/zenodo.5120019 (2021).

Acknowledgements

This work was supported by the Genoscope, the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), France Génomique (ANR-10-INBS-09-08), the Center de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) and Agropolis Fondation (ID 1504-006) 'GenomeHarvest' project through the French Investissements d'avenir program (Labex Agro: ANR- 10-LABX-0001-01). EH and JD were supported by ERDF project 'Plants as a tool for sustainable global development' (No. CZ.02.1.01/0.0/0.0/16_019/0000827). The authors thank the staff of Oxford Nanopore Technology Ltd for technical help, Jitka Weiserová, Eva Jahnová and Dr. Jan Vrána for their help with the material preparation, the CRB Plantes Tropicales Antilles CIRAD-INRA Guadeloupe France for providing the plant materials and the CIRAD – UMR AGAP HPC Data Center of the South Green Bioinformatics platform (http://www.southgreen.fr) for providing computational resources.

Author contributions

K.L. extracted the sequenced DNA. E.H., and J.D. prepared HMW DNA for optical mapping. K.L. extracted the plugs. C.C. realized the bionano experiments. K.L., C.C., and A.L. optimized and performed the sequencing. C.B., B.I., B.N., N.Y., F.C.B., G.M., and J.M.A. performed the bioinformatic analyses. A.D. and J.M.A. conceived the project. C.B., B.I., B.N., K.L., C.C., E.H., G.M., A.D., and J.M.A. wrote the article. A.D., P.W., and J.M.A. supervised the study.

Competing interests

The authors declare the following competing interests, J.M.A. received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences. J.M.A. and C.B. received accommodation expenses to speak during Bionano Genomics user meetings. The authors declare no other competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-02559-3.

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

Peer review information *Communications Biology* thanks David Studholme and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Caitlin Karniski and Brooke LaFlamme. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

9.4.3 Conclusion

L'amélioration des logiciels de basecalling, des flow cells et du taux d'erreur (~3%) ont amélioré la qualité du séquençage obtenu grâce à la technologie Oxford Nanopore.

En utilisant les outils d'assemblage récents, comme Necat[52] qui est particulièrement performant sur les génomes de plantes, il m'a été possible de reconstituer le génome de *Musa acuminata* avec des chromosomes allant d'un télomère à l'autre. Cinq des onze chromosomes sont constitués d'un seul contig nanopore, ce qui montre qu'il sera peut être possible dans le futur de s'affranchir des technologies long range, allégeant le coût de production des génomes.

L'amélioration des protocoles d'extraction d'ADN reste un point essentiel pour séquencer des lectures toujours plus grandes et résoudre les régions les plus complexes des génomes. Dans notre cas, l'utilisation du protocole nanopore (protocole Plant Leave) a permis d'obtenir des fragments de plus de 75Kb.

J'ai effectué la comparaison de la version 4 de l'assemblage avec la version 2 et montré une augmentation de la taille de chaque chromosome ainsi que la présence des répétitions télomériques à chaque extrémité. Nous avons également mis en évidence de nouveaux clusters de gènes rDNA sur les chromosomes 1, 3 et 10 et la présence de repeats Nanica dans tous les centromères. Ces répétitions de type LINE étaient absentes de la version 2 de l'assemblage mais présentes dans les données de séquençage. Elles sont particulièrement difficiles à assembler. De la même façon, des clusters de gènes dupliqués en tandem ont été reconstruits, comme les clusters de gènes codant pour les terpene synthase ou des gènes contenant des motifs NLR (nucleotide-binding leucine-rich repeat) impliqués dans la résistance aux pathogènes.

Les cultures de bananiers sont d'ailleurs sensibles au fusarium qui cause des ravages. Ce champignon est particulièrement invasif et décime les plantations. Il n'existe aucun traitement hormis la mise en quarantaine des sols, provoquant des pertes économiques

considérables. Ceci montre qu'il est essentiel de fournir une référence de qualité pour permettre de rechercher une résistance naturelle à ce champignon. Le spectre de la chlordécone, pesticide largement utilisé dans les plantations de bananiers au XXème siècle pour lutter contre le charançon, est encore présent[122]. Le scandale sanitaire qui en a découlé a mis en lumière la nécessité de ne plus utiliser de produits chimiques hautement nocifs.

Ainsi, ces trois génomes *Musa acuminata*, *Musa schizocarpa* et *Musa balbisiana* forment une ressource très précieuse pour toute la communauté. Récemment le génome d'*Ensete glaucum* ou bananier des neiges a été publié dans la revue *GigaScience*[123]. Les *Ensete*, comme les *Musa* et *Musella* appartiennent aux *Musaceae* dans l'ordre des *Zingiberales* (gingembres et bananes). Le groupe des Musa a divergé des deux autres il y a environ 40 millions d'années[124]. *Ensete glaucum* porte des fruits non comestibles ressemblant à des bananes. Les auteurs ont utilisé le génome de *Musa acuminata* pour mettre en évidence les réarrangements chromosomiques permettant de passer des 11 chromosomes du génome A aux 9 chromosomes du génome des *Ensete*. Des translocations et des fusions dans les chromosomes mais également des modifications dans le contenu en répétitions ont conduit, durant le processus d'évolution, à la divergence entre ces deux groupes génétiques.

Cet article a été largement relayé par la société Oxford Nanopore car il met en avant l'utilisation des très grandes lectures et l'obtention de 5 chromosomes à partir d'un seul contig. J'ai d'ailleurs été invitée à présenter ces résultats lors d'un webinaire[125].

9.5 DISCUSSION ET PERSPECTIVES

Obtenir des génomes de qualité sert de socle aux diverses recherches menées sur les organismes étudiés. Pour reconstruire la séquence de ces génomes, de nombreux développements technologiques et méthodologiques ont vu le jour. Les technologies de séquençage ont évolué afin d'augmenter la taille des lectures obtenues ainsi que le débit, rendant le coût accessible aux équipes de recherche. Les génomes de référence produits se multiplient, accélérant la recherche dans bien des domaines.

Il convient ainsi d'être toujours à la pointe, de tester, de développer, de comparer les outils nécessaires pour obtenir une séquence complète de chacun des chromosomes, de télomère à télomère. En effet, les outils évoluent vite et les standards de qualité ne font que progresser. Il faut sans cesse s'aligner sur ces nouveaux standards qui conditionnent la publication des génomes et garantissent un excellent niveau de qualité

9.5.1 Grandes initiatives de séquençage de génomes de référence

Cette course est illustrée par certaines grandes initiatives telles que le T2T (Telomere To Telomere) qui a publié le premier génome humain de télomère à télomère quasi complet en 2022 dans *Science*[126]. De nombreux génomes humains avaient été publiés par le passé mais, même s'ils étaient de bonne qualité, ils n'étaient pas complets. L'assemblage de référence GRCh38, issu du Human Genome Project et publié en 2003, contient ainsi 151 méga-paires de bases (Mbp) de séquences inconnues réparties dans l'ensemble du génome, notamment dans des régions péri centromériques et subtélomériques, des duplications segmentales récentes, des clusters de gènes amplifiés et des clusters d'ADN ribosomique (ADNr). De plus, la séquence du bras court du chromosome 21 présente des erreurs d'assemblage, et des répétitions en tandem du satellite humain HSat sont absentes[32,126]. Le génome de référence produit par le T2T est le fruit d'une collaboration entre plusieurs laboratoires et a été généré à partir d'une combinaison de toutes les technologies existantes : Pacific Biosciences (High-Fidelity reads), Oxford Nanopore Technology (ultra-long reads), Illumina (banque sans amplification par PCR pour éviter les biais d'amplification c'est à dire banque PCR free), Arima Genomics (Hi-C), Bionano Genomics (carte optique) et séquençage Single-Cell. Une méthodologie particulière a été développée par les auteurs, illustrant le besoin de s'adapter aux organismes étudiés. Même s'il est sans commune mesure avec celui du projet Human Genome de 2003, le coût de cet assemblage parfait reste très élevé et ne peut s'appliquer à tous les organismes. Il ne peut être envisagé que pour les organismes modèles bénéficiant d'une communauté active et de financements adaptés.

Même si ce génome de référence est primordial pour toutes les études

sur le génome humain, il reste essentiel de générer plusieurs génomes de référence pour capturer l'ensemble de la variabilité génétique de l'espèce humaine[127]. L'individu séquencé ne peut pas, à lui seul, représenter l'ensemble de cette variabilité et il faut encore appréhender la diversité des allèles pour chaque gène. Le 1000 Genomes Project a permis de créer un catalogue des variations génétiques humaines courantes, en utilisant des échantillons provenant de personnes volontaires et qui se sont déclarées en bonne santé.

D'autres projets sont, quant à eux, consacrés à la conservation et à l'inventaire de la biodiversité comme ERGA (https://www.erga-biodiversity.eu/), le Darwin Tree of Life (https://www.darwintreeoflife.org/) ou le Vertebrate Genome Project (https://vertebrategenomesproject.org/). Le Genoscope est impliqué dans deux de ces programmes: ERGA et ATLASea. ERGA (European Reference Genome Atlas) regroupe de nombreux laboratoires européens pour la collecte d'échantillons, le séquençage, l'assemblage et l'annotation des génomes de référence. Cet effort a pour but de préserver l'information génétique de la biodiversité européenne dont un cinquième des 200,000 espèces eucaryotes est en danger d'extinction. De son côté, ATLASea a pour objectif d'établir 5000 génomes de référence de la biodiversité eucaryote (plancton, algues, animaux…) du littoral maritime français (pour la métropole et les territoires ultra marins), en ciblant les écosystèmes particulièrement menacés, fragiles, biologiquement importants et économiquement stratégiques. Le but est de mieux comprendre ces biotopes, d'explorer les voies de synthèse de molécules d'intérêt et d'étudier l'impact d'espèces invasives. Suivre la dynamique de ces écosystèmes marins permettra d'anticiper leur préservation. Ce projet se déroulera sur 7 à 10 ans et va nécessiter la mise en place de procédures robustes mais également une veille technologique et une grande adaptabilité (des protocoles d'extraction d'ADN jusqu'aux méthodes d'assemblage et d'annotation). Les organismes ciblés et leur génome seront très divers en termes de complexité. Ce projet représente une collecte de données génomiques essentielles pour les générations futures.

9.5.2 Etude de pan-génomes

Accéder à la structure détaillée des génomes est donc aujourd'hui réalisable et grâce à cela nous sommes en mesure de mener de larges études de pan-génomique. Dans le cas de l'espèce *Brassica rapa*, plusieurs génomes de référence ont été produits. Cette espèce présente une grande variété de morphotypes (chou, navet, sarson,...). Une seule séquence génomique ne permettrait pas de capturer l'ensemble de la diversité ni de satisfaire les besoins de recherche fonctionnelle. Plusieurs génomes de référence au sein des divers morphotypes et écotypes sont nécessaires pour mieux comprendre les bases génétiques de cette différenciation[128]. Dans l'article de Cai *et al.*, 2021, un pan-génome a été établi sur la base de 18 génomes de *Brassica*[129] *rapa*. Des variations structurales ont été identifiées, mettant à jour leur rôle dans les processus de domestication. Un pangénome regroupe les informations de l'ensemble des génomes séquencés pour une espèce donnée. Il permet de capturer toute la diversité allélique. On distingue ainsi le core génome, contenant les gènes communs, et le génome accessoire, contenant les gènes spécifiques d'une variété. Un pan-génome réalisé à partir de 54 lignées de *Brachypodium distachyon* est constitué de deux fois plus de gènes qu'un génome unique[130]. Un grand nombre de gènes trouvés dans le génome accessoire sont impliqués dans la réponse au stress biotique. Chez la tomate, les gènes de résistance sont également contenus dans le génome accessoire[131].

9.5.3 Etude des régions répétées

Assembler les régions répétées des génomes de plantes était un défi jusqu'à l'apparition des longues lectures. Les outils disponibles ne pouvaient pas gérer un trop grand nombre de répétitions parfois en tandem. Aujourd'hui, l'accès à la structure en éléments transposables est une richesse pour comprendre la dynamique des génomes[132]. Ces éléments sont mobiles et peuvent causer des modifications dans la composition des gènes et leur fonction[133]. Des études ont par ailleurs mis en relation la taille des génomes de plantes avec leur contenu en certaines classes de retrotransposons[76]. Des analyses génomiques

comparatives réalisées sur 166 accessions de riz ont montré que les espèces ayant des types de génome différents (AA, BB, CC ou EE) peuvent avoir connu des amplifications différentes de leurs rétrotransposons au cours de leur évolution, ce qui a entraîné des tailles de génome remarquablement différentes.

De même, il était très difficile d'estimer le nombre de copies des gènes dupliqués en tandem. Or, chez les coraux par exemple, les gènes dupliqués en tandem auraient une importance dans la réponse immunitaire et potentiellement la longévité de certaines espèces de coraux[134]. Ces multiples copies de gènes semblent provenir de multiples événements de duplication en tandem dans la branche phylogénétique des coraux. L'analyse des familles de gènes dupliqués, réalisée sur deux nouveaux génomes de coraux, illustre l'importance de générer des assemblages de très grande qualité.

9.5.4 Evolution des technologies de séquençage

Ainsi, les développements méthodologiques et technologiques à venir devront permettre de générer la séquence complète des génomes, même les plus complexes, en utilisant le moins de ressources possible, sans nécessiter l'utilisation de différentes technologies, et à moindre coût. Pour les organismes contenant des génomes plus simples, c'est-à-dire non polyploïdes et homozygotes, les assemblages de télomère à télomère pourront désormais être obtenus à partir d'une seule stratégie de séquençage. C'est ce que nous avons voulu illustrer par la réalisation du génome de référence du bananier (*Musa acuminata*) qui présente 5 chromosomes reconstitués intégralement grâce aux seules lectures Nanopore et à une extraction d'ADN de grande qualité. Pour le reste des organismes, il faudra continuer à combiner différentes stratégies et accentuer les développements méthodologiques.

Les technologies de séquençage à courtes lectures sont en train de vivre une petite révolution avec une perte du monopole de la société Illumina. De nouvelles entreprises commercialisent des instruments qui concurrencent la gamme des séquenceurs Illumina. On peut citer le DNBSEQ de chez MGI qui a connu des débuts difficiles de mise sur le marché, le AVITI de chez Element Biosciences ou le UG100 de chez Ultima Genomics. Toutes ces machines offrent des coûts de

séquençage inférieurs à ceux d'Illumina pour une qualité identique voire supérieure. En réponse, Illumina a présenté fin septembre une nouvelle gamme d'instruments avec un débit accru, une nouvelle chimie et la possibilité de séquencer des fragments plus longs. Le coût du séquençage d'un génome humain serait divisé au moins par 3 sur les machines à plus haut débit. L'entreprise a également beaucoup communiqué autour de la réduction des déchets. Toutes ces innovations sont possibles grâce à l'amélioration des supports de séquençage, de l'optique et à la miniaturisation toujours plus poussée des volumes de réactifs mais également des composants électroniques. Nous aurons donc à évaluer l'intérêt d'utiliser ces nouvelles technologies pour l'ensemble de nos applications.

Du côté du séquençage longues lectures, la société PacBio va proposer au cours de l'année 2023 un nouvel instrument permettant d'accroître son débit. Si on associe ce débit à la qualité des lectures obtenues avec la technologie HiFi, il devrait être plus aisé d'obtenir des assemblages à l'échelle des chromosomes pour des génomes très complexes. L'outil Hifiasm[100] a été développé pour assembler les lectures HiFi et est particulièrement performant pour la séparation des haplotypes.

De son côté, Oxford Nanopore travaille sur la chimie Q20+ (association d'une nouvelle enzyme avec les nouveaux pores R10.4.1). Elle permet d'accroître la qualité des lectures dites simplex (un seul brin séquencé) en abaissant le taux d'erreur à environ 1%. Ces nouvelles conditions favorisent aussi le séquençage en duplex (les deux brins d'un même fragment d'ADN sont séquencés consécutivement). Il est alors possible de générer une séquence consensus à partir des simplex pour obtenir une lecture presque parfaite avec un taux d'erreur proche de 0.1%. Le pourcentage de lectures en duplex est un facteur d'amélioration important pour la technologie Oxford Nanopore. S'il atteint un niveau suffisant il sera possibe sans avoir à multiplier les runs d'obtenir des assemblages de qualité et de continuité à la hauteur des assemblages obtenus avec les lectures HiFi. Il ne sera alors plus nécessaire de corriger les assemblages avec des courtes lectures.

Si on pousse le raisonnement à l'extrême, un développement du protocole Pore-C, qui est actuellement plus difficile à mettre en œuvre que les autres protocoles Hi-C, et des protocoles de préparation de

banques RNAseq (qui nécessite l'obtention d'une quantité d'ARN trop importante pour être possible sur tous les organismes) permettraient de réaliser l'ensemble du séquençage sur les séquenceurs de troisième génération et s'affranchir complètement du séquençage à courtes lectures pour la génération des génomes de référence.

Au cours des années à venir, les méthodes permettant d'obtenir des génomes de référence vont connaître des évolutions. Il est donc à prévoir que les standards de qualité changeront à nouveau.

9.5.5 Perspectives personnelles

Mes perspectives personnelles sont dans un premier temps de mettre en place des méthodes automatisées pour organiser les contigs en scaffolds à l'échelle des chromosomes dans le cadre des grands programmes de séquençage à venir au Genoscope. Il sera essentiel de standardiser les procédures afin de parvenir à générer un grand nombre de génomes de référence. L'automatisation permettra de dégager du temps pour expertiser la qualité et la continuité des séquences assemblées produites. Il est assez évident que ces procédures seront amenées à évoluer au cours de la réalisation des projets. De même, j'organise les réunions de coordination des équipes impliquées dans le programme ATLASea pour la rédaction du cahier des charges, les interactions avec les bases de données du Genoscope et des bases de données publiques, le stockage des métadonnées ou la gestion des flux de production.

Dans un second temps, je pourrais développer des thématiques de recherche pour valoriser certains des génomes de référence produits, en m'orientant en particulier vers des études de génomique comparative. L'analyse et la comparaison de la structure des génomes, de la composition en éléments transposables, de la structure des télomères entre différentes espèces marines pourra être menée. Cette perspective sera affinée en fonction des espèces séquencées.

Enfin, même si je n'ai pas développé cette thématique dans ce manuscrit, je vais travailler sur la mise en place du séquençage de métabarcodes grâce à la technologie Oxford Nanopore. Le séquençage de gènes ribosomaux entiers devrait permettre d'être plus

résolutif pour les assignations taxonomiques d'échantillons environnementaux. La seule limitation reste la complétion des bases de données.

9.6 CONCLUSION

Les évolutions des technologies de séquençage au cours de ces cinquante dernières années ont permis d'accéder à la composition et à la structure d'un grand nombre de génomes. Nous sommes en passe d'obtenir des génomes quasiment parfaits à moindre coût. Les nouveaux standards nous obligent à rester informés, à tester et à développer des approches innovantes pour accéder à un haut niveau d'exigence. Les génomes obtenus pourront alors servir de point de départ pour diverses études: génétiques, médicales, phylogénétiques etc. L'accès à la structure complète des génomes est, par ailleurs, essentiel pour étudier la régulation de la transcription, l'évolution des gènes entre les espèces, l'émergence de nouveaux gènes de résistance ou la variabilité génétique en réponse aux changements climatiques[135]. Il est souvent essentiel d'accumuler des génomes au sein d'une même espèce pour capter toute la diversité allélique des populations. La menace sur la biodiversité a déclenché la mise en place de grands projets internationaux visant à obtenir les génomes de référence d'un très grand nombre d'espèces le long de l'arbre du vivant. La génomique et la bioinformatique associée ont donc encore beaucoup de défis à relever pour répondre aux besoins des scientifiques au service de la biodiversité.

J'ai ainsi acquis une expertise dans le domaine de la reconstruction des chromosomes à partir d'assemblages de longues lectures. Prendre part à la mise en place des technologies en collaborant avec l'équipe de recherche et développement du laboratoire de séquençage m'a permis de comprendre toutes les limites de ces techniques depuis la réalisation des expériences jusqu'à l'utilisation des données elles-mêmes. J'ai ainsi pu développer un rôle de conseil auprès de nos collaborateurs pour la réalisation de leur projet.

Ces dernières années, j'ai participé à la génération de la séquence de

nombreux génomes de référence et à leur valorisation par l'écriture d'articles scientifiques et par des présentations lors de séminaires. Ces valorisations sont l'objet d'études sur la structure de ces génomes et de leur composition.

Les prochaines années seront consacrées aux programmes sur la biodiversité, sujet absolument passionnant, qui me demanderont de maintenir ce niveau d'expertise en continuant à développer de nouvelles stratégies adaptées à chaque type d'organisme.

9.7 BIBLIOGRAPHIE

1 Della-Negra, O. *et al.* Transformation of the recalcitrant pesticide chlordecone by Desulfovibrio sp.86 with a switch from ring-opening dechlorination to reductive sulfidation activity. *Sci Rep* **10**, 13545, doi:10.1038/s41598-020-70124-9 (2020).

2 Morgan, T. H. Heredity and Sex. *Columbia University Press* (1913).

3 Sturtevant, A. H. *Contributions to the genetics of Drosophila simulans and Drosophila melanogaster*. (1929).

4 Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**, 137-158, doi:10.1084/jem.79.2.137 (1944).

5 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738, doi:10.1038/171737a0 (1953).

6 Briggs, R. & King, T. J. Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. *Proc Natl Acad Sci U S A* **38**, 455-463, doi:10.1073/pnas.38.5.455 (1952).

7 Gurdon, J. B. The transplantation of nuclei between two species of Xenopus. *Developmental Biology* **5**, 68-83, doi:10.1016/0012-1606(62)90004-0 (1962).

8 Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-356, doi:10.1016/s0022-2836(61)80072-7 (1961).

9 Nirenberg, M. W. The genetic code. II. *Sci Am* **208**, 80-94, doi:10.1038/scientificamerican0363-80 (1963).

10 Khorana, H. G. *et al.* Polynucleotide synthesis and the genetic code.

Cold Spring Harb Symp Quant Biol **31**, 39-49, doi:10.1101/sqb.1966.031.01.010 (1966).
11 Wahba, A. J. *et al.* Synthetic polynucleotides and the amino acid code. VI. *Proc Natl Acad Sci U S A* **48**, 1683-1686, doi:10.1073/pnas.48.9.1683 (1962).
12 Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695, doi:10.1038/265687a0 (1977).
13 Sanger, F. & Thompson, E. O. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J* **53**, 353-366, doi:10.1042/bj0530353 (1953).
14 Sanger, F. & Thompson, E. O. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem J* **53**, 366-374, doi:10.1042/bj0530366 (1953).
15 Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147**, 1462-1465, doi:10.1126/science.147.3664.1462 (1965).
16 Bkownlee, G. G., Sanger, F. & Barrell, B. G. The sequence of 5 s ribosomal ribonucleic acid. *Journal of Molecular Biology* **34**, 379-412, doi:10.1016/0022-2836(68)90168-x (1968).
17 Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**, 82-88, doi:10.1038/237082a0 (1972).
18 Gilbert, W. & Maxam, A. The nucleotide sequence of the lac operator. *Proc Natl Acad Sci U S A* **70**, 3581-3584, doi:10.1073/pnas.70.12.3581 (1973).
19 Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500-507, doi:10.1038/260500a0 (1976).
20 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467, doi:10.1073/pnas.74.12.5463 (1977).
21 Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560-564, doi:10.1073/pnas.74.2.560 (1977).
22 Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res* **15**, 1767-1776, doi:10.1101/gr.3770505 (2005).
23 *Applied Biosystems 3730/3730xl*, <https://assets.thermofisher.com/TFS-Assets/LSG/manuals/cms_041259.pdf>
24 Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**, 2601-2610, doi:10.1093/nar/6.7.2601 (1979).
25 Gardner, R. C. *et al.* The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun

sequencing. *Nucleic Acids Res* **9**, 2871-2888, doi:10.1093/nar/9.12.2871 (1981).
26 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
27 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).
28 Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601-607, doi:10.1038/nature01348 (2003).
29 Shampo, M. A. & Kyle, R. A. J. Craig Venter--The Human Genome Project. *Mayo Clin Proc* **86**, e26-27, doi:10.4065/mcp.2011.0160 (2011).
30 Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254, doi:10.1371/journal.pbio.0050254 (2007).
31 Venter, J. C., Smith, H. O. & Hood, L. A new strategy for genome sequencing. *Nature* **381**, 364-366, doi:10.1038/381364a0 (1996).
32 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53, doi:10.1126/science.abj6987 (2022).
33 Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-141, doi:10.1016/j.tig.2007.12.007 (2008).
34 *History of sequencing by synthesis*, <https://www.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html>
35 *Introduction to NGS*, <www.illumina.com/technology/next-generation-sequencing.html>
36 *DNA-Sequencing-Costs-Data*.
37 Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138, doi:10.1126/science.1162986 (2009).
38 *HIFI SEQUENCING*, <https://www.pacb.com/technology/hifi-sequencing/>
39 Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155-1162, doi:10.1038/s41587-019-0217-9 (2019).
40 *Pacific Biosciences Terminology*, <http://files.pacb.com/software/smrtanalysis/2.2.0/doc/smrtportal/help/!SSL!/Webhelp/Portal_PacBio_Glossary.htm>
41 *Q20+ chemistry*, <https://nanoporetech.com/q20plus-chemistry>
42 *Nanopore products*, <https://nanoporetech.com/products>

43 Xu, Y., Luo, H., Wang, Z., Lam, H. M. & Huang, C. Oxford Nanopore Technology: revolutionizing genomics research in plants. *Trends Plant Sci* **27**, 510-511, doi:10.1016/j.tplants.2021.11.004 (2022).
44 El-Metwally, S., Hamza, T., Zakaria, M. & Helmy, M. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol* **9**, e1003345, doi:10.1371/journal.pcbi.1003345 (2013).
45 Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr Protoc Bioinformatics* **70**, e102, doi:10.1002/cpbi.102 (2020).
46 Bruijn, N. G. d. « A Combinatorial Problem ». *Koninklijke Nederlandse Akademie v. Wetenschappen* **49**, 758–764 (1946).
47 De bruijn graph. doi:https://towardsdatascience.com/genome-assembly-using-de-bruijn-graphs-69570efcc270.
48 Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**, 36-46, doi:10.1038/nrg3117 (2011).
49 *Mate Pair Sequencing*, <https://emea.illumina.com/science/technology/next-generation-sequencing/mate-pair-sequencing.html>
50 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).
51 Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte* **2021**, 1-9, doi:10.46471/gigabyte.15 (2021).
52 Chen, Y. *et al.* Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun* **12**, 60, doi:10.1038/s41467-020-20236-7 (2021).
53 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540-546, doi:10.1038/s41587-019-0072-8 (2019).
54 Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, 155-158, doi:10.1038/s41592-019-0669-3 (2020).
55 Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737-746, doi:10.1101/gr.214270.116 (2017).
56 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).
57 Aury, J. M. & Istace, B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genom Bioinform* **3**, lqab034,

doi:10.1093/nargab/lqab034 (2021).
58 Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J* **18**, 9-19, doi:10.1016/j.csbj.2019.11.002 (2020).
59 *Platform technology*, <https://bionanogenomics.com/technology/platform-technology/>
60 Liu, N. *et al.* Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C. *Epigenetics Chromatin* **14**, 41, doi:10.1186/s13072-021-00417-4 (2021).
61 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).
62 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
63 *What's Involved In Running A Dovetail Hi-C Assay?*, <https://dovetailgenomics.com/2021/02/01/blog_running-a-dovetail-assay-part-1/>
64 Hansen, P. *et al.* Computational Processing and Quality Control of Hi-C, Capture Hi-C and Capture-C Data. *Genes (Basel)* **10**, doi:10.3390/genes10070548 (2019).
65 Sur, A., Noble, W. S. & Myler, P. J. A benchmark of Hi-C scaffolders using reference genomes and de novo assemblies. *BioRXiV*, doi:10.1101/2022.04.20.488415 (2022).
66 *VGP project phase I*, <https://vertebrategenomesproject.org/phase-one>
67 Pore-C protocole. https://nanoporetech.com/resource-centre/pore-c-using-nanopore-reads-delineate-long-range-interactions-between-genomic-0
68 Jeck, W. R. *et al.* A Nanopore Sequencing-Based Assay for Rapid Detection of Gene Fusions. *J Mol Diagn* **21**, 58-69, doi:10.1016/j.jmoldx.2018.08.003 (2019).
69 Katsman, E. *et al.* Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. *Genome Biol* **23**, 158, doi:10.1186/s13059-022-02710-1 (2022).
70 Fortin, J. P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* **16**, 180, doi:10.1186/s13059-015-0741-y (2015).
71 Li, X. *et al.* Genomic analyses of wild argali, domestic sheep, and their hybrids provide insights into chromosome evolution, phenotypic variation, and germplasm innovation. *Genome Res*,

doi:10.1101/gr.276769.122 (2022).
72 Edwards, D., Batley, J. & Snowdon, R. J. Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* **126**, 1-11, doi:10.1007/s00122-012-1964-x (2013).
73 Zhang, W. *et al.* Investigation of the Genetic Diversity and Quantitative Trait Loci Accounting for Important Agronomic and Seed Quality Traits in Brassica carinata. *Front Plant Sci* **8**, 615, doi:10.3389/fpls.2017.00615 (2017).
74 Jackson, S. A., Iwata, A., Lee, S. H., Schmutz, J. & Shoemaker, R. Sequencing crop genomes: approaches and applications. *New Phytol* **191**, 915-925, doi:10.1111/j.1469-8137.2011.03804.x (2011).
75 Jiao, Y. *et al.* Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524-527, doi:10.1038/nature22971 (2017).
76 Macas, J. *et al.* In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae. *PLoS One* **10**, e0143424, doi:10.1371/journal.pone.0143424 (2015).
77 Zhao, M. *et al.* Shifts in the evolutionary rate and intensity of purifying selection between two Brassica genomes revealed by analyses of orthologous transposons and relics of a whole genome triplication. *Plant J* **76**, 211-222, doi:10.1111/tpj.12291 (2013).
78 Andersen, E. J. *et al.* Wheat Disease Resistance Genes and Their Diversification Through Integrated Domain Fusions. *Front Genet* **11**, 898, doi:10.3389/fgene.2020.00898 (2020).
79 Kim, S. *et al.* New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol* **18**, 210, doi:10.1186/s13059-017-1341-9 (2017).
80 Formenti, G. *et al.* The era of reference genomes in conservation genomics. *Trends Ecol Evol* **37**, 197-202, doi:10.1016/j.tree.2021.11.008 (2022).
81 *Vertebrate Genomes Project special number*, <https://www.nature.com/collections/cabiagjdfj> (2021).
82 A reference standard for genome biology. *Nat Biotechnol* **36**, 1121, doi:10.1038/nbt.4318 (2018).
83 Jones, A. *et al.* High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. *PLoS One* **16**, e0253830, doi:10.1371/journal.pone.0253830 (2021).
84 Russo, A. *et al.* Low-Input High-Molecular-Weight DNA Extraction for Long-Read Sequencing From Plants of Diverse Families. *Front Plant Sci* **13**, 883897, doi:10.3389/fpls.2022.883897 (2022).
85 Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M. & Olmo, E.

Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenet Genome Res* **147**, 217-239, doi:10.1159/000444429 (2015).
86 Lertzman-Lepofsky, G., Mooers, A. O. & Greenberg, D. A. Ecological constraints associated with genome size across salamander lineages. *Proc Biol Sci* **286**, 20191780, doi:10.1098/rspb.2019.1780 (2019).
87 GM., C. *The Cell: A Molecular Approach 2nd edition. Sunderland (MA): Sinauer Associates; 2000. The Complexity of Eukaryotic Genomes.*, <https://www.ncbi.nlm.nih.gov/books/NBK9846/>
88 Lallemand, T., Leduc, M., Landes, C., Rizzon, C. & Lerat, E. An Overview of Duplicated Gene Detection Methods: Why the Duplication Mechanism Has to Be Accounted for in Their Choice. *Genes (Basel)* **11**, doi:10.3390/genes11091046 (2020).
89 Zhu, Z., Tan, S., Zhang, Y. & Zhang, Y. E. LINE-1-like retrotransposons contribute to RNA-based gene duplication in dicots. *Sci Rep* **6**, 24755, doi:10.1038/srep24755 (2016).
90 Kong, H. *et al.* Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. *Plant J* **50**, 873-885, doi:10.1111/j.1365-313X.2007.03097.x (2007).
91 Seo, E., Kim, S., Yeom, S. I. & Choi, D. Genome-Wide Comparative Analyses Reveal the Dynamic Evolution of Nucleotide-Binding Leucine-Rich Repeat Gene Family among Solanaceae Plants. *Front Plant Sci* **7**, 1205, doi:10.3389/fpls.2016.01205 (2016).
92 Hayashi, K. & Yoshida, H. Refunctionalization of the ancient rice blast disease resistance gene Pit by the recruitment of a retrotransposon as a promoter. *Plant J* **57**, 413-425, doi:10.1111/j.1365-313X.2008.03694.x (2009).
93 Janko, K. *et al.* Genome Fractionation and Loss of Heterozygosity in Hybrids and Polyploids: Mechanisms, Consequences for Selection, and Link to Gene Function. *Mol Biol Evol* **38**, 5255-5274, doi:10.1093/molbev/msab249 (2021).
94 Van de Peer, Y., Ashman, T. L., Soltis, P. S. & Soltis, D. E. Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* **33**, 11-26, doi:10.1093/plcell/koaa015 (2021).
95 Kagale, S. *et al.* Polyploid evolution of the Brassicaceae during the Cenozoic era. *Plant Cell* **26**, 2777-2791, doi:10.1105/tpc.114.126391 (2014).
96 Pucker, B., Irisarri, I., de Vries, J. & Xu, B. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology* **3**, doi:10.1017/qpb.2021.18 (2022).
97 Koren, S. *et al.* De novo assembly of haplotype-resolved genomes

with trio binning. *Nat Biotechnol*, doi:10.1038/nbt.4277 (2018).
98 Campoy, J. A. *et al.* Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol* **21**, 306, doi:10.1186/s13059-020-02235-5 (2020).
99 Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* **5**, 833-845, doi:10.1038/s41477-019-0487-8 (2019).
100 Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170-175, doi:10.1038/s41592-020-01056-5 (2021).
101 Chen, H. *et al.* Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat Commun* **11**, 2494, doi:10.1038/s41467-020-16338-x (2020).
102 Sun, H. *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat Genet* **54**, 342-348, doi:10.1038/s41588-022-01015-0 (2022).
103 Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**, 1103-1110, doi:10.1038/s41592-020-00971-x (2020).
104 Yang, C. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput Struct Biotechnol J* **19**, 6301-6314, doi:10.1016/j.csbj.2021.11.028 (2021).
105 Duncan, A. *et al.* Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. *Microbiome* **10**, 67, doi:10.1186/s40168-022-01254-7 (2022).
106 Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* **2**, 100123, doi:10.1016/j.xgen.2022.100123 (2022).
107 Cheng, F., Wu, J. & Wang, X. Genome triplication drove the diversification of Brassica plants. *Hortic Res* **1**, 14024, doi:10.1038/hortres.2014.24 (2014).
108 Murat, F. *et al.* Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol* **16**, 262, doi:10.1186/s13059-015-0814-y (2015).
109 Lu, K. *et al.* Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. *Nat Commun* **10**, 1154, doi:10.1038/s41467-019-09134-9 (2019).
110 Chalhoub, B. *et al.* Plant genetics. Early allopolyploid evolution in the

post-Neolithic Brassica napus oilseed genome. *Science* **345**, 950-953, doi:10.1126/science.1253435 (2014).

111 N. U. Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization. *Jpn J Bot* **7**, 389–452 (1935).

112 Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97-100, doi:10.1038/nature09916 (2011).

113 Rousseau-Gueutin, M. *et al.* Long-read assembly of the Brassica napus reference genome Darmor-bzh. *Gigascience* **9**, doi:10.1093/gigascience/giaa137 (2020).

114 *Triangle de U*, <https://fr.wikipedia.org/wiki/Triangle_de_U>

115 Manzo-Sánchez, G. *et al.* Genetic Diversity in Bananas and Plantains (Musa spp.). doi:10.5772/59421 (2015).

116 Martin, G. *et al.* Genome ancestry mosaics reveal multiple and cryptic contributors to cultivated banana. *Plant J* **102**, 1008-1025, doi:10.1111/tpj.14683 (2020).

117 Boideau, F. *et al.* Epigenomic and structural events preclude recombination in Brassica napus. *New Phytol* **234**, 545-559, doi:10.1111/nph.18004 (2022).

118 Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. . *bioRxiv* doi:10.1101/2022.06.09.495093 (2022).

119 Wang, Z. *et al.* Musa balbisiana genome reveals subgenome evolution and functional divergence. *Nat Plants* **5**, 810-821, doi:10.1038/s41477-019-0452-6 (2019).

120 D'Hont, A. *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213-217, doi:10.1038/nature11241 (2012).

121 Martin, G. *et al.* Improvement of the banana "Musa acuminata" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* **17**, 243, doi:10.1186/s12864-016-2579-4 (2016).

122 Sabatier, P. *et al.* Evidence of Chlordecone Resurrection by Glyphosate in French West Indies. *Environ Sci Technol* **55**, 2296-2306, doi:10.1021/acs.est.0c05207 (2021).

123 Wang, Z. *et al.* A chromosome-level reference genome of Ensete glaucum gives insight into diversity and chromosomal and repetitive sequence evolution in the Musaceae. *Gigascience* **11**, doi:10.1093/gigascience/giac027 (2022).

124 Zhao, T. *et al.* Whole-genome microsynteny-based phylogeny of angiosperms. *Nat Commun* **12**, 3498, doi:10.1038/s41467-021-23665-0 (2021).

125 C, B. *Generating gapless, telomere-to-telomere plant genome*

assemblies, <https://nanoporetech.com/resource-centre/video/webinar-gapless-telomere-to-telomere-plant-genome-assemblies?utm_campaign=Platform&utm_content=220093072&utm_medium=social&utm_source=twitter&hss_channel=tw-37732219> (2021).

126 Mao, Y. & Zhang, G. A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics. *Nat Methods* **19**, 635-638, doi:10.1038/s41592-022-01512-4 (2022).

127 Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biol* **20**, 159, doi:10.1186/s13059-019-1774-4 (2019).

128 Wu, J. *et al.* A Chromosome Level Genome Assembly of a Winter Turnip Rape (Brassica rapa L.) to Explore the Genetic Basis of Cold Tolerance. *Front Plant Sci* **13**, 936958, doi:10.3389/fpls.2022.936958 (2022).

129 Cai, X. *et al.* Impacts of allopolyploidization and structural variation on intraspecific diversification in Brassica rapa. *Genome Biol* **22**, 166, doi:10.1186/s13059-021-02383-2 (2021).

130 Gordon, S. P. *et al.* Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat Commun* **8**, 2184, doi:10.1038/s41467-017-02292-8 (2017).

131 Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet* **51**, 1044-1051, doi:10.1038/s41588-019-0410-2 (2019).

132 Quesneville, H. Twenty years of transposable element analysis in the Arabidopsis thaliana genome. *Mob DNA* **11**, 28, doi:10.1186/s13100-020-00223-x (2020).

133 Lisch, D. How important are transposons for plant evolution? *Nat Rev Genet* **14**, 49-61, doi:10.1038/nrg3374 (2013).

134 Noel, B. *et al.* Pervasive gene duplications as a major evolutionary driver of coral biology. *bioRxiv* doi:10.1101/2022.05.17.492263 (2022).

135 Chen, Y. *et al.* Large-scale genome-wide study reveals climate adaptive variability in a cosmopolitan pest. *Nat Commun* **12**, 7206, doi:10.1038/s41467-021-27510-2 (2021).

10 ANNEXES :

10.1 ANNEXE 1 : LISTE DE MES PUBLICATIONS

ORCID : 0000-0002-8108-9910

Publications en premier et second auteur :

***** contributed equally to the work

2022 :

Belser C*, Poulain J*, ... & Wincker P. Integrative omics framework for characterization of coral reef ecosystems from the Tara Pacific expedition https://doi.org/10.48550/arXiv.2207.02475 - en cours de publication dans *Scientific data*.

2021 :

Belser, C*, Baurens, FC*, Noel, B, ... & Aury JM. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Commun Biol* 4, 1047 (2021). https://doi.org/10.1038/s42003-021-02559-3

2020 :

Rousseau-Gueutin M*, Belser C*, Da Silva C*, Richard G, Istace B, Cruaud C , Falentin C, Boideau F, Boutte J, Delourme R, Deniot G, Engelen S, Ferreira de Carvalho J, Lemainque A, Maillet L, Morice J, Wincker P, Denoeud F, Chèvre AM, Aury JM. Long-read assembly of the Brassica napus reference genome Darmor-bzh, *GigaScience*, Volume 9, Issue 12, December 2020, giaa137, https://doi.org/10.1093/gigascience/giaa137

Istace B, Belser C, Aury JM. Improving Bionano scaffolding with BiSCoT. *PeerJ*, DOI 10.7717/peerj.10150 (2020)

2018 :

Belser C*, Istace B*, **Denis E*** , ... & Aury JM.. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nature Plants 4, 879–887 (2018). https://doi.org/10.1038/s41477-018-0289-4

2014 :

Alberti A*, Belser C*, Engelen S, ... & Wincker P. Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics*. 2014;15(1):912. Published 2014 Oct 20. doi:10.1186/1471-2164-15-912

Autres publications :

2022 :

Guérin N, Ciccarella M, Flamant E, Frémont P, Mangenot S, Istace B, Noel B, **Belser C**, Bertrand L, Labadie K, Cruaud C, Romac S, Bachy C, Gachenot M, Pelletier E, Alberti A, Jaillon O, Wincker P, Aury JM, Carradec Q. Genomic adaptation of the picoeukaryote Pelagomonas calceolata to iron-poor oceans revealed by a chromosome-scale genome sequence. *Commun Biol*. 2022 Sep 16;5(1):983. doi: 10.1038/s42003-022-03939-z.

Eleftheriou E, Aury J-M, Vacherie B, Istace B, **Belser C**, Noel B, Moret Y, Rigaud T, Berro F, Gasparian S, Labadie-Bretheau K, Lefebvre T, Madoui M-A. Chromosome-scale assembly of the yellow mealworm genome [version 3; peer review: 2 approved]. *Open Res Europe* 2022, 1:94 (https://doi.org/10.12688/openreseurope.13987.3)

Boideau F, Richard G, Coriton O, Huteau V, **Belser C,** Deniot G, Eber F, Falentin C, Ferreira de Carvalh, J, Gilet M, Lodé-Taburel M, Maillet L, Morice J, Trotoux G, Aury J.-M, Chèvre A.-M and Rousseau-Gueutin M. (2022), Epigenomic and structural events preclude recombination in *Brassica napus*. New Phytol, 234: 545-559. https://doi.org/10.1111/nph.18004

Canaguier A, Guilbaud R, Denis E, Magdelenat G, **Belser C**, Istace B, ... & Barbe V. (2022). Oxford Nanopore and Bionano Genomics technologies evaluation for plant structural variation detection. *BMC genomics*, *23*(1), 1-17.. https://doi.org/10.1186/s12864-022-08499-4

Aury JM, Engelen S, Istace B, Monat C, Lasserre-Zuber P, **Belser C**, Cruaud C, Rimbert H, Leroy P, Arribat S, Dufau I, Bellec A, Grimbichler D, Papon N, Paux E, Ranoux M, Alberti A, Wincker P, Choulet F, Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding, *GigaScience*, Volume 11, 2022, giac034, https://doi.org/10.1093/gigascience/giac034

Lang-Yona N, Flores J. M, Haviv R, Alberti A, Poulain J, **Belser C**, ... & Vardi A. (2022). Terrestrial and marine influence on atmospheric bacterial diversity over the north Atlantic and Pacific Oceans. *Communications Earth & Environment*, *3*(1), 1-10. https://doi.org/10.1038/s43247-022-00441-6

2021 :

Brandt M I, Pradillon F, Trouche B, Henry N, Liautard-Haag C, Cambon-Bonavita M. A, ..., **Belser C**, ... & Zeppilli D (2021). Evaluating sediment and water sampling methods for the estimation of deep-sea biodiversity using environmental DNA. *Scientific Reports*, *11*(1), 1-14. https://doi.org/10.1038/s41598-021-86396-8

Groppi A, Liu S, Cornille A, Decroocq S, Bui Q. T., Tricon D, ..., **Belser C**, ... & Decroocq V (2021). Population genomics of apricots unravels domestication history and adaptive events. *Nature Communications*, *12*(1), 1-16. https://doi.org/10.1038/s41467-021-24283-6

Mahé F, Henry N, Berney C, Poulain J, Romac S, Ruscheweyh HJ, Salazar G, **Belser C**, Clayssen Q, Hume B.C.C, Boissin E, Galand PE, Pesant S, Lombard F, Armstrong E, Lang Yona N, Klinges G, McMinds R, Vega Thurber R, ... & de Vargas C. (2021). Tara Pacific V9 18S rDNA metabarcoding dataset [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.5041070

Trouche B, Brandt MI, **Belser C**, ... & Maignien L. Diversity and Biogeography of Bathyal and Abyssal Seafloor Bacteria and Archaea Along a Mediterranean-Atlantic Gradient. *Frontiers in Microbiology*. 2021 ;12:702016. DOI: 10.3389/fmicb.2021.702016. PMID: 34790173; PMCID: PMC8591283.

2020 :

Hume B.C.C., Poulain J, Pesant S, **Belser C**, Ruscheweyh HJ, Moulin C, ... & Voolstra C R. (2020). Tara Pacific metabarcoding sequencing (16S, 18S, ITS2) reference & replication tables version 1 (Version 1) [Data set]. *Zenodo*. http://doi.org/10.5281/zenodo.4073035

Ruscheweyh HJ, Salazar G, Poulain J, **Belser C**, Clayssen Q, Hume B.C.C., ... & Sunagawa S. (2020). Tara Pacific 16S rRNA data analysis release (Version 1.0.0) [Data set]. *Zenodo*. http://doi.org/10.5281/zenodo.4073269

Hume B.C.C., Poulain J, Pesant S, **Belser C**, Ruscheweyh HJ, Forcioli D, ... & Voolstra C R. (2020). Tara Pacific ITS2 Symbiodiniaceae data release version 1 (Version 1) [Data set]. *Zenodo*. http://doi.org/10.5281/zenodo.4061797

Hume B.C.C., Poulain J, Pesant S, **Belser C**, Ruscheweyh HJ, Boissin E, Armstrong E, Clayssen Q, Henry N, Klinges G, McMinds R, Paoli L, Pogoreutz C, Salazar G, Ziegler M, Moulin C, Bourdin G, Iwankow G, Romac S, … & Voolstra C R. (2020). Tara Pacific 18S-based coral host genetic analysis data release version 1 (Version 1) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.4265266

Brandt M, Pradillon F, Trouche B, Henry N, Cambon Bonavita MA, Cueff-Gauchard V, Wincker P, **Belser C**, Poulain J, Arnaud-Haond S, Zeppilli D (2020). ABYSS sampling comparisons. IFREMER. https://doi.org/10.12770/2deb785a-74c5-4b9d-84d6-82a81e0dda6d

Boutte, J, Maillet, L, Chaussepied, T, Letort, S, Aury, JM, **Belser, C**, Boideau, F, Brunet, A, Coriton, O, Deniot, G, Falentin, C, Huteau, V, Lodé, M, Morice, J, Trotoux, G, Chèvre, AM, Rousseau-Gueutin, M, Ferreira de Carvalho, J. Genome Size Variation and Comparative Genomics Reveal Intraspecific Diversity in Brassica rapa. *Frontiers in Plant Science* 2020; https://doi.org/10.3389/fpls.2020.577536

Leroy, T, Rougemont, Q, Dupouey, J.-L, Bodénès, C, Lalanne, C, **Belser, C**, Labadie, K, Le Provost, G, Aury, J.-M, Kremer, A and Plomion, C. (2020), Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. *New Phytol*, 226: 1183-1197. doi:10.1111/nph.16039

Polonais V, Niehus S, Wawrzyniak I, Franchet A, Gaspin C, Belkorchia A, Reichstadt M, **Belser C**, Labadie K, Couloux A, Delbac F, Peyretaillade E, Ferrandon D. 2019. Draft genome sequence of Tubulinosema ratisbonensis, a microsporidian species infecting the model organism Drosophila melanogaster. *Microbiol Resour Announc* 8:e00077-19. https://doi.org/10.1128/MRA.00077-19

2019 :

Gorsky G, Bourdin G, Lombard F, Pedrotti M. L, Audrain S, Bin N., … , **Belser C**, … & Karsenti E. (2019). Expanding Tara oceans protocols for underway, ecosystemic sampling of the ocean-atmosphere interface during Tara Pacific expedition (2016–2018). *Frontiers in Marine Science*, 750. https://doi.org/10.3389/fmars.2019.00750

Kreplak J, Madoui M. A, Cápal P, Novák P, Labadie K, Aubert G, Bayer P. E, Gali K. K, Syme R. A, Main D, Klein A, Bérard A, Vrbová I, Fournier C, d'Agata L, **Belser C**, Berrabah W, Toegelová H, Milec Z, Vrána J, … Burstin J. (2019). A reference genome for pea provides insight into legume genome evolution.

Nature genetics, 51(9), 1411–1422. https://doi.org/10.1038/s41588-019-0480-1

Planes S, Allemand D, Agostini S, Banaigs B, Boissin E, Boss E, ... , **Belser C** , ... & Tara Pacific Consortium. (2019). The Tara Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. *PLoS biology*, *17*(9), e3000483. https://doi.org/10.1371/journal.pbio.3000483

2018 :

Plomion C, Aury J. M, Amselem J, Leroy T, Murat F, Duplessis S, ..., **Belser C**, ... & Salse J. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, *4*(7), 440-452. https://doi.org/10.1038/s41477-018-0172-3

2017 :

Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, ..., **Belser C**, ... & Wincker P. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific data*, *4*(1), 1-20. https://doi.org/10.1038/sdata.2017.93

Gouin A, Bretaudeau A, Nam K, Gimenez S, Aury J. M, Duvic B, ..., **Belser C,** ... & Fournier P. (2017). Two genomes of highly polyphagous lepidopteran pests (Spodoptera frugiperda, Noctuidae) with different host-plant ranges. *Scientific reports*, *7*(1), 1-12. doi:10.1038/s41598-017-10461-4

2016 :

Plomion C, Aury J. M, Amselem J, Alaeitabar T, Barbe V, **Belser C**, ... & Kremer A. (2016). Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Molecular ecology resources*, *16*(1), 254-265. doi:10.1111/1755-0998.12425

2015 :

Alberti A, Briñas L, Orvain C, **Belser C**, Cruaud C, Labadie K, Bertrand L, Barbe V, Aury JM, Wincker P. BAC ends library generation for Illumina sequencing. Published in *Protocol Exchange* on June 02, 2015. doi.org/10.1038/PROTEX.2015.048

Lesur I, Le Provost G, Bento P, ..., **Belser C,** ... & Plomion C. The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery

of genes regulated during bud dormancy release. *BMC Genomics* 16, 112 (2015). https://doi.org/10.1186/s12864-015-1331-9

Madoui M. A, Engelen S, Cruaud C, **Belser C**, Bertrand L, Alberti A, ... & Aury J. M. (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC genomics*, *16*(1), 1-11.https://doi.org/10.1186/s12864-015-1519-z

2003 :

Heilig R, Eckenberg R, Petit J. L, Fonknechten N, Da Silva C, Cattolico L, ..., **Belser C**, ... & Weissenbach J. (2003). The DNA sequence and analysis of human chromosome 14. *Nature*, *421*(6923), 601-607. https://doi.org/10.1038/nature01348

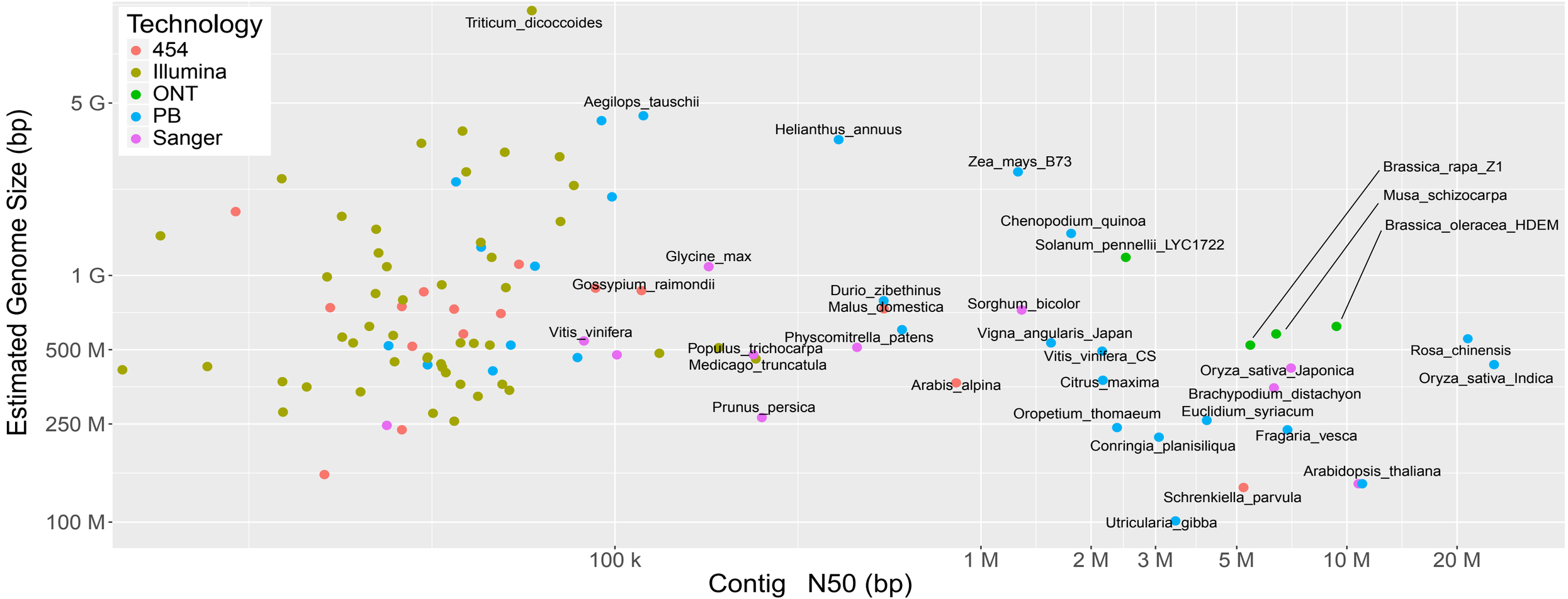
2002 :

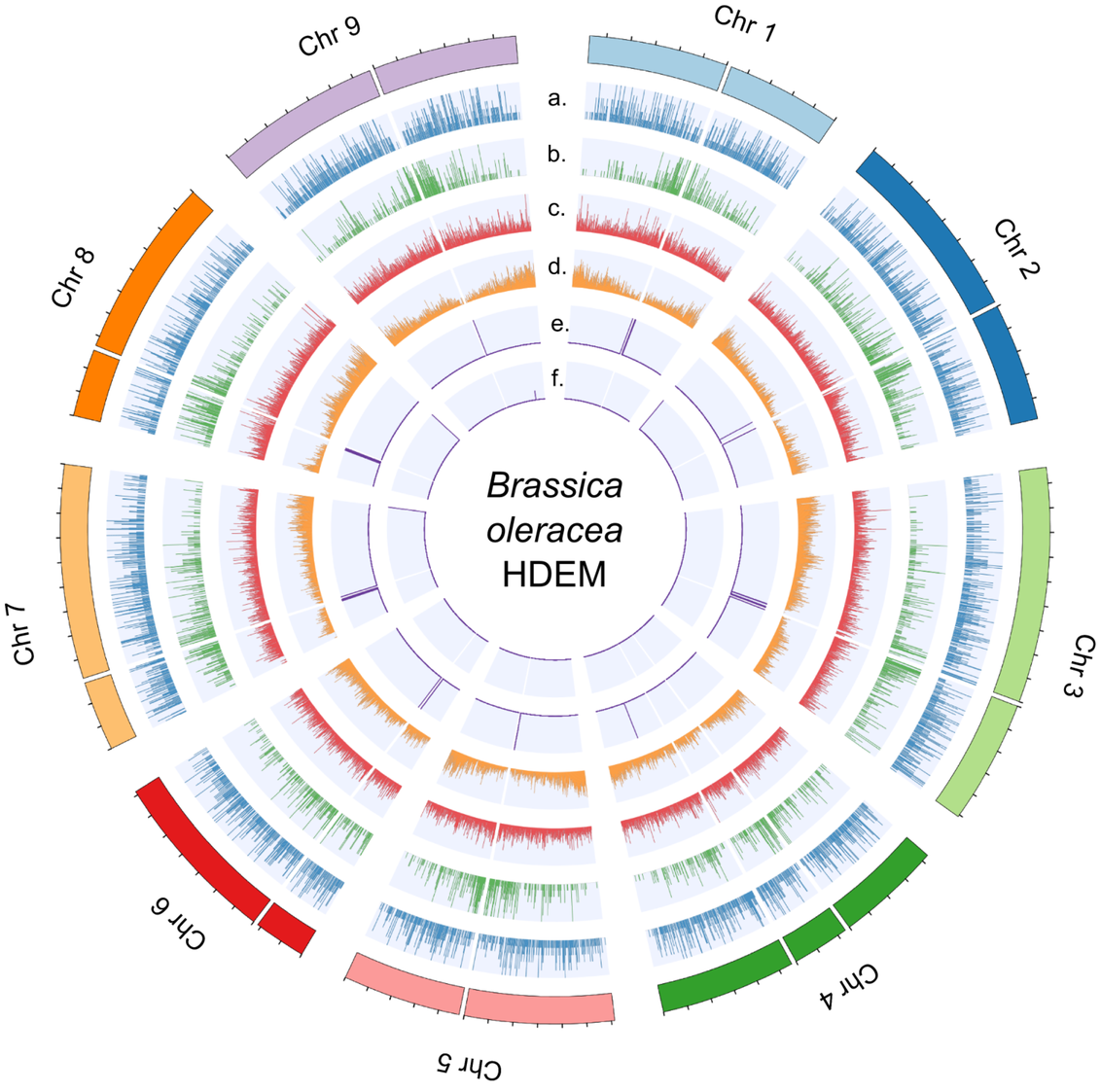
Lefebvre S, Burlet P, Viollet L, Bertrandy S, Huber C, **Belser C**, Munnich A. (2002). A novel association of the SMN protein with two major non-ribosomal nucleolar proteins and its implication in spinal muscular atrophy. *Human Molecular Genetics*, *11*(9), 1017-1027. doi:10.1093/hmg/11.9.1017

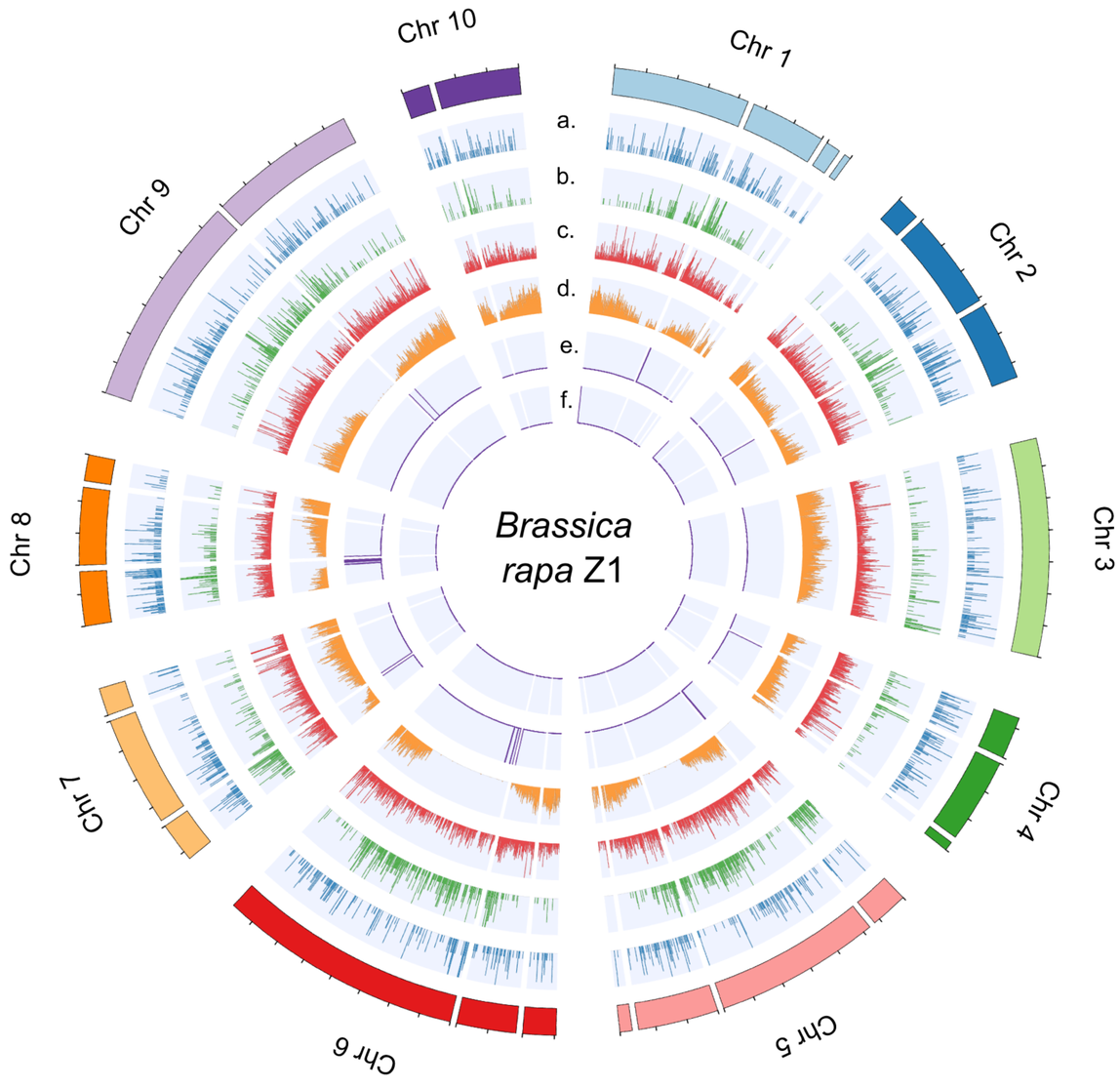
2000 :

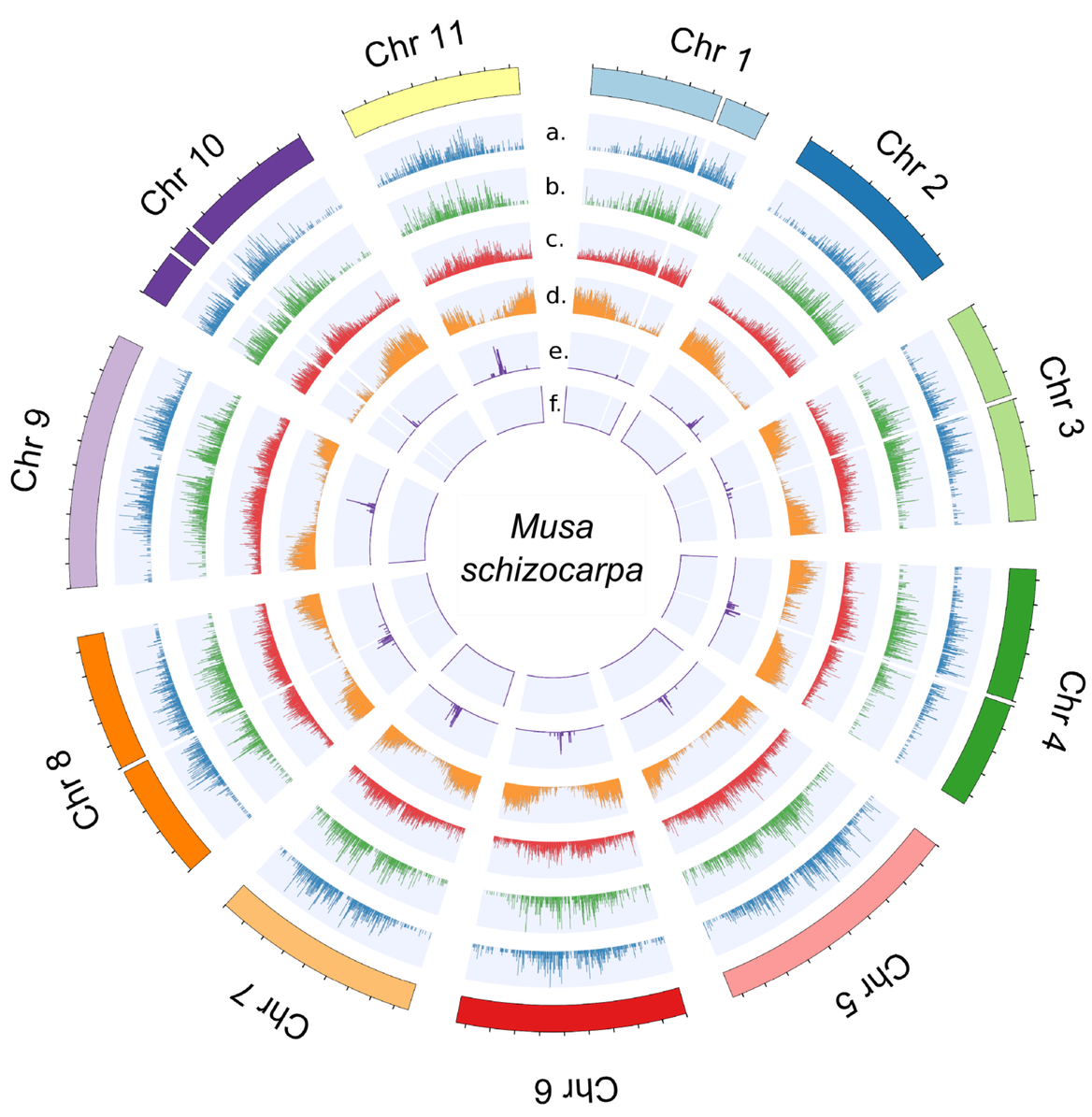
Viollet L, Leclair-Richard D, Burlet P, **Belser C**, Vial E, Bertrandy S, ... & Munnich A. (2000, October). A new variant for autosomal recessive spinal muscular atrophy in childhood. In *The American Journal of Human Genetics* (Vol. 67, No. 4, pp. 113-113).

10.2 Annexe 2 : "Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps" Belser, C., Istace, B., Denis, E. et al. Nature Plants. 2018 Supplementary Data.


Technology
454
Illumina
ONT
PB
Sanger
Triticum_dicoccoides
Aegilops_tauschii
Helianthus_annuus
Zea_mays_B73
Brassica_rapa_Z1
Musa_schizocarpa
Brassica_oleracea_HDEM
Chenopodium_quinoa
Solanum_pennellii_LYC1722
Glycine_max
Gossypium_raimondii
Durio_zibethinus
Malus_domestica
Sorghum_bicolor
Vitis_vinifera
Physcomitrella_patens
Vigna_angularis_Japan
Populus_trichocarpa
Medicago_truncatula
Vitis_vinifera_CS
Rosa_chinensis
Oryza_sativa_Japonica
Oryza_sativa_Indica
Arabis_alpina
Citrus_maxima
Brachypodium_distachyon
Prunus_persica
Oropetium_thomaeum
Euclidium_syriacum
Fragaria_vesca
Conringia_planisiliqua
Arabidopsis_thaliana
Schrenkiella_parvula
Utricularia_gibba
5 G
1 G
500 M
250 M
100 M
100 k
1 M
2 M
3 M
5 M
10 M
20 M
Estimated Genome Size (bp)
Contig N50 (bp)


Chr 9
Chr 1
Chr 8
Chr 2
Chr 7
Chr 3
Chr 6
Chr 4
Chr 5
a.
b.
c.
d.
e.
f.
Brassica
oleracea
HDEM


Chr 10
Chr 1
Chr 9
Chr 2
Chr 8
Chr 3
Chr 7
Chr 4
Chr 6
Chr 5
a.
b.
c.
d.
e.
f.
Brassica
rapa Z1


Chr 11
Chr 1
Chr 10
Chr 2
Chr 9
Chr 3
Chr 8
Chr 4
Chr 7
Chr 5
Chr 6
a.
b.
c.
d.
e.
f.
Musa
schizocarpa


Brassica oleracea
Brassica rapa
Musa sp.
Proportion of estimated genome size occupied (%)
100
90
80
70
60
50
40
30
20
10
0
Reference
Assembly
7 %
19 %
27 %
7 %
10 %
9 %
5 %
10 %
7 %
69 %
11 %
4 %
12 %
8 %
64 %
26 %
37 %
10 %
9 %
4 %
31 %
24 %
7 %
9 %
8 %
5 %
7 %
9 %
70 %
7 %
4 %
9 %
10 %
70 %
6 %
28 %
29 %
8 %
9 %
8 %
17 %
27 %
16 %
6 %
16 %
15 %
68 %
16 %
19 %
64 %
17 %
47 %
16 %
6 %
Other
Ns
Unannotated regions
Repeats
All classes
DNA transposons
LINEs
Copia
Gypsy
Known TEs
Gene structure
Introns
Exons

Supplementary Table 1. Brassica diversity. Positioning of the new assembled genomes among the diversity described for each *B. rapa* and *B. oleracea* species (Cheng et al. 2016)

	Groups as defined by Cheng et al. 2016	Morphotypes	ssp	Reference genomes
B. rapa	1. Turnips	European and Chinese turnips	*rapa*	
	2. Sarson	Sarson	*trilocularis*	Z1
	3. Turnip rapes	Turnip rape	*oleifera*	
	4. Japanese group	Komatsuna	*perviridis*	
		Mizuna	*nipposinica*	
	5. Pak choi	Pak choi	*chinensis*	
		Wutacai	*narinosa*	
		Caixin	*parachinensis*	
		Zicaitai	*chinensis var purpurea* Bailey	
		Taicai	*chinensis var tai-tsai Lin*	
	6. Chinese cabbages	Chinese cabbages	*pekinensis*	Chiifu (Wang et al. 2011)
B. oleracea	1. Kohlrabi	Kohlrabi	*gongylodes*	
	2. Brussel sprouts	Brussel sprouts	*gemmifera*	
	3. Kale	Kale	*acephala*	
		Curly kale	*sabellica*	
	4. Cabbage	Cabbage	*capitata*	
		Pointed cabbage	*capitata*	
		White cabbage	*capitata*	
	5. Broccoli	Broccoli	*italica*	HDEM
	6. Cauliflower	Cauliflower	*botrytis*	
	7. Chinese kale	Chinese kale	*alboglabra*	To1000 (Parkin et al. 2014)

Supplementary Table 2. Statistics of the ONT datasets. Standard metrics were computed using Nanopore reads larger than 1Kb.

	Brassica oleracea	***Brassica rapa***	***Musa sp.***	
	HDEM	Z1	*M. schizocarpa* (MinION)	*M. schizocarpa* (PromethION)
# Reads (>1Kb)	1,242,230	3,763,611	2,554,197	1,240,257
Cumulative size (bp)	20,081,668,921	31,496,053,430	25,852,838,525	17,643,414,337
Average size (bp)	16,165.8	8,368.57	10,121.7	14,225.6
N50 (bp)	32,339	15,042	24,443	26,273
N90 (bp)	7,718	3,570	4,106	6,826
Max size (bp)	607,633	582,421	929,542	272,895
# of reads > 50Kb	77,223	30,453	41,480	38,419
Coverage	31.87X	79X	44X	30X
Coverage (reads >50Kb)	8.23X	5.08X	4.36X	3.99X
Number of flowcells	14	12	18	1

Supplementary Table 3. *Brassica oleracea* HDEM assembly statistics. Three assembly tools were used (with different input datasets) and compared.

	Ra		SMARTDeNovo		Wtdbg
Subset of reads used	All raw reads	All Canu corrected reads	All raw reads	All Canu corrected reads	All Canu corrected reads
Coverage	32x	25x	32x	25x	25x
# contigs	**244**	908	622	425	7,350
Cumulative size	546,379,674	521,156,753	515,008,582	505,564,541	**613,251,567**
N50 (L50)	**7,277,585** (23)	2,385,213 (65)	1,853,586 (80)	3,873,401 (32)	91,340 (2,457)
N90 (L90)	**1,296,241** (85)	447,359 (214)	413,768 (294)	509,533 (164)	54,742 (5,817)
Max size	**25,371,342**	11,765,384	10,639,203	21,968,009	494,298
# of N's	0	0	0	0	0
CPU time (h)	261.4	861.2	363.1	321.6	**56.2**

Supplementary Table 4. *Brassica rapa* Z1 assembly statistics. Three assembly tools were used (with different input datasets) and compared.

	Ra				SMARTDeNovo				Wtdbg
Subset of reads used	All raw reads	Filtlong	Longest	All Canu corrected reads	All raw reads	Filtlong	Longest	All Canu corrected reads	All Canu corrected reads
Coverage	58x	30x	30x	47x	58x	30x	30x	47x	47x
# contigs	**544**	674	842	2,003	1,037	836	835	720	1,652
Cumulative size	**375,275,633**	364,619,539	374,160,667	288,463,898	335,630,604	327,770,952	327,611,620	302,406,720	365,998,479
N50 (L50)	**3,799,257** (25)	1,360,290 (71)	1,207,008 (79)	192,074 (477)	2,273,748 (41)	2,031,767 (45)	2,004,761 (44)	3,467,078 (20)	540,558 (173)
N90 (L90)	**202,877** (265)	195,135 (400)	158,697 (496)	74,570 (1,404)	93,651 (337)	129,313 (289)	128,579 (287)	153,081 (192)	113,199 (657)
Max size	**21,625,070**	5,058,985	9,887,946	1,021,377	8,757,174	7,323,734	8,239,632	15,954,262	2,972,523
# of N's	0	0	0	0	0	0	0	0	0
CPU time (h)	315.7	264.4	256.3	272.1	431.6	313.5	326.8	365.3	**155.3**

Supplementary Table 5. *Musa schizocarpa* assembly statistics. Three assembly tools were used (with different input datasets) and compared.

	Ra				SMARTDeNovo				Wtdbg
Subset of reads used	All raw reads	Filtlong	Longest	All Canu corrected reads	All raw reads	Filtlong	Longest	All Canu corrected reads	All Canu corrected reads
Coverage	45x	30x	30x	41x	45x	30x	30x	41x	41x
# contigs	615	437	455	**433**	801	757	724	507	1,181
Cumulative size	522,028,523	**527,161,307**	515,523,264	506,159,756	496,199,770	495,616,075	486,814,010	497,262,053	267,936,515
N50 (L50)	2,134,507 (59)	**4,036,202** (33)	3,644,647 (35)	1,953,748 (63)	1,561,622	1,520,953 (79)	1,303,135 (83)	2,599,794 (51)	554,240 (132)
N90 (L90)	333,206 (292)	**558,882** (180)	474,353 (179)	326,879 (286)	272,502	297,191 (339)	279,363 (372)	470,097 (229)	103,732 (556)
Max size	12,763,834	15,097,746	**17,454,872**	10,846,287	8,691,051	14,410,166	10,972,794	10,725,756	3,572,747
# of N's	0	0	0	0	0	0	0	0	0
CPU time (h)	345.7	245.6	231.2	746.6	435.4	378.2	371.3	363.6	**143.4**

Supplementary Table 6. Polishing of the *B. oleracea* assembly. Statistics of the best assembly of *B. oleracea* HDEM after each round of polishing.

	Base	Nanopolish	Racon			Pilon		
Iteration	0	1	1	2	3	1	2	3
# contigs	244	244	244	244	244	244	244	244
Cumulative size	546,379,674	547,031,160	546,913,461	547,026,192	547,091,095	547,943,904	547,194,066	547,048,580
N50 (L50)	7,277,585 (23)	7,284,156 (23)	7,283,685 (23)	7,284,263 (23)	7,284,706 (23)	7,298,775 (23)	7,295,653 (23)	7,294,619 (23)
N90 (L90)	1,296,241 (85)	1,298,063 (85)	1,297,835 (85)	1,298,173 (85)	1,298,496 (85)	1,305,142 (85)	1,304,970 (85)	1,304,967 (85)
Max size	25,371,342	25,394,918	25,391,398	25,394,738	25,396,225	25,436,773	25,407,931	25,406,906
BUSCO (complete)	74.3 %	76.3 %	75.0 %	76.0 %	76.0 %	96.3 %	97.4 %	97.3 %
BUSCO (missing)	19.2 %	17.2 %	18.5 %	17.5 %	17.3 %	2.8 %	2.0 %	1.9 %
CPU time (h)	NA	1,533.6	103.7	96.5	92.1	216.3	214.2	217.8

Supplementary Table 7. Polishing of the *B. rapa* assembly. Statistics of the best assembly of *B. rapa* Z1 before and after each round of polishing.

	Base	Racon			Pilon		
Iteration	0	1	2	3	1	2	3
# contigs	544	544	544	544	544	544	544
Cumulative size	375,275,633	375,251,823	375,181,714	375,103,314	373,871,141	373,598,242	373,437,357
N50 (L50)	3,799,257 (25)	3,802,388 (25)	3,802,955 (25)	3,802,820 (25)	3,793,313 (25)	3,793,078 (25)	3,793,063 (25)
N90 (L90)	202,877 (265)	203,333 (265)	203,295 (265)	203,694 (294)	202,177 (264)	202,035 (264)	202,023 (264)
Max size	21,625,070	21,626,948	21,626,315	21,626,623	21,556,942	21,548,434	21,547,698
BUSCO (complete)	79.7 %	81.7 %	82.1 %	82.4 %	96.6 %	97.5 %	97.8 %
BUSCO (missing)	13.0 %	12.2 %	11.6 %	11.0 %	2.0 %	1.6 %	1.4 %

Supplementary Table 8. Polishing of the *M. schizocarpa* assembly. Statistics of the best assembly of *M. schizocarpa* before and after each round of polishing.

	Base	Racon			Pilon		
Iteration	0	1	2	3	1	2	3
# contigs	437	437	437	437	437	437	437
Cumulative size	527,161,307	528,155,060	528,253,158	528,193,136	525,214,663	523,293,606	522,734,513
N50 (L50)	4,036,202 (33)	4,056,238 (33)	4,049,711 (33)	4,053,660 (33)	4,031,517 (33)	4,019,977 (33)	4,019,832 (33)
N90 (L90)	558,882 (180)	560,158 (180)	564,416 (179)	563,775 (179)	557,976 (180)	555,379 (180)	554,125 (180)
Max size	15,097,746	15,119,911	15,124,595	15,127,107	15,005,349	14,927,788	14,918,243
BUSCO (complete)	53.8 %	58.8 %	58.2 %	58.9 %	89.7 %	93.4 %	93.4 %
BUSCO (missing)	40.6 %	35.4 %	37.2 %	36.8 %	8.0 %	5.5 %	5.4 %

Supplementary Table 9. BioNano raw data. Statistics of the data produced with the Saphyr system.

	B. rapa Z1		*B. oleracea* HDEM		*M. schizocarpa*	
	*Bsp*QI	DLE	*Bsp*QI	DLE	*Bsp*QI	DLE
Total number of molecules	3,487,923	12,184,456	6,489,688	9,118,635	8,107,001	19,663,778
Total length (Mbp)	345,728	1,175,790	481,008	480,100	636,930	1,383,815
Molecule N50 (kbp)	176	128	177.9	156.3	120.8	105.27
Label density (/100kb)	9.09	14.29	9.75	13.56	8.90	18.98

Supplementary Table 10. Optical Maps. Statistics of the six genome maps produced in this study.

	B. rapa Z1		*B. oleracea* HDEM		*M. schizocarpa*	
	*Bsp*QI	DLE	*Bsp*QI	DLE	*Bsp*QI	DLE
Number Genome Maps	337	381	445	116	266	197
Total Genome Map Length (Mbp)	474.953	584.226	609.464	639.650	565.271	643.321
Mean Genome Map Length (Mbp)	1.409	1.533	1.370	5.514	2.125	3.266
Median Genome Map Length (Mbp)	0.838	0.464	0.845	0.606	0.671	0.583
Genome Map N50 (Mbp)	2.166	10.705	2.264	32.187	5.099	28.779

Supplementary Table 11. Gaps validation. Statistics of the three assemblies before and after the GapChecker process.

	Brassica oleracea		***Brassica rapa***		***Musa sp.***			
	HDEM		Z1		*M. schizocarpa* (MinIon)		*M. schizocarpa* (PromethIon)	
	Input	Output	Input	Output	Input	Output	Input	Output
Estimated genome size	630 Mb	630 Mb	529 Mb	529 Mb	587 Mb	587 Mb	587 Mb	587 Mb
# scaffolds (>=2Kb)	140	140	335	335	227	227	199	199
Cumulative size	557,055,344	555,091,784	406,461,934	401,924,207	530,631,668	525,589,058	525,348,044	519,519,437
N50 (L50)	29,588,783 (8)	29,526,370 (8)	15,479,745 (8)	15,385,215 (8)	37,464,497 (6)	36,782,414 (6)	37,356,153 (6)	36,870,750 (6)
N90 (L90)	13,953,050 (17)	13,883,595 (17)	1,748,645 (31)	1,671,465 (31)	9,742,099 (15)	9,701,981 (15)	20,101,261 (14)	19,789,157 (14)
Max size	48,464,554	48,264,641	39,998,724	38,870,275	53,362,269	52,773,321	52,779,685	52,129,314
# of N's	1.79%	1.79%	8.12%	8.2%	1.48%	1.48%	1.46%	1.46%
# contigs (>=500bp)	342	269	766	634	558	383	523	332
Cumulative size	547,048,190	545,121,057	373,436,138	368,963,829	522,733,752	517,778,641	517,637,207	511,903,909
N50 (L50)	6,977,344 (23)	9,493,263 (19)	3,603,274 (26)	5,519,976 (17)	3,802,356 (35)	6,493,560 (24)	3,811,725 (36)	9,982,791 (17)
N90 (L90)	1,183,704 (88)	2,202,256 (59)	154,330 (309)	184,937 (212)	527,432 (190)	1,054,407 (84)	528,173 (178)	1,007,446 (67)
Max size	25,406,906	26,715,931	21,547,698	22,127,281	14,918,243	18,138,444	13,768,943	27,050,776

Supplementary Table 12. Final genome assemblies. Statistics of the three assemblies generated using genetic map (*Brassica oleracea*) or comparative genomics compared to current reference genomes (*Brassica rapa* and *Musa schizocarpa*).

	Brassica oleracea		***Brassica rapa***		***Musa sp.***	
	To1000	HDEM	Chiifu	Z1	*Musa acuminata*	*Musa schizocarpa*
Reference	Parkin 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	61,279	41,019	46,721	36,542	32,809
% of anchored genes	91.39%	98.25%	96.56%	98.14%	91.98%	98.38%

Supplementary Table 13. Gene prediction results. Statistics of gene predictions compared to existing gene catalogues.

	Brassica oleracea		***Brassica rapa***		***Musa sp.***	
	To1000	HDEM	Chiifu	Z1	*Musa acuminata*	*Musa schizocarpa*
Reference	Parkin et al.	This study	Cai et al.	This study	Martin et al.	This study
# number of genes	59,225	61,279	41,019	46,721	36,542	32,809
#genes no introns	13,094	16,047	8,867	11,597	4,515	7,038
genes size (avg :med)	1,749.70 : 1,347	1,969.14 : 1,171	2,018.61 : 1,558	2,043.81 : 1,373	3,595.14 : 2,268	4,040.61 : 2,206
#exons/gene (avg :med)	4.54 : 3	4.30 : 3	5.04 : 3	4.72 : 3	5.43 : 4	5.08 : 3
#exons/gene pluri (avg :med)	5.54 : 4	5.47 : 4	6.15 : 5	5.94 : 4	6.05 : 5	6.19 : 5
CDS size (avg :med)	1,042.26 : 837	937.60 : 699	1,173.19 : 981	1,062.68 : 861	1,035.67 : 861	1,126.84 : 939
CDS size pluri (avg :med)	1,146.79 : 1,038	1,061.29 : 969	1,268.80 : 1,170	1,183.09 : 1,110	1,114.41 : 975	1,226.26 : 1,149
#introns	209,371	202,246	165,565	173,667	161,762	133,709
introns size (avg :med)	200.10 : 97	312.55 : 96	209.46 : 95	263.95 : 94	578.18 : 147	714.97 : 166
BUSCO (complete)	95.1%	95.8%	96.3%	96.6%	86.8%	92.3%

Supplementary Table 14. Gene prediction validation. Annotation Edit Distance (AED) of the three gene prediction compared to the reference annotations.

	Brassica oleracea			***Brassica rapa***			***Musa schizocarpa***		
	SN	SP	Accuracy	SN	SP	Accuracy	SN	SP	Accuracy
AED	92.69%	83.37%	88.03%	95.69%	85.86%	90.73%	93.78%	89.56%	91.67%

Supplementary Table 15. **Complete TE copies annotation**. Number and cumulative size occupied by different classes of TEs for *Brassica rapa*, *Brassica oleracea* and *Musa spp*. Only alignments that covered >=90% of the TEs were kept.

	Brassica oleracea		***Brassica rapa***		***Musa sp.***	
	To1000	HDEM	Chiifu	Z1	Musa acuminata	Musa schizocarpa
# Copia	5,789 (20,766,122)	7,986 (31,124,810)	2,950 (8,858,730)	3,391 (11,874,419)	5,141 (27,570,474)	8,069 (44,491,821)
# Gypsy	4,321 (14,682,965)	5,217 (20,459,038)	3,144 (11,483,563)	3,246 (13,567,209)	4,433 (25,824,554)	8,806 (53,755,507)
# known TEs	29,768 (141,052,776)	28,909 (167,305,803)	24,892 (85,474,058)	25,550 (91,053,084)	22,021 (119,972,693)	29,610 (183,436,600)
# centromeric repeats	171 (27,612)	1,406 (237,497)	1,595 (272,835)	8,711 (1,425,573)	115 (603,126)	535 (2,808,232)
# telomeric repeats	3,980 (479,248)	4,812 (635,821)	1,360 (128,237)	1,190 (112,592)	65 (8,515)	57 (519,989)

Supplementary Table 16. **TE elements annotation**. Raw RepeatMasker output, showing number and cumulative size occupied by different classes of TEs for *Brassica rapa*, *Brassica oleracea* and *Musa spp*.

	Brassica oleracea		***Brassica rapa***		***Musa sp.***	
	To1000	HDEM	Chiifu	Z1	Musa acuminata	Musa schizocarpa
# DNA transposons	49221 (17,753,560)	51,496 (29,553,144)	31,308 (9,715,893)	25,361 (9,084,906)	6,490 (1,557,033)	5,929 (1,808,547)
# LINES	16,358 (10,654,646)	15,258 (11,720,018)	9,582 (6,100,845)	7,036 (5,211,332)	2,114 (745,713)	1,763 (724,733)
# Copia	20,206 (22,014,628)	20,748 (37,023,752)	11,731 (9,344,058)	9,422 (13,270,803)	43,716 (28,397,593)	65,470 (53,766,258)
# Gypsy	23,871 (16,365,284)	25,320 (23,476,470)	22,818 (10,801,776)	16,909 (16,234,922)	49,952 (27,196,813)	101,894 (45,836,858)
# known TEs	260,430 (152,921,361)	232,354 (210,596,602)	260,430 (152,921,361)	277,209 (142,708,090)	306,035 (166,541,343)	487,427 (286,509,336)
# centromeric repeats	4,855 (686,342)	38,663 (6,671,809)	32,897 (4,161,745)	63,147 (10,930,312)	2,044 (1,542,559)	4,026 (5,915,074)
# telomeric repeats	3,980 (479,248)	4,812 (635,821)	1,360 (128,237)	1,190 (112,592)	65 (8,515)	57 (519,989)

Supplementary Table 17. **Comparison of PACBIO and ONT datasets.** Standard metrics of PACBIO and ONT datasets used for the comparison of read length and coverage.

	Vigna angularis	*Vitis vinifera*	*Citrus maxima*	*Rosa chinensis*	*Fragaria vesca*	*Arabidopsis thaliana*	*Brassica oleracea*	*Brassica rapa*	*Musa schizocarpa*
	PACBIO						ONT		
# reads	15,537,722	14,711,534	6,093,228	7,886,587	5,217,686	7,806,180	1,582,528	4,656,569	4,143,878
Assembly size (bp)	522,761,097	591,420,921	345,757,338	514,324,162	220,357,314	127,419,454	554,975,960	401,923,810	525,280,193
Cumulative size (Gbp)	67	136	43	115	37	36	20	32	27
Coverage	127	231	125	224	168	283	36	79	51
N50 (bp)	10,921	20,241	11,985	25,271	13,518	13,947	32,095	14,782	23,532
Avg size (bp)	4,287.76	9,275.65	7,085.07	14,584.90	7,108.05	4,624.80	12,787.0	6,854.17	6,445.47
# Reads > 50Kbp	277	59,188	54	134,459	7,133	5,415	77,223	30,453	41,480
Coverage (>50Kbp)	0.02	5.81	0.00	16.20	1.89	2.42	9.34	5.08	4.88
N90 (bp)	2,100	6,775	3,563	10,460	4,416	1,593	7,269	3,298	3,033t

Supplementary Table 18. **Alignment of reads from various morphotypes.** Proportion of aligned illumina reads from 318 accessions of *Brassica* on the Chiifu (Chinese cabbage) and To1000 (Chinese kale) reference genomes and on our Z1 (sarsons) and HDEM (broccoli) genome assemblies

Species	Morphotype	Number of accessions	Average % of aligned reads on reference genome	Average % of aligned reads on our assembly	Difference
Brassica rapa	Yellow Sarson	1	94.7	95.6	0.9
	Caixin	30	97.2367	97.7867	0.55
	Edible Flower	1	95.2	95.8	0.6
	Turnip	53	96.2679	97.3434	1.07547
	Pak choi	25	95.104	95.848	0.744
	Wutacai	7	95.0286	95.8857	0.857143
	Oil seeds	12	94.975	95.4583	0.483333
	Komatsuna	2	95.35	96.4	1.05
	Mizuna	2	95.4	96.6	1.2
	Chinese cabbage (Chiifu)	46	95.9304	95.8304	-0.1
	Taicai	4	95.125	95.925	0.8
	Sarsons, rapid cycling (Z1)	2	93.2	93.95	0.75
	Zicaitai	13	96.9615	97.8385	0.876923
Brassica oleracea	Brussels sprouts	2	96.7	98.3	1.6
	Pointed Cabbage	3	95.6667	98.1	2.43333
	Kohlrabi	19	96.2263	98.3053	2.07895
	Cauliflower	20	96.58	98.53	1.95
	Kale	2	97.35	98.45	1.1
	Curly Kale	2	95.55	98.1	2.55
	Cabbage	31	96.9194	98.4097	1.49032
	Broccoli (HDEM)	23	96.2043	98.9739	2.76957
	Chinese kale (To1000)	4	97.4	98.35	0.95
	White Cabbage	11	95.9727	98.1364	2.16364
	Wild	2	96.45	98	1.55

Supplementary Table 19. **Uniquely aligned reads from various morphotypes.** Proportion of uniquely aligned illumina reads from 318 accessions of *Brassica* on the Chiifu (Chinese cabbage) and To1000 (Chinese kale) reference genomes and on our Z1 (sarsons) and HDEM (broccoli) genome assemblies

Species	Morphotype	Number of accessions	Average % of uniquely aligned reads on reference genome	Average % of uniquely aligned reads on our assembly	Difference
Brassica rapa	Yellow Sarson	1	54.44	52.85	1.59
	Caixin	30	62.08	56.80	5.28
	Edible Flower	1	56.85	58.45	-1.6
	Turnip	53	63.03	56.54	6.49
	Pak choi	25	55.74	53.69	2.05
	Wutacai	7	56.43	53.29	3.14
	Oil seeds	12	67.26	58.96	8.3
	Komatsuna	2	54.16	53.38	0.78
	Mizuna	2	50.36	49.75	0.61
	Chinese cabbage (Chiifu)	46	55.87	53.43	2.44
	Taicai	4	53.33	53.59	-0.26
	Sarsons, rapid cycling (Z1)	2	54.44	52.85	1.59
	Zicaitai	13	63.35	53.37	9.98
Brassica oleracea	Brussels sprouts	2	72.52	62.80	9.72
	Pointed Cabbage	3	74.54	66.02	8.52
	Kohlrabi	19	75.19	67.92	7.27
	Cauliflower	20	75.47	69.92	5.55
	Kale	2	74.18	68.89	5.29
	Curly Kale	2	74.63	66.00	8.63
	Cabbage	31	74.48	68.25	6.23
	Broccoli (HDEM)	23	74.79	67.98	6.81
	Chinese kale (To1000)	4	75.28	68.91	6.37
	White Cabbage	11	75.08	67.80	7.28
	Wild	2	75.52	70.85	4.67

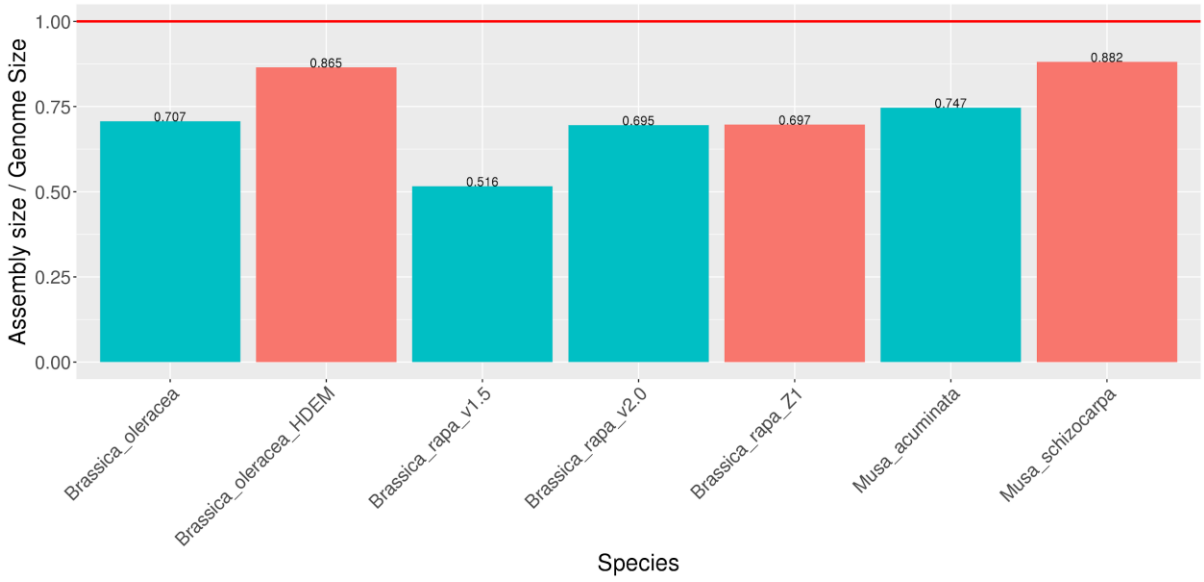
Supplementary Table 20. **Annotation of Z1 and HDEM specific regions.** Regions that are either specific to the HDEM or Z1 accessions or absent from the assembled To1000 or Chiifu chromosomes.

Brassica oleracea				***Brassica rapa***			
Total	Coding exon	Introns	TEs	Total	Coding exon	Introns	TEs
5,262,968	408,723 (7.8%)	732,457 (13.9%)	2,139,201 (40.6%)	7,609,638	454,628 (6.0%)	1,090,175 (14.3%)	3,977,523 (52.3%)

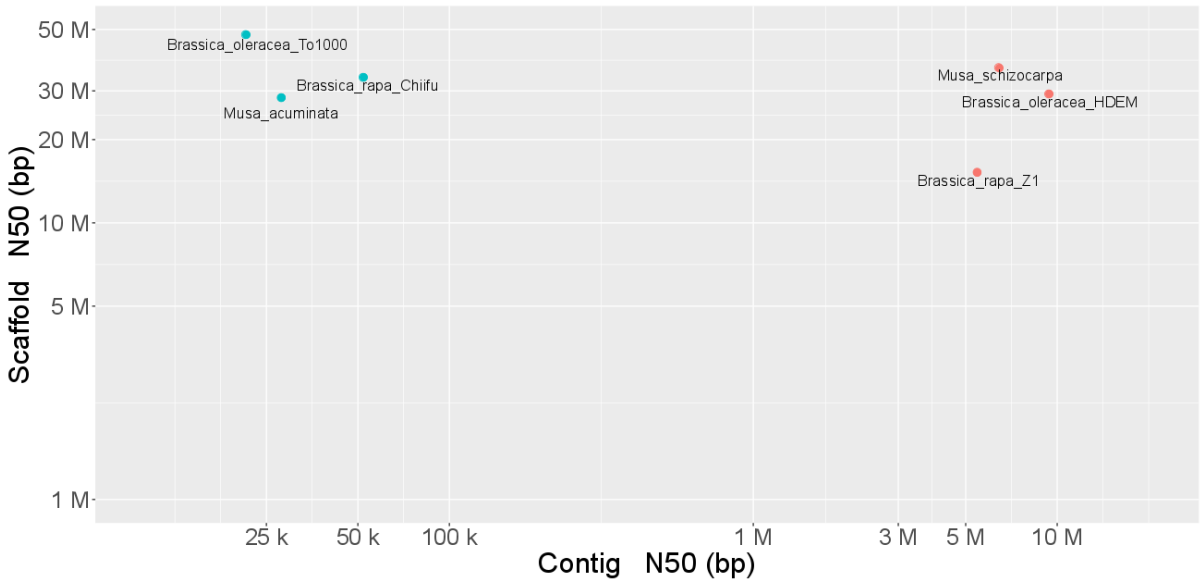
Supplementary Table 21. **R-gene clusters comparison.** Comparison of N proportions between clusters of orthologous disease resistance-like genes in *M. acuminata* and *M. schizocarpa genome assemblies*.

	M. acuminata (Martin et al.)			*M. schizocarpa*		
	Cluster1	Cluster2	Cluster3	Cluster1	Cluster2	Cluster3
Chromosome	chr03	chr03	chr10	chr03	chr03	chr10
Cluster start	27,854,579	31,896,453	22,396,030	37,793,114	41,876,149	27,240,711
Cluster end	27,992,273	32,022,883	22,572,229	37,822,002	41,960,108	27,258,327
Fragment size	137,695	126,431	176,200	28,889	83,960	17,617
Number of N	5255	10372	13145	0	0	0
Proportion of N (%)	3.82	8.20	7.46	0	0	0

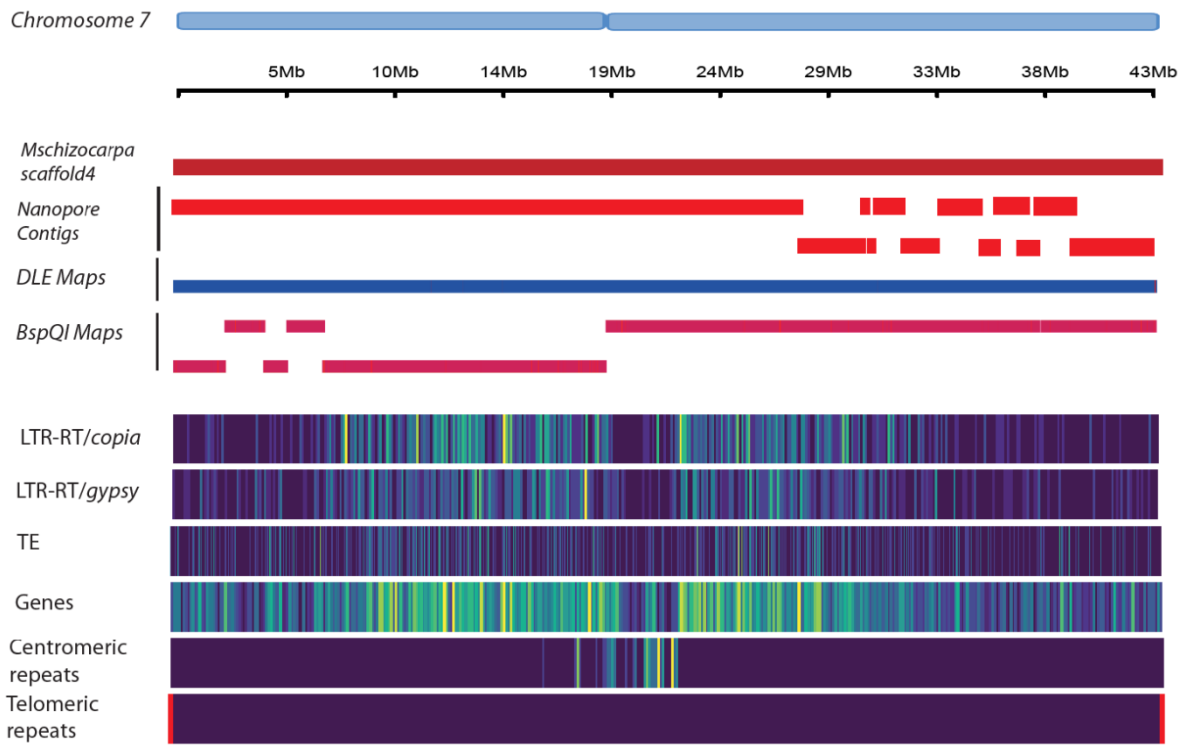
Supplementary Figure 1A. **Assembly sizes comparison.** Comparison of assembly size and estimated genome size of *B. rapa* Chiifu, *B. oleracea* To1000 and *M. acuminata* available assemblies (blue) with our assemblies (orange).



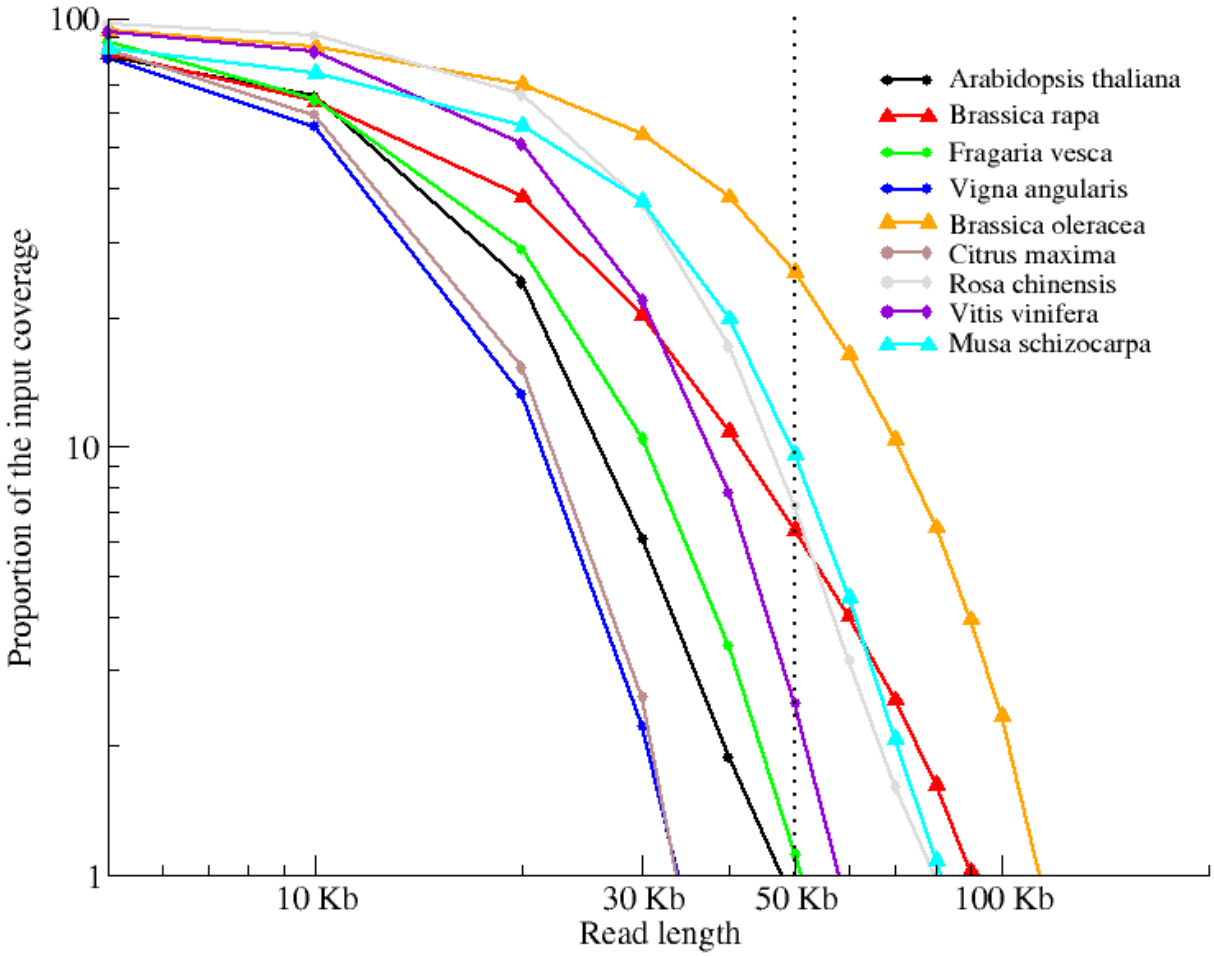
Supplementary Figure 1B. **Assembly N50 comparison.** Comparison of contig and scaffold N50 sizes of available assemblies (blue, short-read strategy) with our assemblies (orange, long-read strategy).



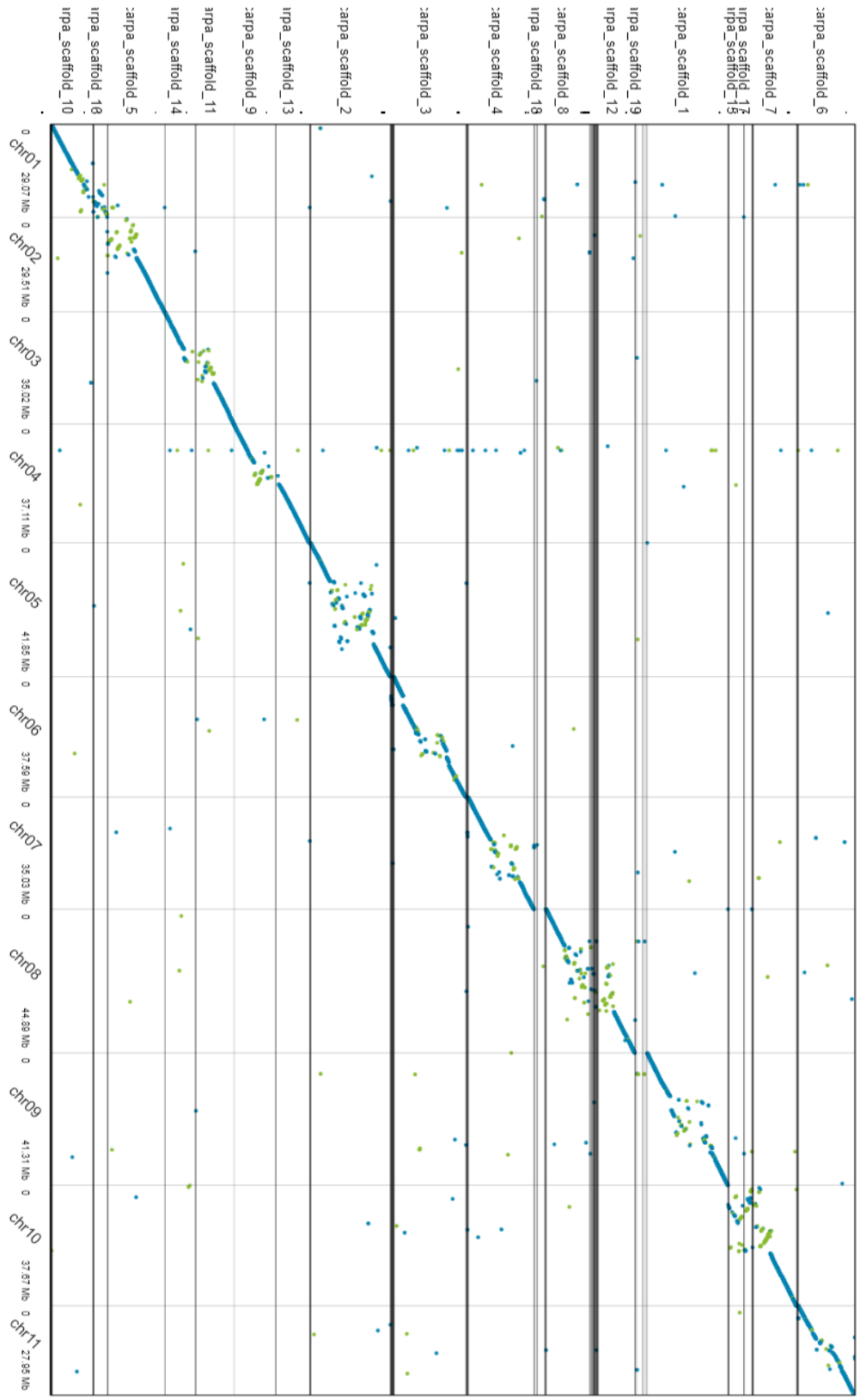
Supplementary Figure 2. Schematic view of Chromosome 7 from the *Musa schizocarpa* genome. The centromere on the ideogram was arbitrarily located in the middle of the region which harbours centromeric repeats (16-22 Mb).



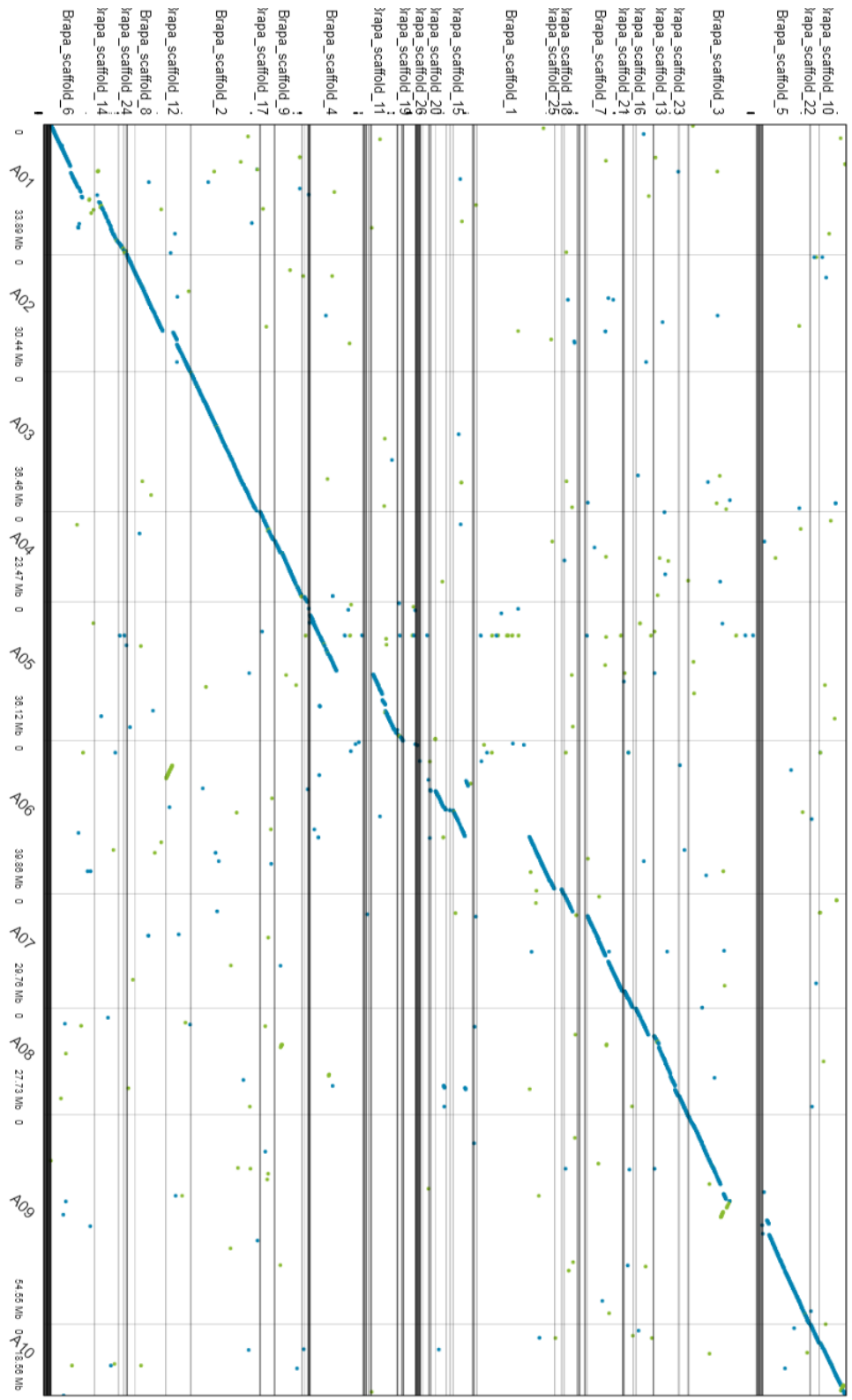
Supplementary Figure 3. Comparison of PACBIO and ONT datasets of nine recent genome sequencing projects. *B. rapa* Z1 (red), *B. oleracea* HDEM (orange) and *M. schizocarpa* (cyan) were sequenced using the ONT platform and *C. maxima* (brown), *V. angularis* (blue), *V. vinifera* (purple), *R. chinensis* (grey), *A. thaliana* (black) and *F. vesca* (green) were sequenced using the PACBIO technology.



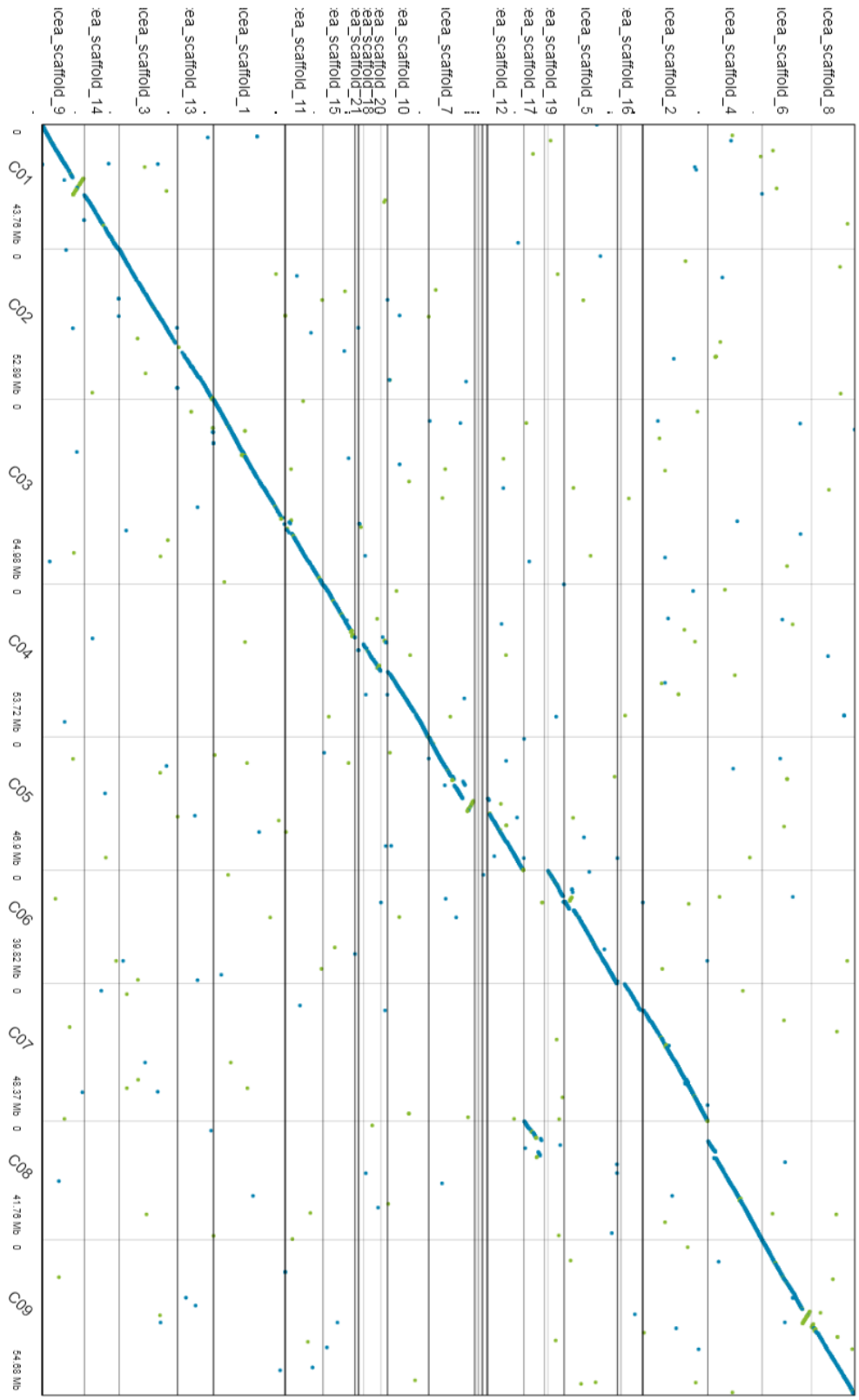
Supplementary Figure 4. Comparison of *Musa* sp. assemblies. Nucmer comparison of *Musa schizocarpa* scaffolds with *Musa acuminata* chromosomes.



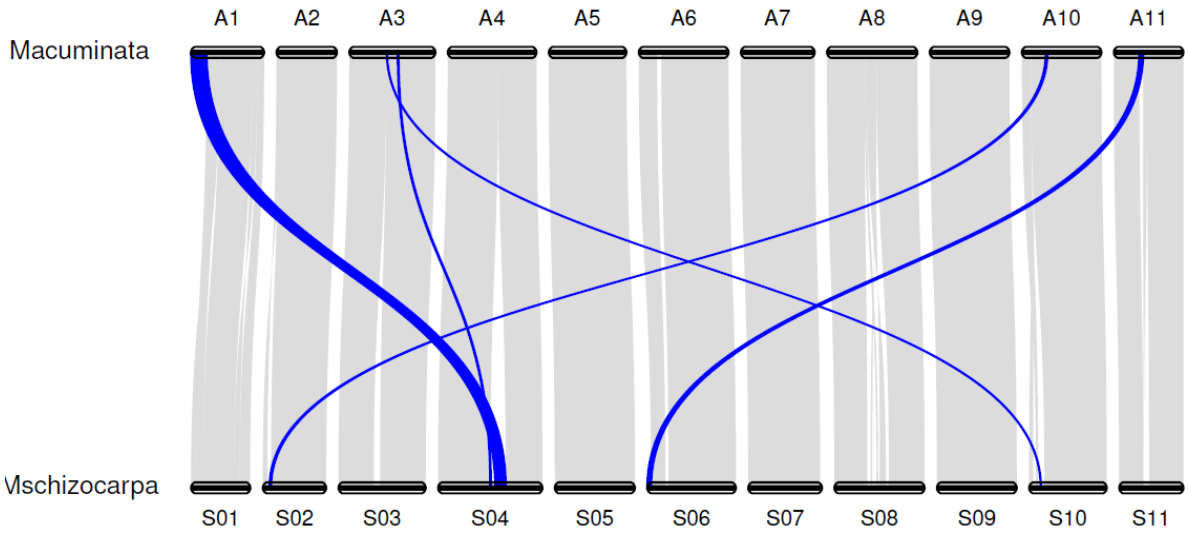
Supplementary Figure 5. Comparison of *B. rapa* assemblies. Nucmer comparison of *Brassica rapa* Z1 scaffolds with *Brassica rapa* Chiifu chromosomes.



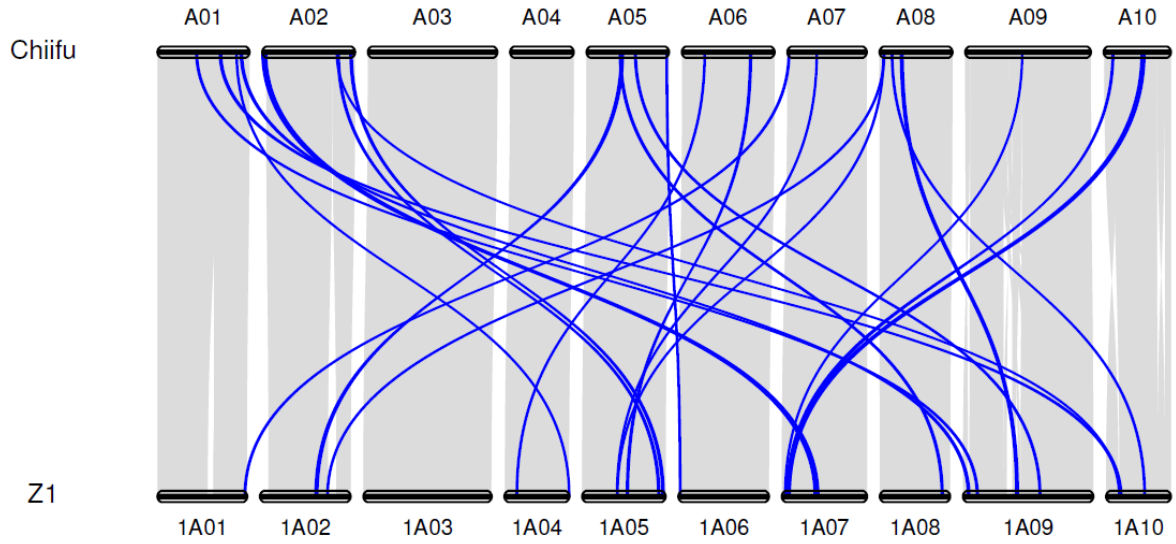
Supplementary Figure 6. Comparison of *B. oleracea* assemblies. Nucmer comparison of *Brassica oleracea* HDEM scaffolds with *Brassica oleracea* To1000 chromosomes.



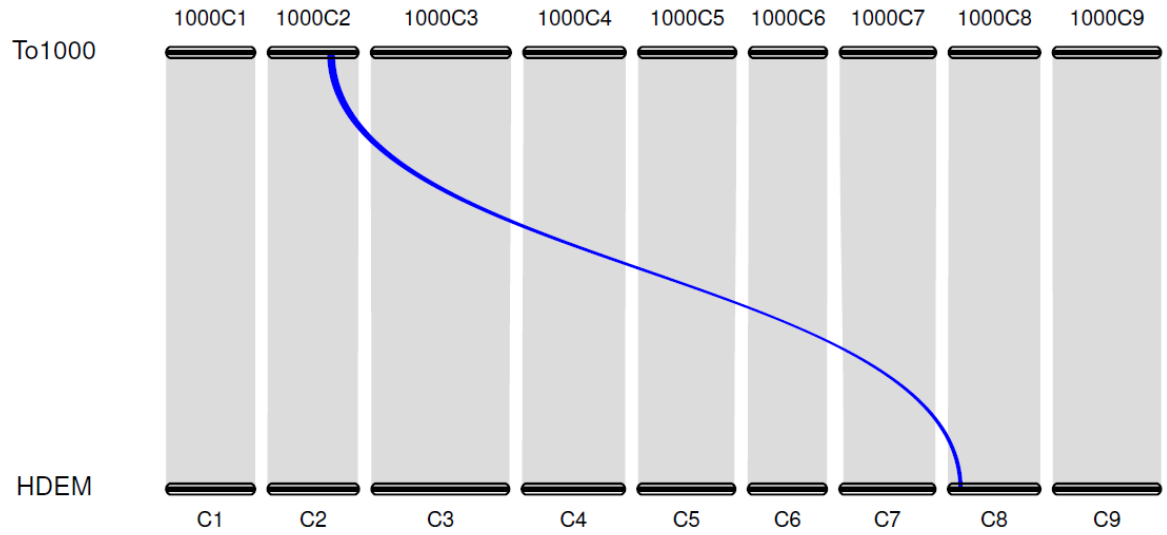
Supplementary Figure 7. Gene order comparison. Synteny visualization between *Musa schizocarpa* and *Musa acuminata* (assembly version1) chromosomes.



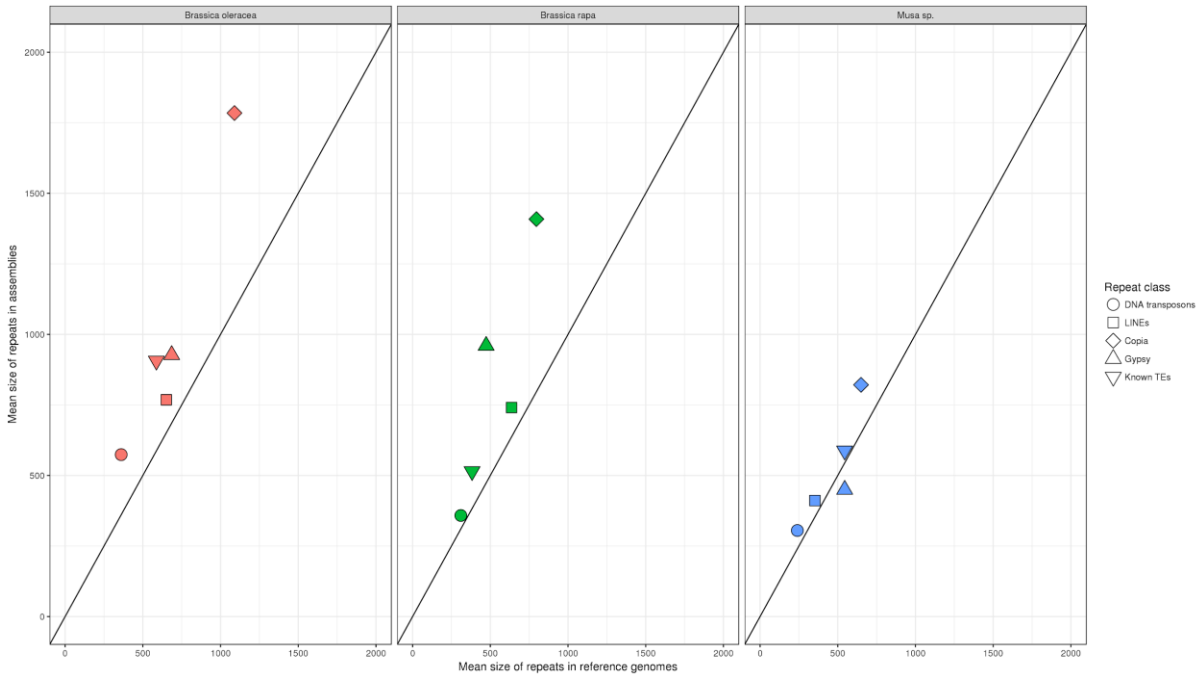
Supplementary Figure 8. Gene order comparison. Synteny visualization between *Brassica rapa* Z1 and *Brassica rapa* Chiifu chromosomes.



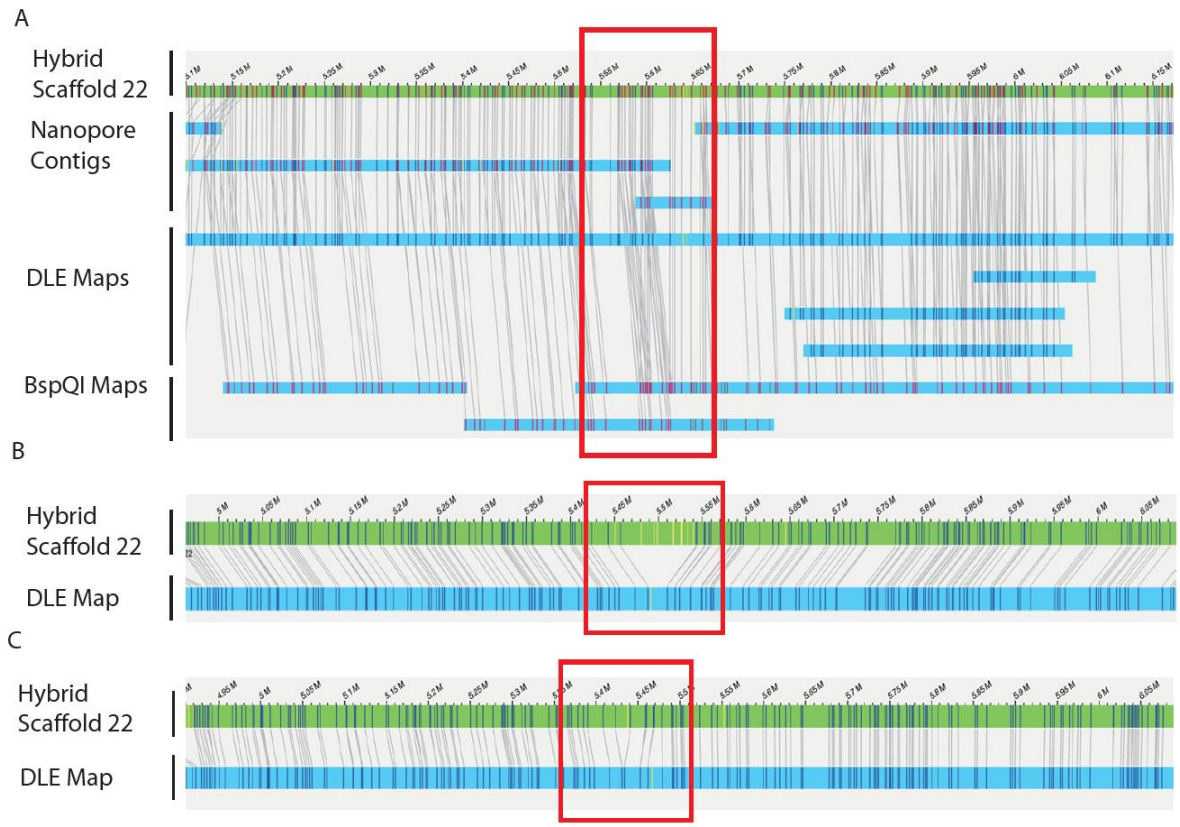
Supplementary Figure 9. Gene order comparison. Synteny visualization between *Brassica oleracea* HDEM and *Brassica oleracea* To1000 chromosomes.



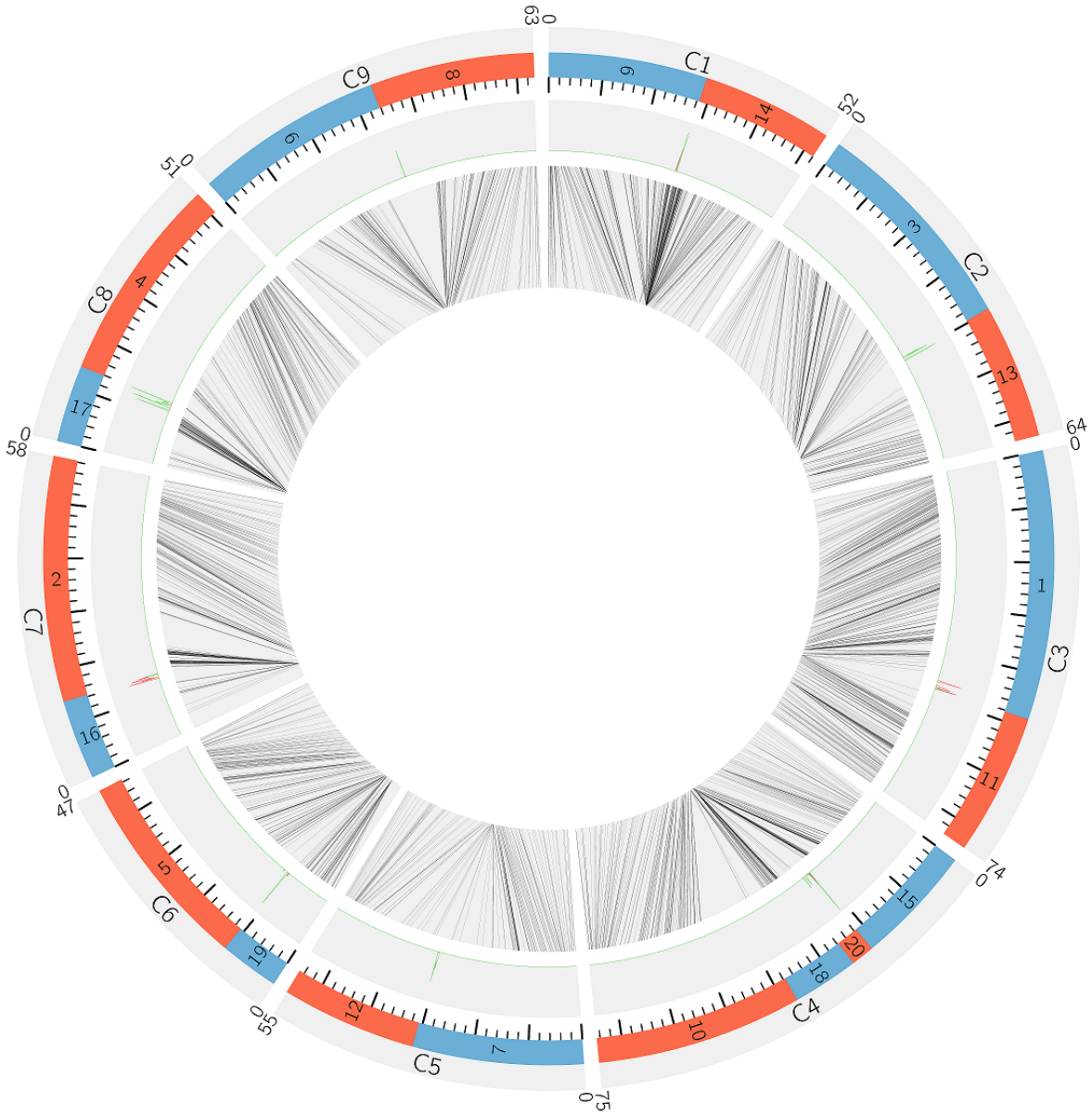
Supplementary Figure 10. Comparison of TE sizes. Mean size allocation for several classes of repeated elements in assemblies compared to reference genomes.



Supplementary Figure 11. Gaps validation. Schematic view of the scaffold 4 of *Musa schizocarpa* showing the overlap between two nanopore contigs (A), that corresponds to a duplicated regions (B). After applying the GapChecker process only one version of the duplicated blocks has been kept (C).

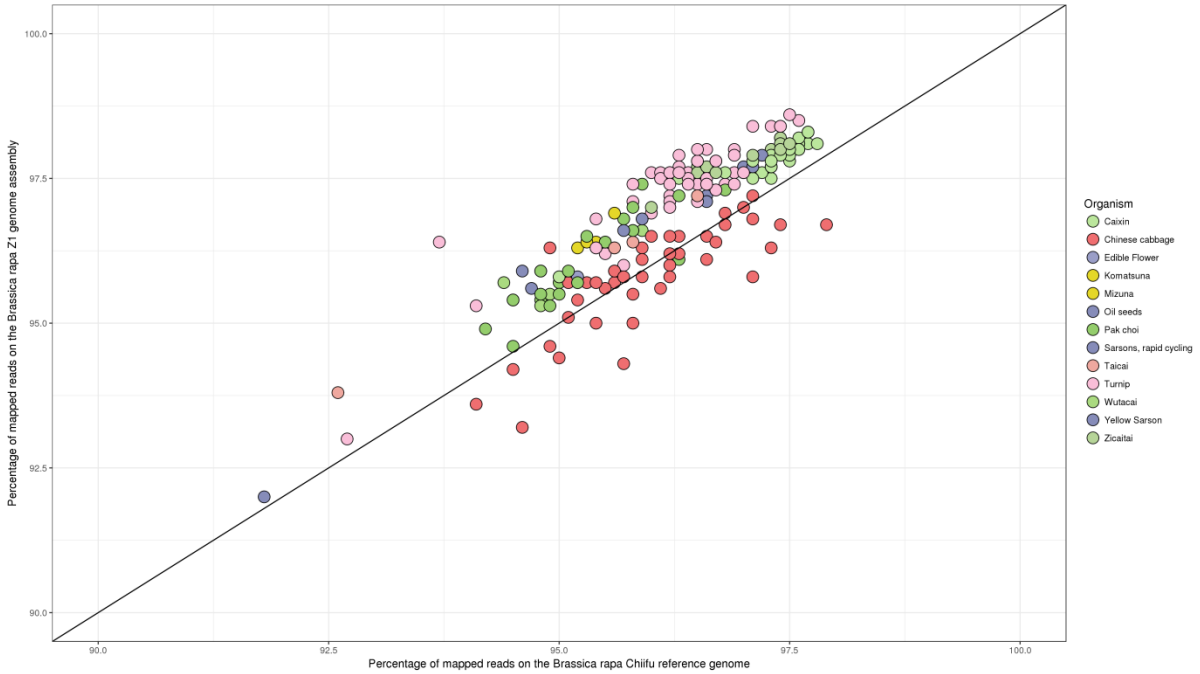


Supplementary Figure 12. Comparison of *B. oleracea* assembly (var HDEM) and *B. oleracea* genetic map (Richelain * HDEM). On this circular representation are represented the 20 scaffolds that were assigned and oriented on the nine *B. oleracea* chromosomes using the SNP markers of the genetic map. The links on the inner circle represent the physical (top) and genetic (bottom) positions of each SNP. Additionally, the position of the centromeres (CenBr1 in red and CenBr2 in green) is indicated in the first circle.

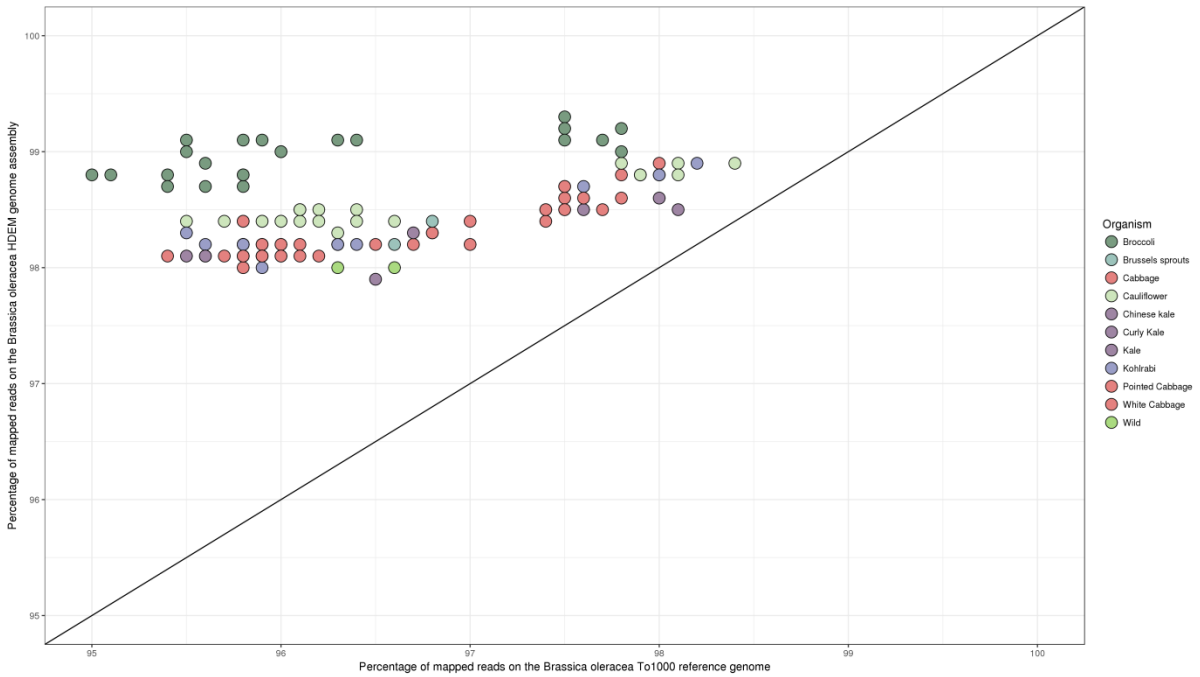


Supplementary Figure 13. Alignment of reads from various morphotypes. A. Proportion of aligned illumina reads from 199 accessions of *B. rapa* on the Chiifu (Chinese cabbage) reference genome (x axis) and on our Z1 (sarsons) genome assembly (y axis). **B.** Proportion of aligned illumina reads from 119 accessions of *B. oleracea* on the TO1000 (Chinese kale) reference genome (x axis) and on our HDEM (broccoli) genome assembly (y axis).

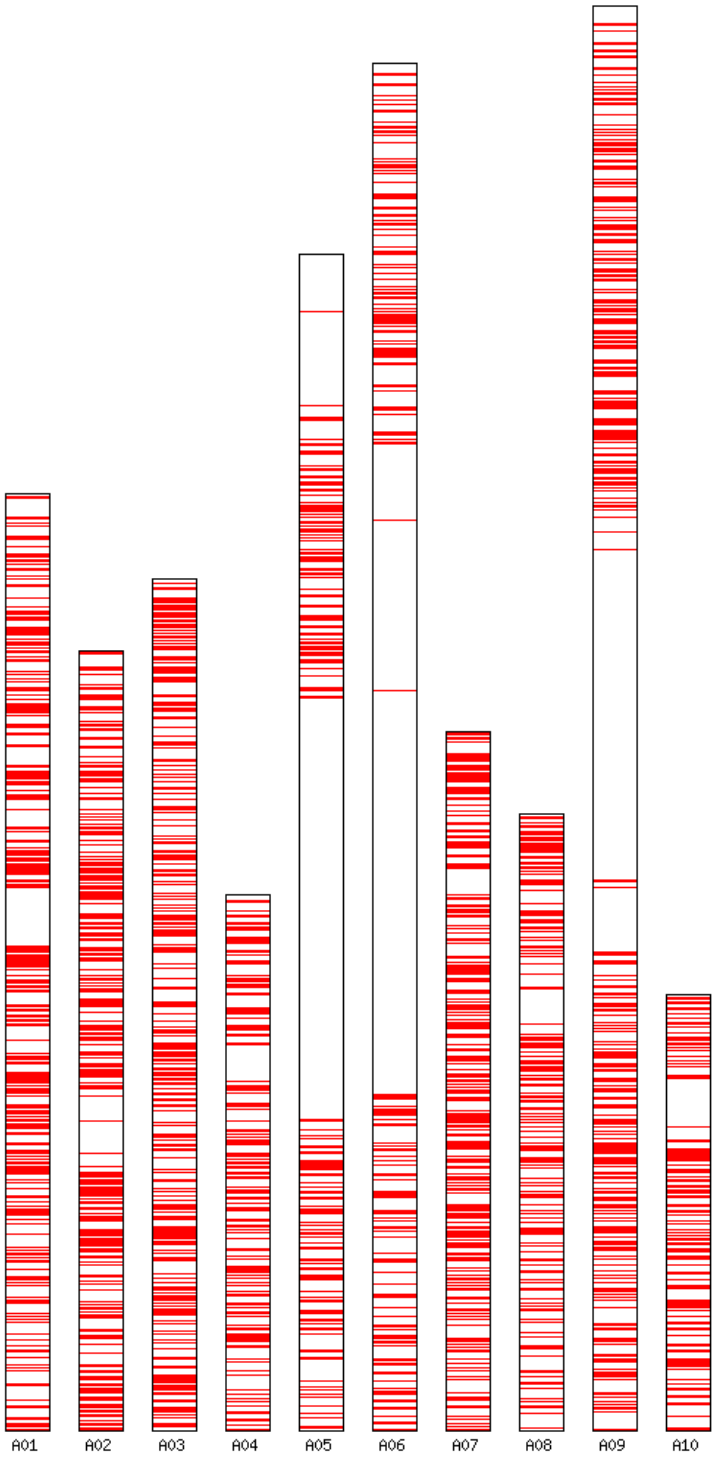
A.



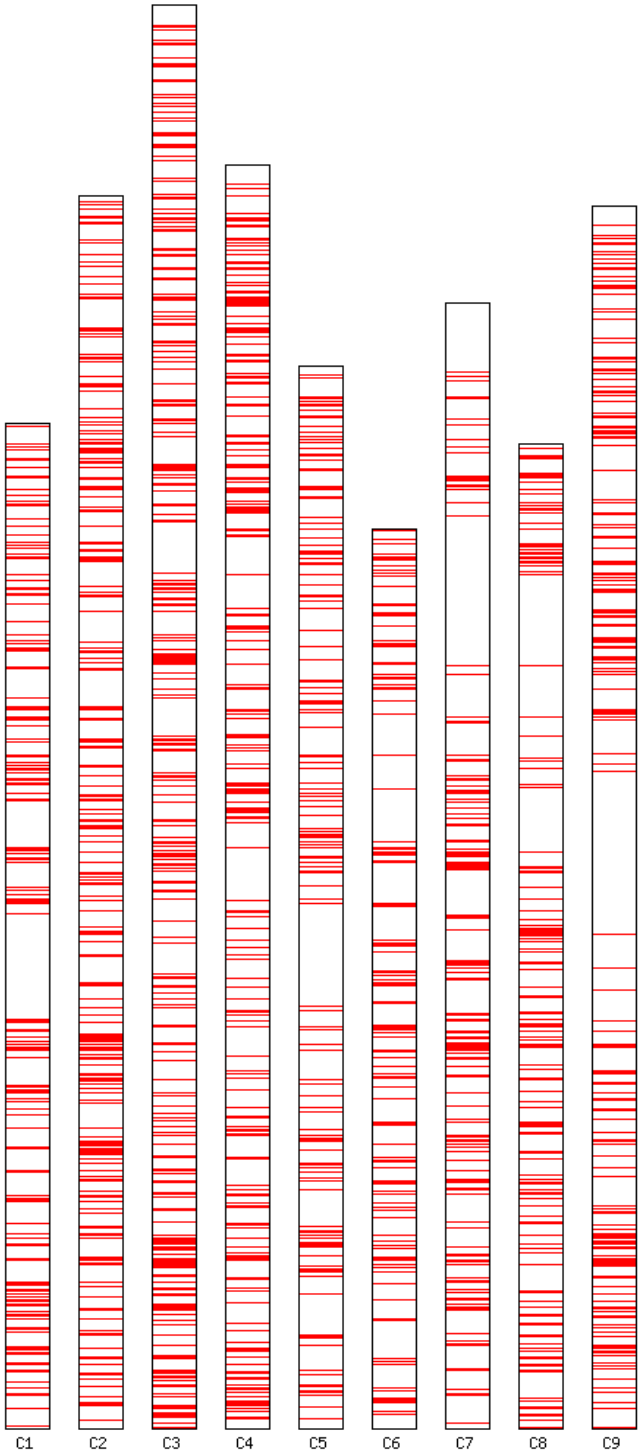
B.



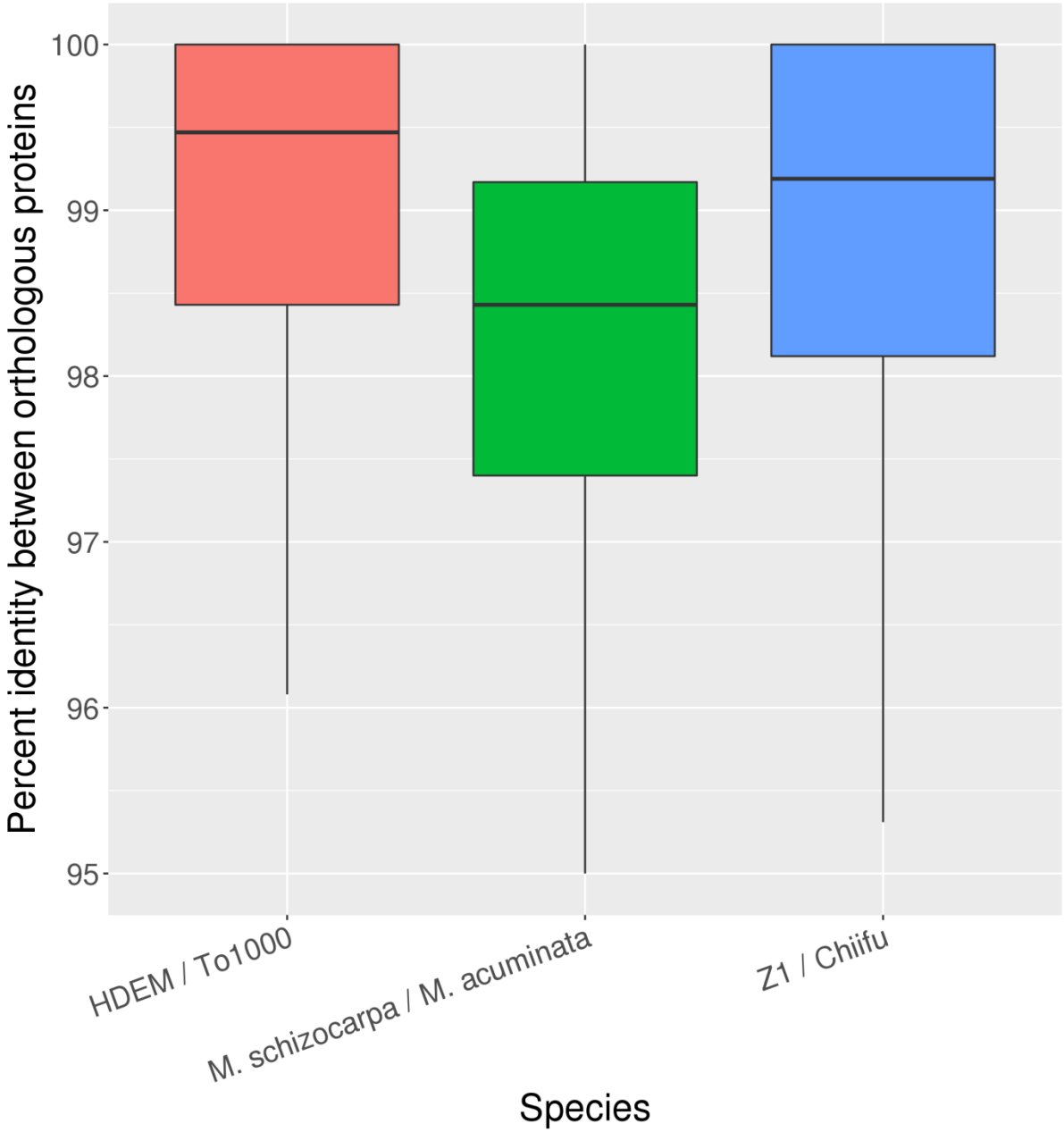
Supplementary Figure 14. Z1 specific regions. Red bands represent regions that are specific to the Z1 accession or absent from the assembled Chiffu chromosomes.



Supplementary Figure 15. HDEM specific regions. Red bands represent regions that are specific to the HDEM accession or absent from the assembled To1000 chromosomes.



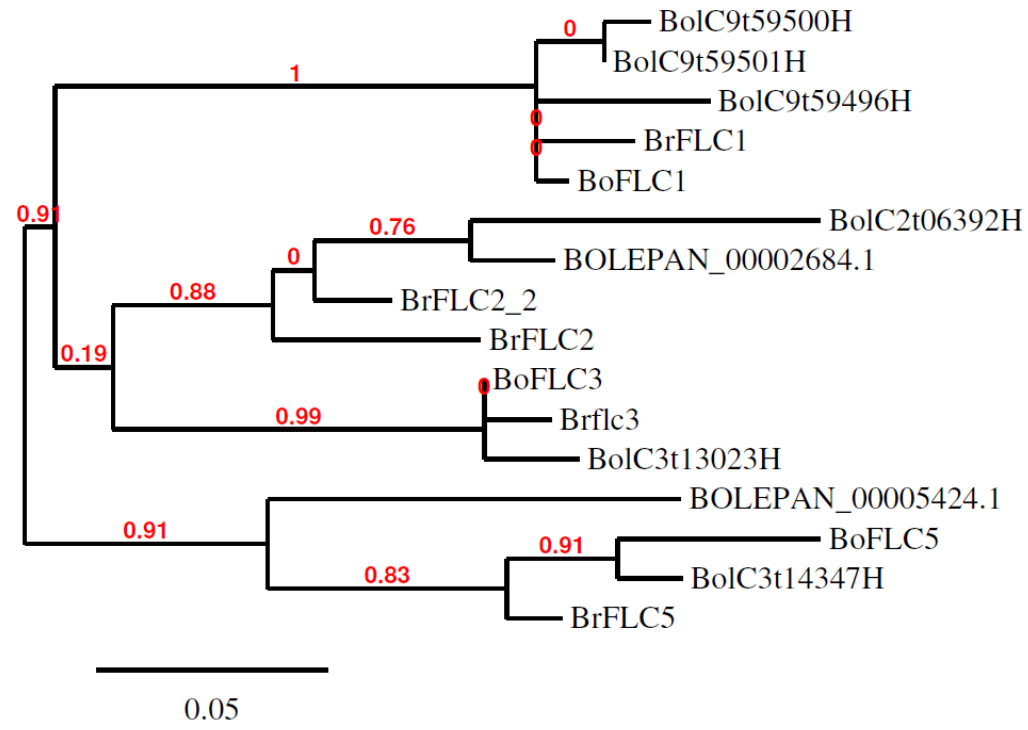
Supplementary Figure 16. Comparison of orthologous proteins. Distribution of the percent identity between proteins of each pair of genomes (N= 39,765 27,333 and 33,290 respectively for HDEM/To1000, M.schizocarpa/M.acuminata and Z1/Chiifu).



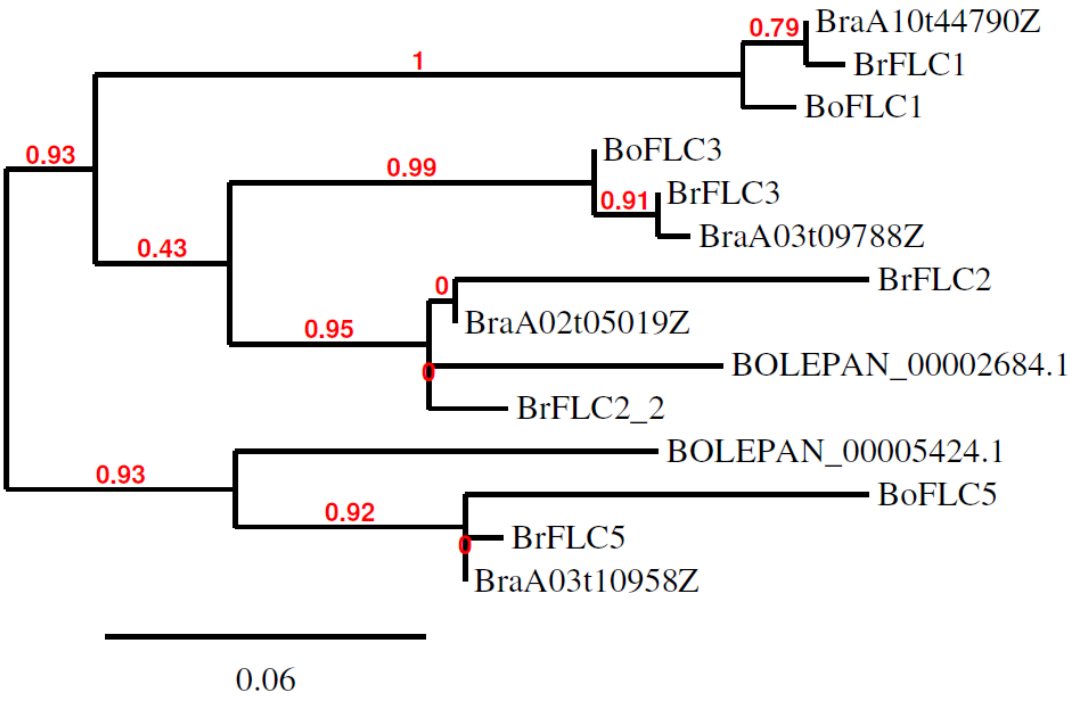
Species	N	Minima	1st quartile	Median	3rd quartile	Maxima
HDEM / To1000	39,765	96.10	98.44	99.48	100.00	100.00
Musa sp.	27,333	95.00	97.40	98.43	99.17	100.00
Z1 / Chiifu	33,290	95.35	98.14	99.20	100.00	100.00

Supplementary Figure 17. Phylogenies of the FLC genes from *B. oleracea* (A) and *B. rapa* (B) annotations. The annotated genes from HDEM and Z1 are prefixed with Bol and Bra respectively.

A.

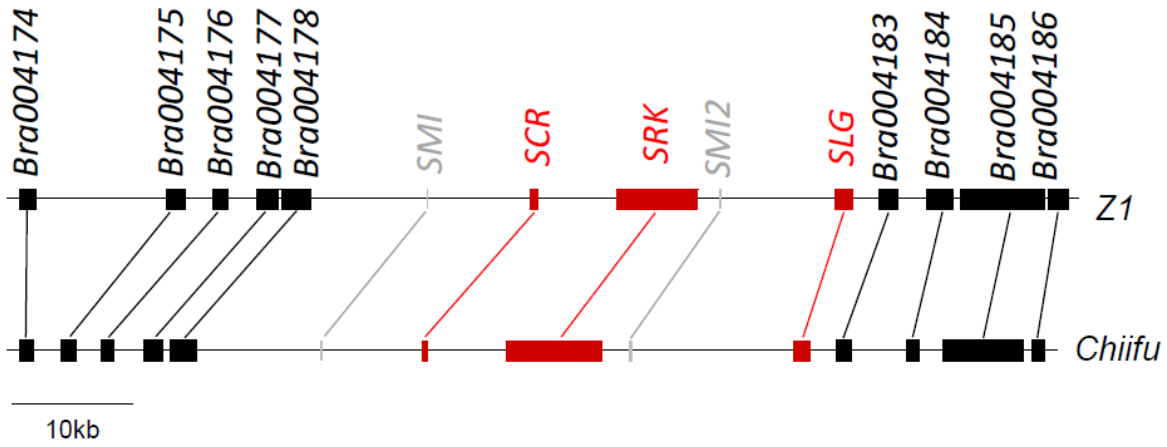


B.

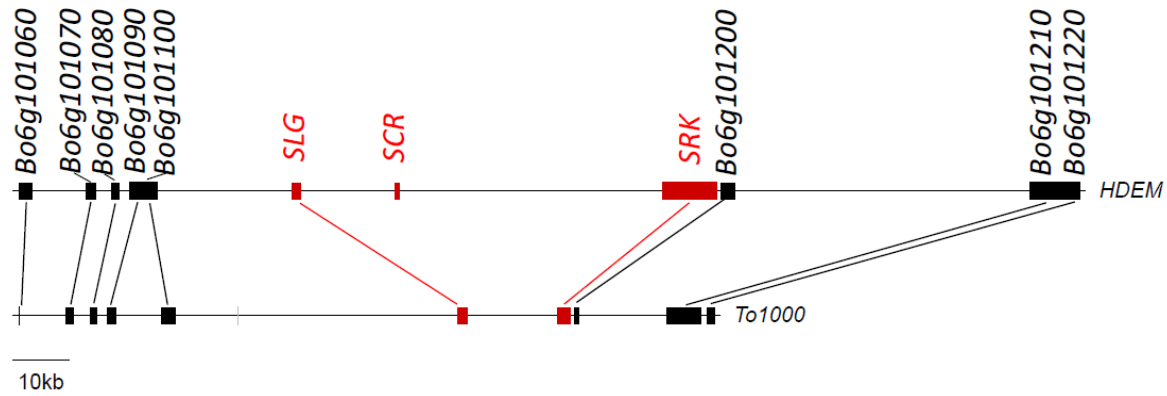


Supplementary Figure 18. S-locus region. A. Synteny in the S-locus region between the two *B. rapa* assemblies. The SI genes *SCR*, *SRK* and *SLG* are shown in red; flanking genes in black; the two S-haplotype-linked small RNAs involved in dominance interactions are shown in grey. **B.** Synteny in the S-locus region between the two *B. oleracea* assemblies. The SI genes *SCR*, *SRK* and *SLG* are shown in red; flanking genes in black. Note that To1000 apparently lacks the *SCR* gene as well as the first two exons of *SRK*. Note also the difference in scale between the *B. rapa* and *B. oleracea* S-locus region.

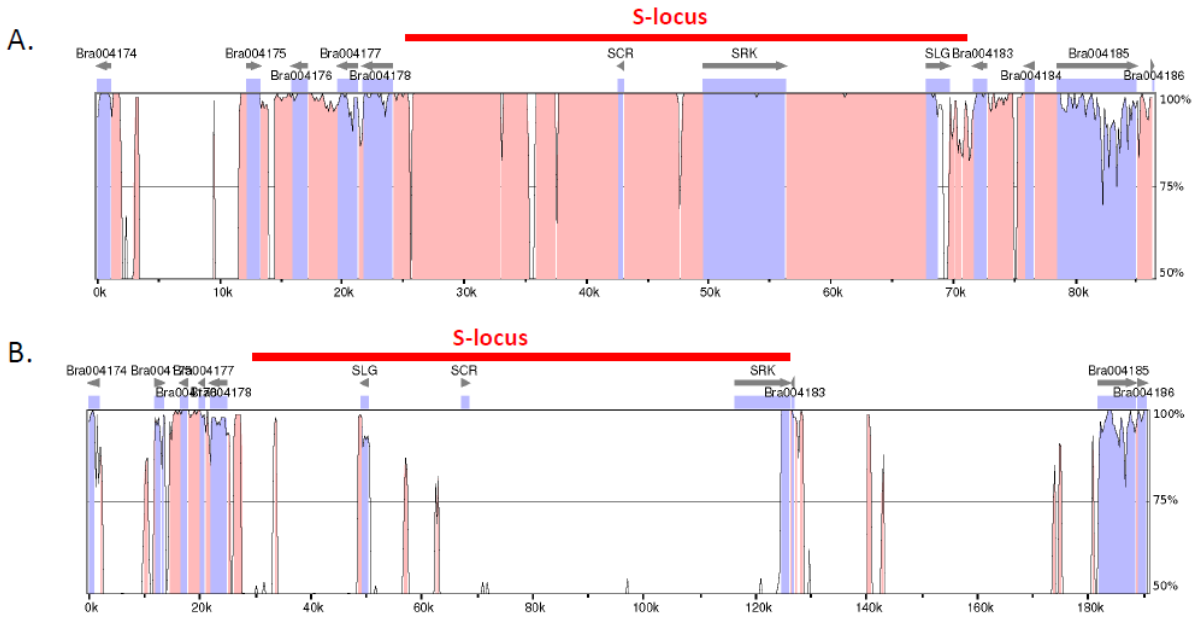
A.



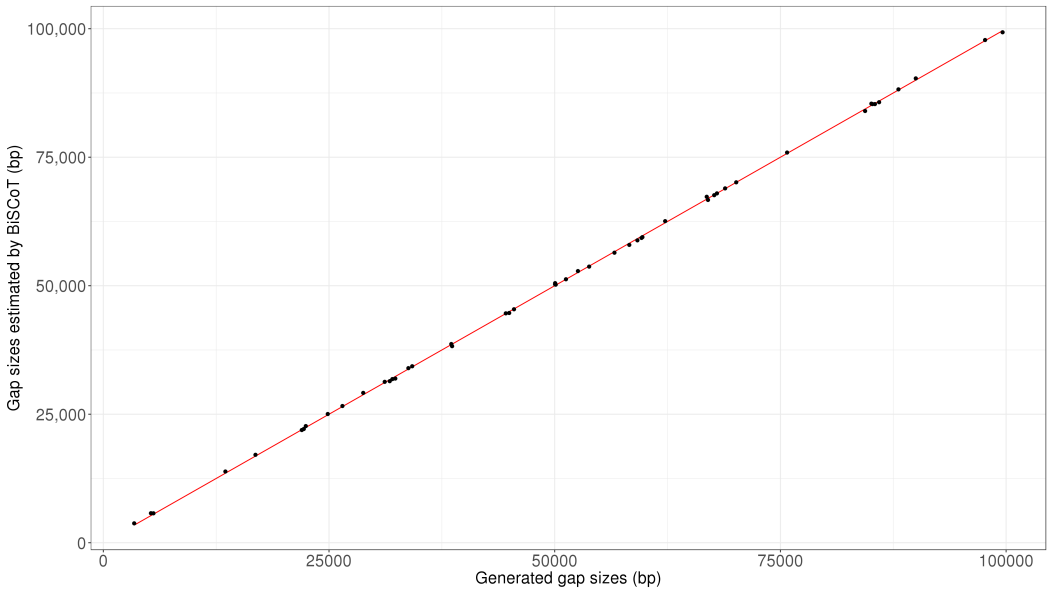
B.



Supplementary Figure 19. Sequence homology of the S-locus region. Analysis of sequence homology in the self-incompatibility locus region using mVISTA between the two *Brassica rapa* cultivars Z1 and Chiifu (A.) and the two *Brassica oleracea* cultivars HDEM and To1000 (B.). The annotations correspond to the locations of the flanking genes as well as the three S-locus genes (*SRK*, *SCR* and *SLG*) in the Z1 and HDEM cultivars, respectively. Note that the sequence homology is much higher in *B. rapa* than in *B. oleracea*, because of S-allele sharing in the former and distinct S-alleles in the latter. Note also the difference in size of the S-locus region between the two species, in agreement with differences in genome size.



10.3 Annexe 3 : "BiSCoT: improving large eukaryotic genome assemblies with optical maps." Istace B, Belser C, Aury JM. PeerJ. 2020 Supplementary Data.



Supplementary Figure 1. Distribution of gap sizes estimated by BiSCoT using the optical maps against the gap sizes introduced in the simulated Human chromosome 1 assembly

	Overlaps	Gaps	Contained contigs
Generated number of events	50	50	5
Cumulative size of events before BiSCoT	2,206,983	2,547,879	1,168,275
Mean size of events before BiSCoT	44,139	50,957	233,655
Min size of events before BiSCoT	278	3,420	215,581
Max size of events before BiSCoT	98,683	99,611	283,879
Remaining number of events after BiSCoT	11	50	0
Cumulative size of events after BiSCoT	30,587	2,549,286	-
Mean size of events after BiSCoT	2680	50,985	-
Min size of events after BiSCoT	278	3,392	-
Max size of events after BiSCoT	5,959	99,792	-

Supplementary Table 1. Metrics of the overlaps, gaps and contained contigs introduced in the simulated Human genome's chromosome 1 assembly before and after applying BiSCoT.

	Simulated contigs	Bionano		BiSCoT	
		Contigs	Scaffolds	Contigs	Scaffolds
Cumulative size	231,215,374	231,215,374	252,023,317	227,870,703	248,278,703
N50	2,429,486	2,363,641	62,178,504	**3,612,465**	61,255,664
L50	46	43	2	**19**	2
N90	**1,677,340**	1,621,176	2,792,747	1,658,986	2,872,288
L90	95	90	10	**55**	9
auN	2,347,866	2,260,168	62,408,635	**4,569,037**	62,312,336
# Ns	0	0	20,789,288	0	20,408,000
NGA50	2,354,935	2,287,164	36,067,187	**2,931,345**	42,227,440
NGA75	1,943,500	1,803,364	2,537,018	**2,034,390**	7,636,847
# misassemblies	**0**	**0**	38	**0**	12

Supplementary Table 2. Metrics of the simulated contigs of the NA12878 chromosome scaffolds and contigs before or after BiSCoT treatment. Bold formatting indicates the best scoring assembly among contigs.

	Nanopore contigs	Bionano		BiSCoT	
		Contigs	Scaffolds	Contigs	Scaffolds
Cumulative size	547,048,580	547,048,579	557,059,506	544,091,840	553,538,342
N50	7,294,619	6,977,344	29,588,784	**12,590,381**	29,458,880
L50	23	23	8	**15**	8
N90	1,304,967	1,183,704	13,953,050	**2,419,034**	13,949,143
L90	85	88	17	**52**	17
auN	9,034,610	8,946,525	29,924,521	**12,646,724**	29,817,651
# Ns	0	0	10,010,927	0	9,446,502
Complete BUSCOs	**1,601 (99.2%)**	**1,601 (99.2%)**	1,598 (99.0%)	**1,601 (99.2%)**	1,600 (99.2%)
Duplicated BUSCOs	235 (14.6%)	235 (14.6%)	232 (14.4%)	**234 (14.5%)**	232 (14.4%)
Missing BUSCOs	10 (0.6%)	**9 (0.6%)**	10 (0.6%)	**9 (0.6%)**	10 (0.6%)

Supplementary Table 3. Metrics of the *Brassica oleracea* HDEM scaffolds and contigs before or after BiSCoT treatment. Bold formatting indicates the best scoring assembly among contigs.

	Nanopore contigs	Bionano		BiSCoT	
		Contigs	Scaffolds	Contigs	Scaffolds
Cumulative size	373,437,357	373,437,357	406,471,180	369,747,840	402,627,824
N50	3,793,063	3,603,274	15,479,745	**5,519,975**	15,275,286
L50	25	26	8	**17**	8
N90	**202,023**	154,330	1,748,645	181,213	1,674,920
L90	264	309	31	**221**	31
auN	5,532,997	5,453,354	18,883,302	**7,237,727**	18,700,050
# Ns	0	0	33,033,823	0	32,879,984
Complete BUSCOs	1,604 (99.4%)	**1,605 (99.5%)**	1,604 (99.4%)	**1,605 (99.5%)**	1,604 (99.4%)
Duplicated BUSCOs	**233 (14.4%)**	235 (14.6%)	234 (14.5%)	**233 (14.4%)**	233 (14.4%)
Missing BUSCOs	**7 (0.5%)**	**7 (0.5%)**	7 (0.5%)	**7 (0.5%)**	7 (0.5%)

Supplementary Table 4. Metrics of the *Brassica rapa* Z1 scaffolds and contigs before or after BiSCoT treatment. Bold formatting indicates the best scoring assembly among contigs.

	Nanopore contigs	Bionano		BiSCoT	
		Contigs	Scaffolds	Contigs	Scaffolds
Cumulative size	518,619,765	518,619,765	526,521,784	517,940,161	525,719,686
N50	4,019,832	2,097,979	36,762,080	**7,987,169**	36,858,856
L50	33	60	6	**24**	6
N90	554,125	292,444	9,697,206	**888,370**	9,721,221
L90	180	310	15	**92**	15
auN	5,390,023	5,285,943	33,951,065	**7,477,787**	33,460,868
# Ns	0	0	7,902,019	0	7,779,525
Complete BUSCOs	1,558 (96.6%)	**1,562 (96.8%)**	1,561 (96.7%)	1,560 (96.6%)	1,559 (96.6%)
Duplicated BUSCOs	69 (4.3%)	**68 (4.2%)**	68 (4.2%)	**68 (4.2%)**	70 (4.3%)
Missing BUSCOs	**34 (2.0%)**	**34 (2.0%)**	34 (2.0%)	**34 (2.0%)**	34 (2.0%)

Supplementary Table 5. Metrics of the *Musa schizocarpa* scaffolds and contigs before or after BiSCoT treatment. Bold formatting indicates the best scoring assembly among contigs.

	Nanopore contigs	Bionano		BiSCoT	
		Contigs	Scaffolds	Contigs	Scaffolds
Cumulative size	652,555,937	652,555,937	665,966,510	649,440,360	662,857,763
N50	2,985,938	2,985,938	31,920,664	**3,969,296**	31,819,818
L50	51	51	10	**42**	10
N90	488,936	485,536	13,186,102	**612,779**	13,076,771
L90	267	261	21	**216**	21
auN	4,975,870	4,969,973	29,298,774	**5,711,235**	29,207,200
# Ns	0	0	13,410,573	0	13,417,403
Complete BUSCOs	1,569 (97.3%)	1,572 (97.4%)	1,576 (97.6%)	**1,576 (97.6%)**	1,573 (97.4%)
Duplicated BUSCOs	**30 (1.9%)**	31 (1.9%)	31 (1.9%)	31 (1.9%)	31 (1.9%)
Missing BUSCOs	26 (1.5%)	24 (1.5%)	23 (1.5%)	**23 (1.5%)**	24 (1.5%)

Supplementary Table 6. Metrics of the *Sorghum bicolor Tx430* scaffolds and contigs before or after BiSCoT treatment. Bold formatting indicates the best scoring assembly among contigs.

10.4 Annexe 4 : "Sequencing and Chromosome-Scale Assembly of Plant Genomes, Brassica rapa as a Use Case" Istace, B.; Belser, et al. *Biology* 2021 Supplementary Data.

Table S1. Metrics of Nanopore datasets

Readset	All reads	Longest reads	Filtlong reads
Cumulative size	93,615,101,387	15,000,004,992	15,000,009,495
Coverage	186x	30x	30x
Number of reads	13,094,187	214,687	316,605
N50	26,978	68,127	50,558
Number of reads > 50kb	296,164	214,687	117,159
coverage > 50kb	38x	30x	15x

Table S2. Metrics of raw Smartdenovo Nanopore assemblies. Statistics were generated on contigs of more than 30kb in size.

Readset	All reads	Longest reads	Filtlong reads
Cumulative size	413,393,283	354,455,409	320,273,956
Number of contigs	596	323	429
N50 (L50)	10,591,557 (14)	4,616,132 (19)	2,264,044 (35)
N90 (L90)	246,123 (185)	326,576 (146)	263,776 (203)
auN	9,674,642	6,312,816	3,679,656
Max. size	27,174,673	18,229,467	17,762,612
Complete buscos (%)	1,576 (97.6%)	1,436 (88.9%)	1,557 (96.5%)
Merqury score	25.1333	20.3158	23.8738

Table S3. Metrics of raw Flye Nanopore assemblies. Statistics were generated on contigs of more than 30kb in size.

Readset	All reads	Longest reads	Filtlong reads
Cumulative size	308,405,611	340,681,490	320,203,564
Number of contigs	195	266	382
N50 (L50)	8,877,964 (12)	13,456,414 (10)	6,180,078 (13)
N90 (L90)	974,980 (48)	497,751 (58)	310,436 (106)
auN	9,118,954	12,335,383	8,103,042
Max. size	21,466,788	27,098,069	22,540,294
Complete buscos (%)	1,585 (98.2%)	1,584 (98.1%)	1,594 (98.8%)
Merqury score	28.7338	27.4808	28.0033

Table S4. Metrics of raw Wtdbg2 Nanopore assemblies. Statistics were generated on contigs of more than 30kb in size.

Readset	All reads	Longest reads	Filtlong reads
Cumultive size	728,902,491	541,375,314	381,524,645
Number of contigs	9,550	5,095	3,213
N50 (L50)	87,609 (2,265)	142,868 (574)	206,244 (357)
N90 (L90)	38,374 (7,405)	42,678 (3,598)	44,750 (2,179)
auN	144,875	654,911	450,988
Max. size	1,853,134	6,900,900	3,496,279
Complete buscos (%)	1,455 (90.1%)	1,365 (84.6%)	1,402 (86.9%)
Merqury score	17.4181	15.9445	18.4839

Table S5. Metrics of the raw Necat Nanopore assembly. Statistics were generated on contigs of more than 30kb in size.

Readset	All reads
Cumulative size	442,919,932
Number of contigs	299
N50 (L50)	10,445,217 (12)
N90 (L90)	856,761 (58)
auN	14,175,039
Max. size	45,040,585
Complete buscos (%)	1,567 (97.1%)
Merqury score	27.513

Table S6. Metrics of the Necat assembly after each polishing step. Statistics were generated on contigs of more than 30kb in size.

	Raw	**Racon**	**Medaka**	**Hapo-G x2**
Cumulative size	442,919,932	443,505,353	443,619,512	443,649,441
Number of contigs	299	299	299	299
N50 (L50)	10,445,217 (12)	10,458,156 (12)	10,461,728 (12)	10,461,875 (12)
N90 (L90)	856,761 (58)	857,014 (58)	850,928 (58)	857,267 (58)
auN	14,175,039	14,197,548	14,204,585	14,202,687
Max. size	45,040,585	45,103,281	45,117,601	45,115,632
Complete buscos (%)	1,567 (97.1%)	1,592 (98.6%)	1,602 (99.3%)	1,604 (99.4%)
Merqury score	27.513	29.8573	31.0464	36.4423

Table S7. Hybrid scaffolding results.

	Nanopore contigs	Hybrid scaffolds	Contigs left	Assembly after Hybrid scaffolding	Scaffolds after BisCoT treatment	Contigs after BisCoT treatment	Scaffolds after polishing with Hapo-G
Number	299	53	208	261	236	296	236
N50 (L50)	10,461,875 (12)	16,816,852 (8)	154,636 (53)	16,816,852 (8)	17,017,530 (8)	10,256,302 (14	17,017,634 (8)
N90 (L90)	857,267 (58)	4,034,000 (45)	67,775 (152)	1,954,922 (32)	3,409,196 (30)	981,751 (59)	3,409,175 (30)
Min size	30,349	118,769	30,349	30,349	30,347	30,347	30,348
Max size	45,115,632	44,052,815	978,115	44,052,815	44,068,965	27,443,240	44,069,534
Cumulative size	443,649,441	420,416,306	26,002,488	446,418,794	443,948,706	441,033,740	443,951,349
%N	0%	0.7%	0%	0.66%	0.66%	0%	0.66%

Table S8. Metrics of the Pore-C raw PromethION run with (left column) or without (right column) reads of more than 100kb in size.

Readset	All reads	Reads < 100kb
Cumulative size	15,551,590,510	15,502,009,110
Number of reads	5,894,741	5,894,524
N50	3,990	3,990
Number of reads > 100kb	217	0
Cumulative size > 100kb	49,581,400	0

Figure S1. Dotplots of polished Necat contigs aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right).

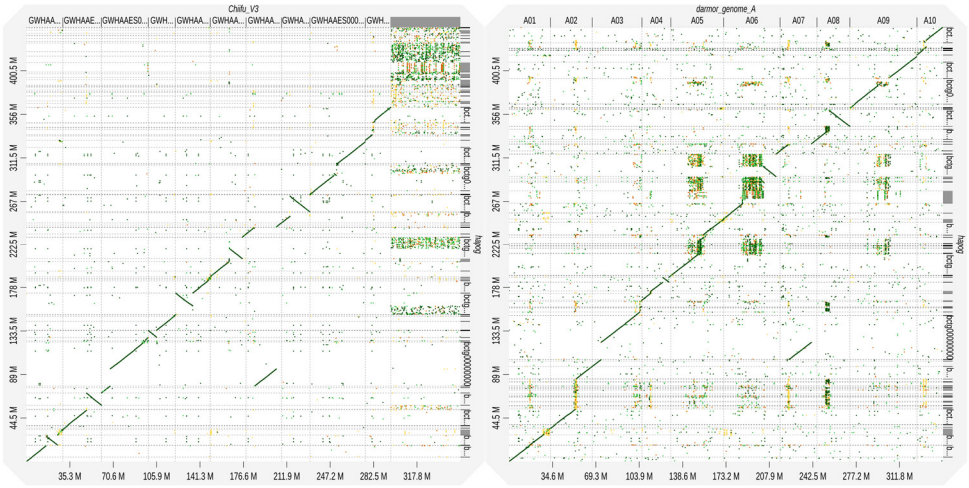


Figure S2. Dotplots of Bionano scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right).

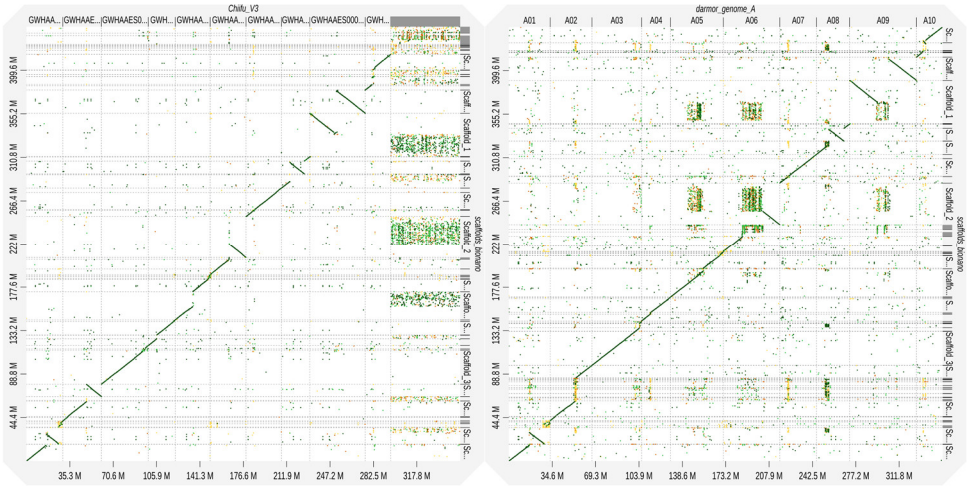


Figure S3. Dotplots of Omni-C scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right).

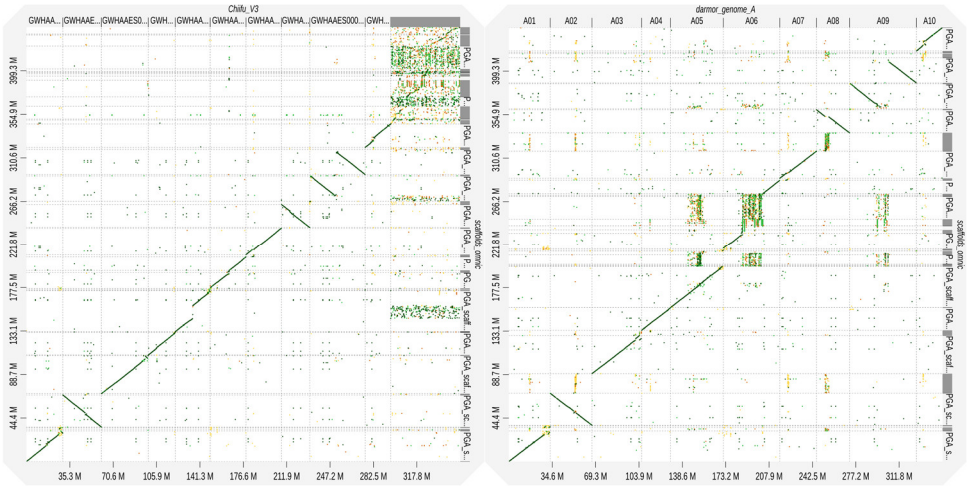


Figure S4. Dotplots of Pore-C scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right).

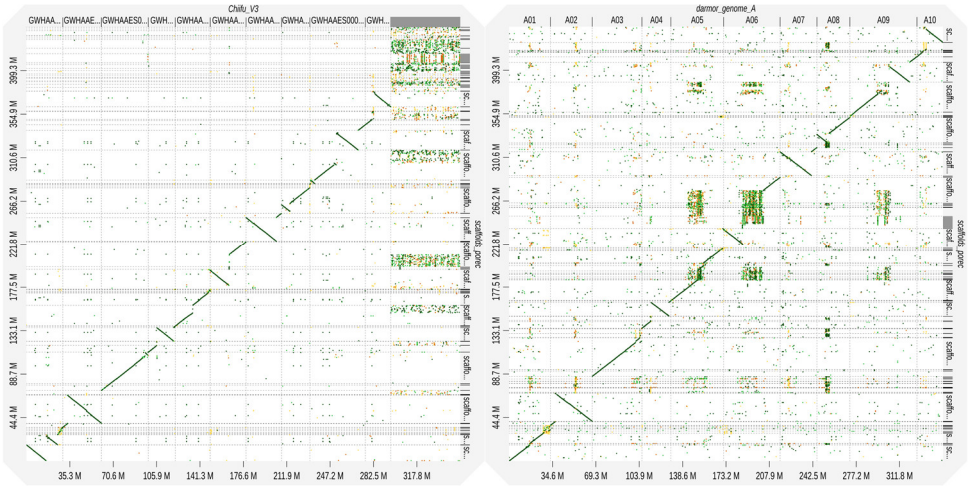


Figure S5. Dotplots of Bionano + Omni-C scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right).

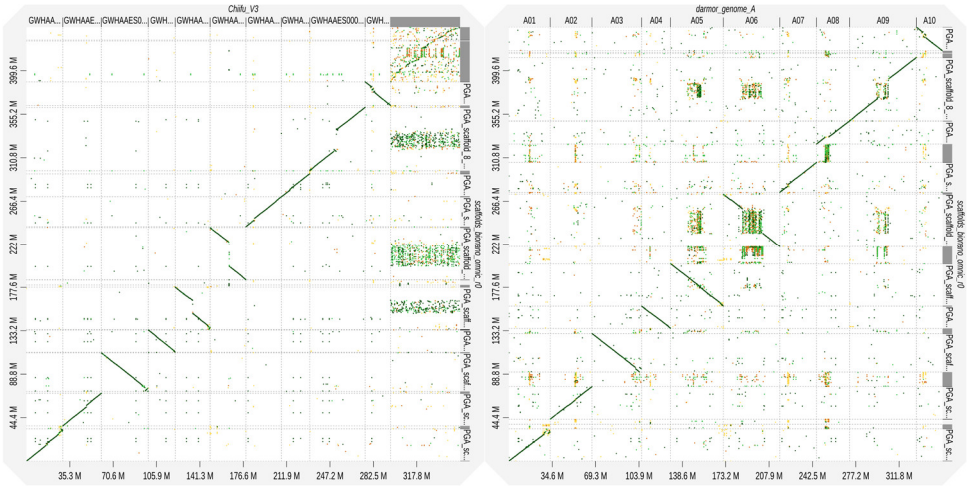


Figure S6. Dotplots of Bionano + Pore-C scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right).

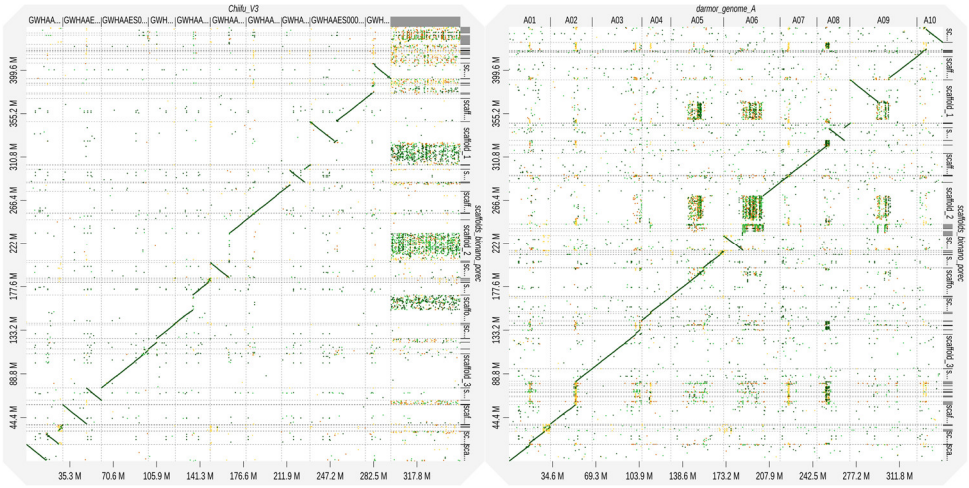


Figure S7. Dotplots of Omni-C + Bionano scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right).

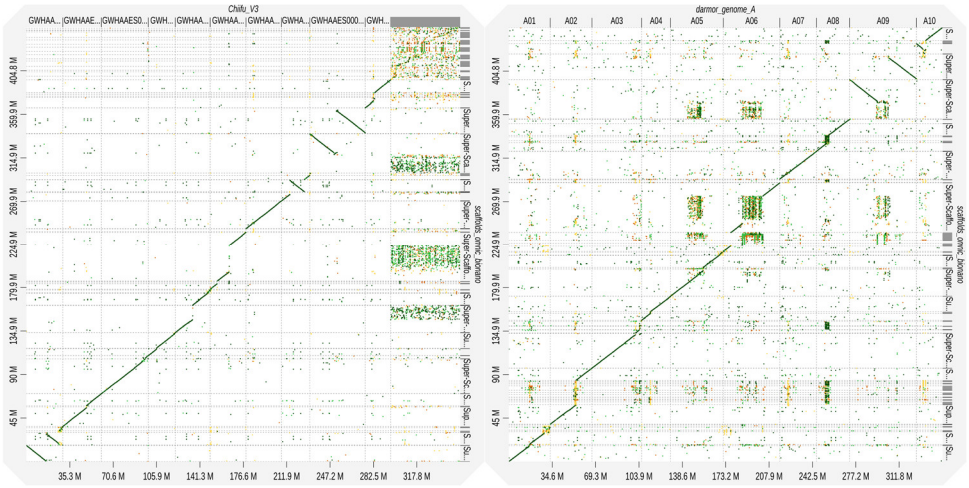


Figure S8. Dotplots of Pore-C + Bionano scaffolds aligned to *Brassica rapa* Chiifu (left) or *Brassica napus* Darmor-BZH A genome (right).

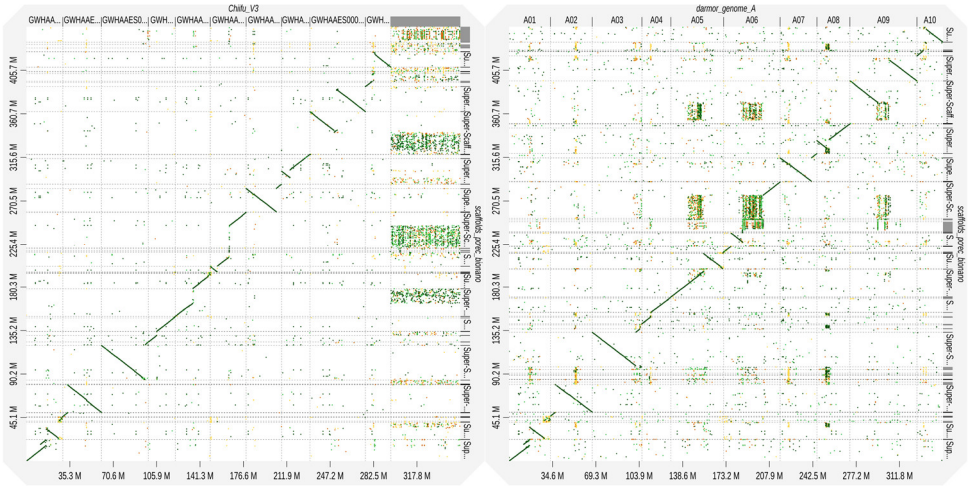


Figure S9. Contact Map generated with the Omni-C library (mapping on the nanopore contigs). The left panel shows the contact before review. The right panel shows the contact map after review.

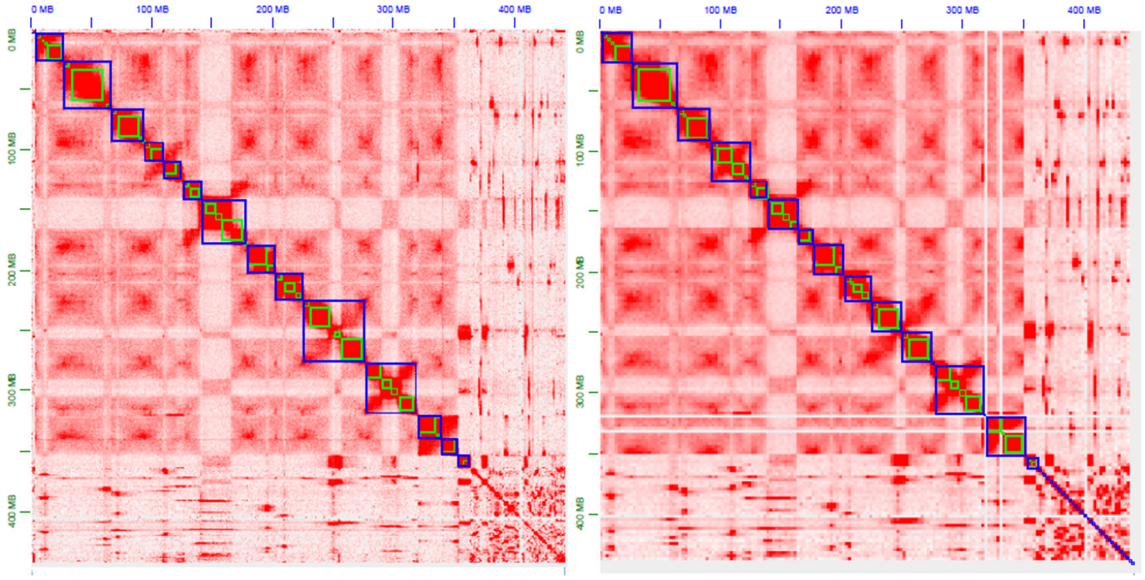


Figure S10. Contact Map generated with the Omni-C library (mapping on the bionano scaffolds). The left panel shows the contact before review. The right panel shows the contact map after review.

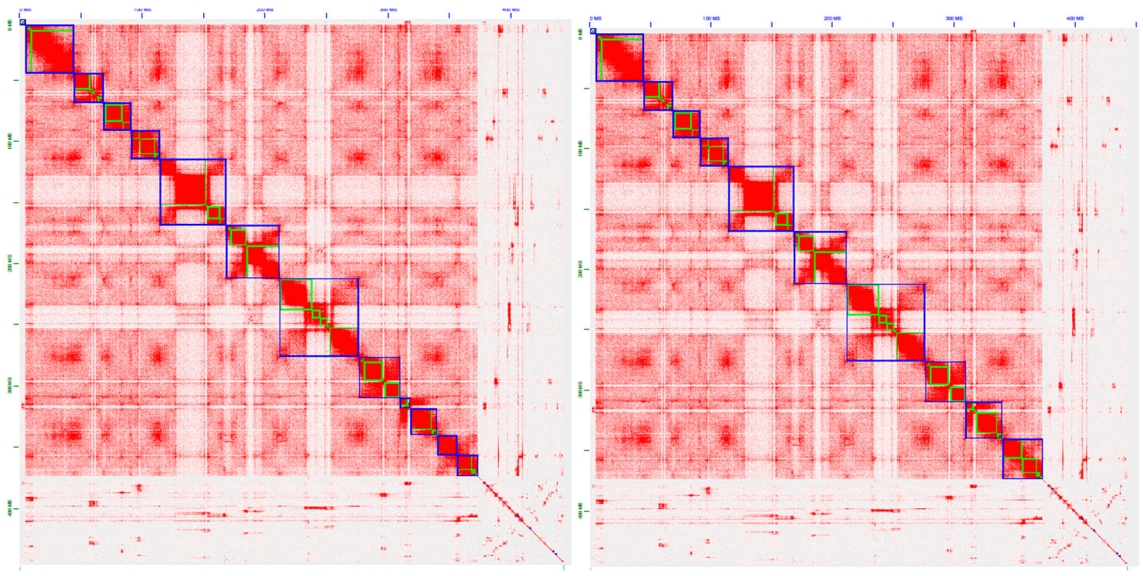


Figure S11. Conflict example detected by the optical maps on the Omni-C scaffolds. The two first lines represent the DLE-1 maps. The third line represents one Omni-C scaffold. The two last lines represent the BspQI maps. The vertical lines show the correspondence between the labels of the scaffold and the labels of the maps. In the middle, we can see a region without any correspondence which leads to a break of the scaffold.

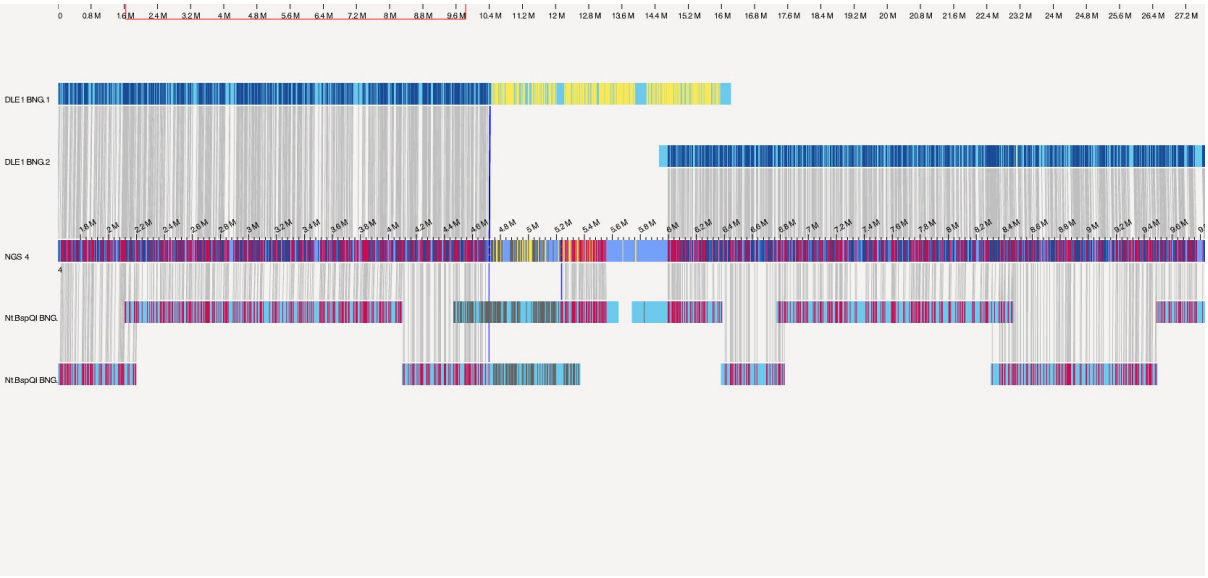
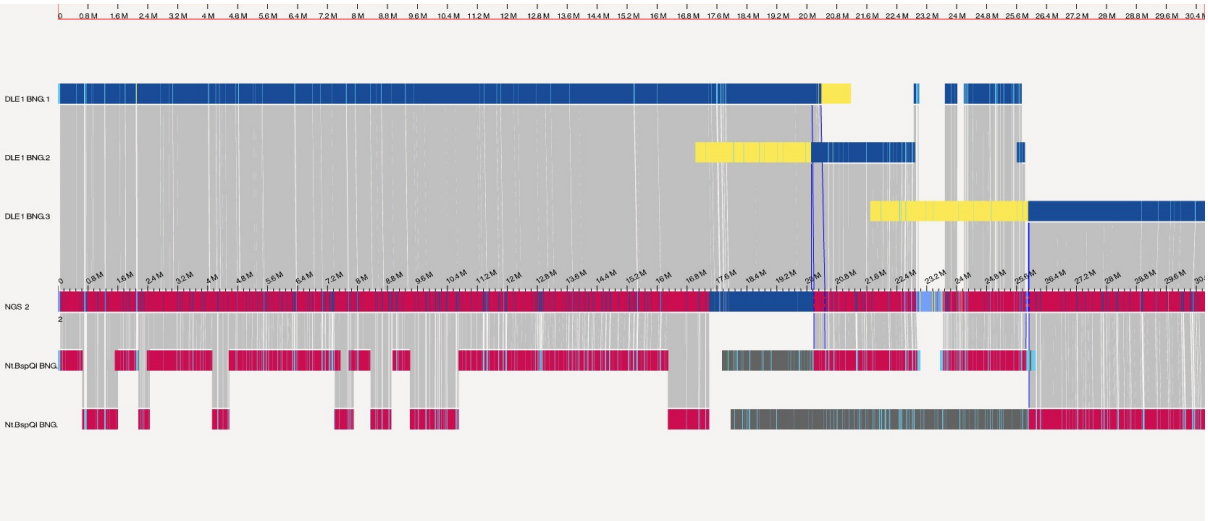


Figure S12. Conflict example detected by the optical maps on the Pore-C scaffolds. The three first lines represent the DLE-1 maps. The fourth line represents one Pore-C scaffold. The two last lines represent the BspQI maps. The vertical lines show the correspondence between the labels of the scaffold and the labels of the maps. We can see regions without any correspondence (the yellow portion of the DLE-1 maps and the grey portions of the BspQI maps) which leads to several breaks of the scaffold.



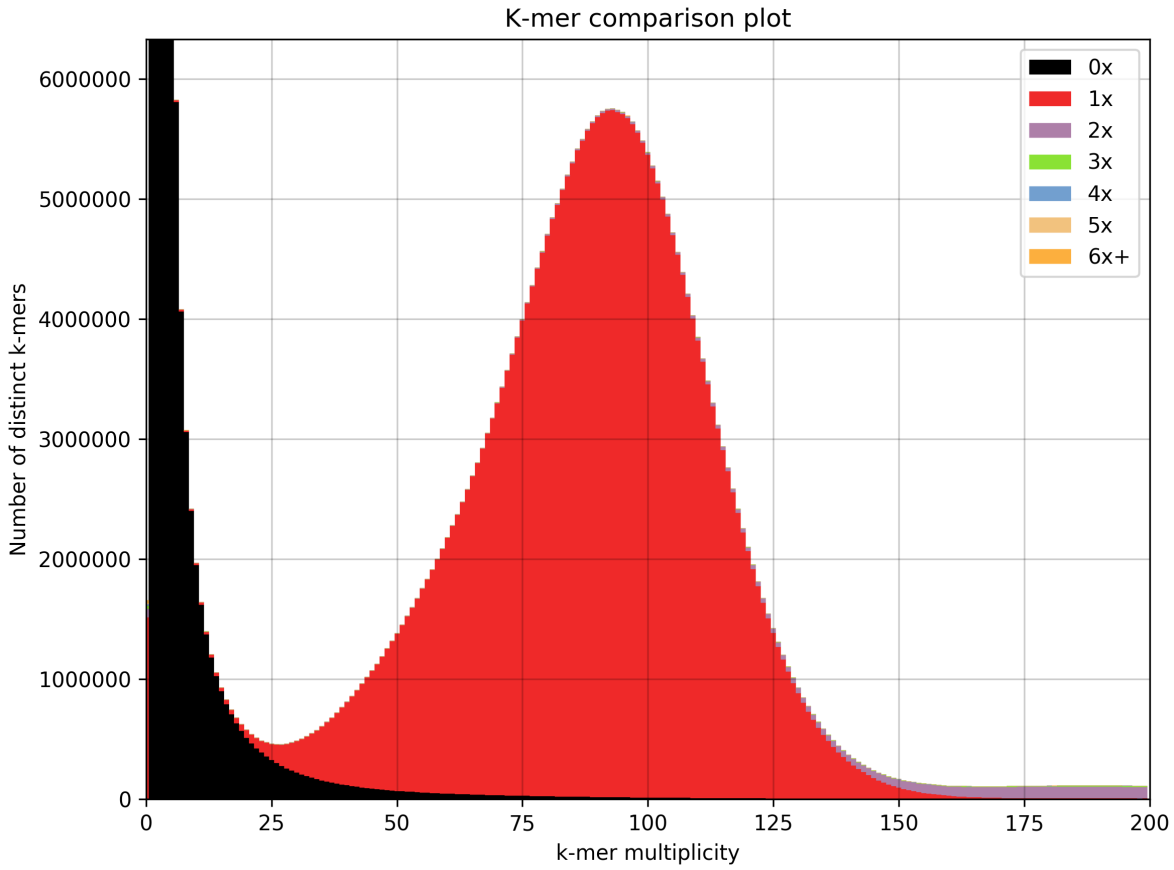
10.5 ANNEXE 5 : "TELOMERE-TO-TELOMERE GAPLESS CHROMOSOMES OF BANANA USING NANOPORE SEQUENCING" BELSER, C., BAURENS, FC., ET AL. *COMMUNICATIONS BIOLOGY*. 2021 SUPPLEMENTARY DATA.

Les Supplementary Data 1 et 2 sont disponibles au téléchargement sur le site de *Communications Biology* à l'adresse suivante : https://www.nature.com/articles/s42003-021-02559-3

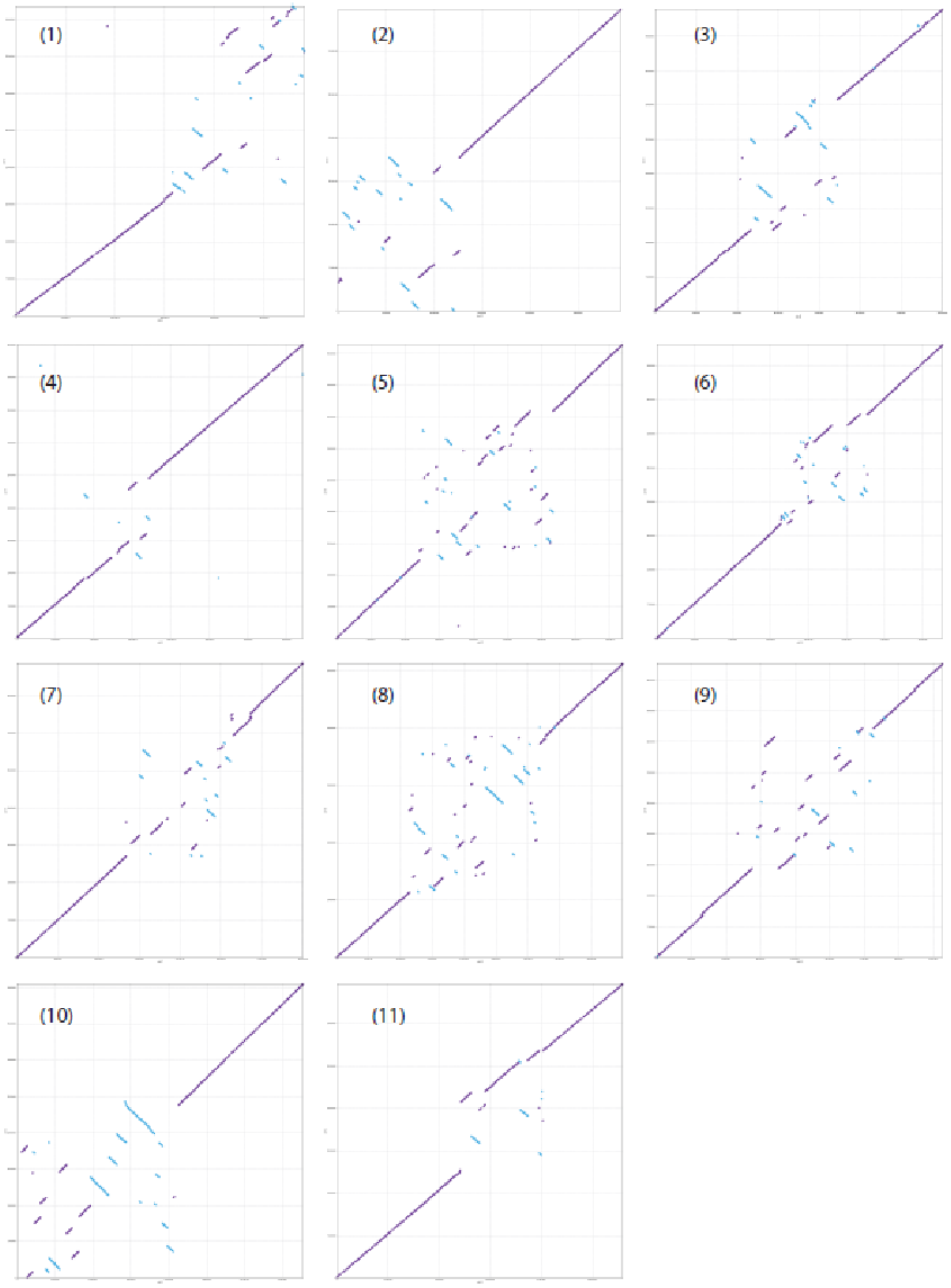
(ou https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-021-02559-3/MediaObjects/42003_2021_2559_MOESM4_ESM.xlsx

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-021-02559-3/MediaObjects/42003_2021_2559_MOESM5_ESM.xlsx)

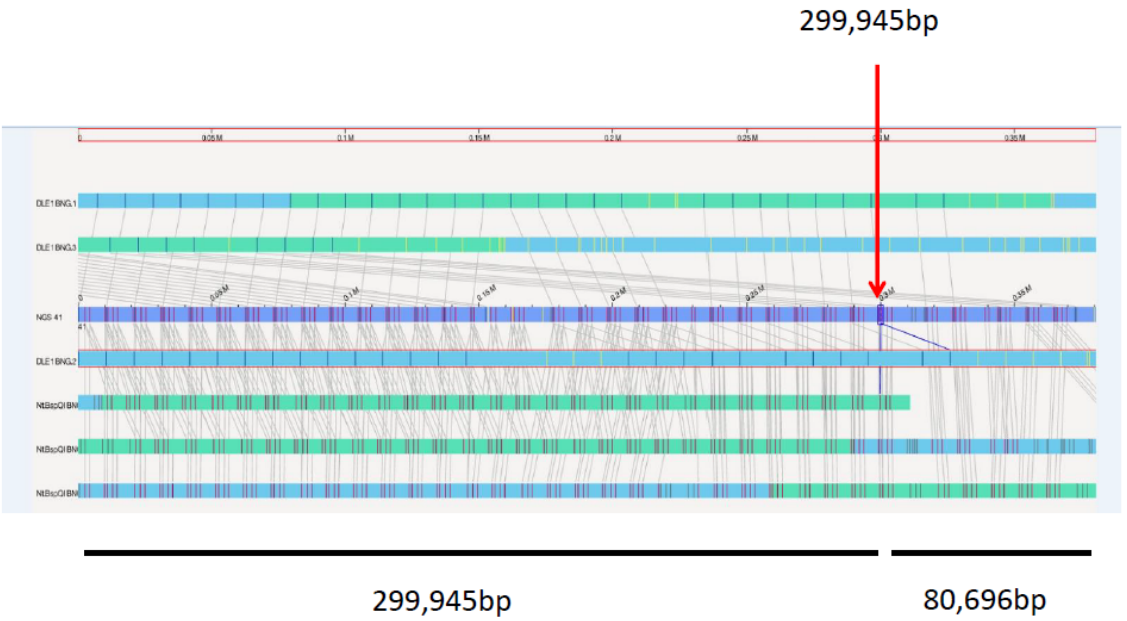
Supplementary Figure 1: **KAT plot of *Musa acuminata* V4 assembly.** K-mer multiplicity in the assembly is represented by colors (black:0, red:1, purple:2, green: 3, blue: 4, orange:5).



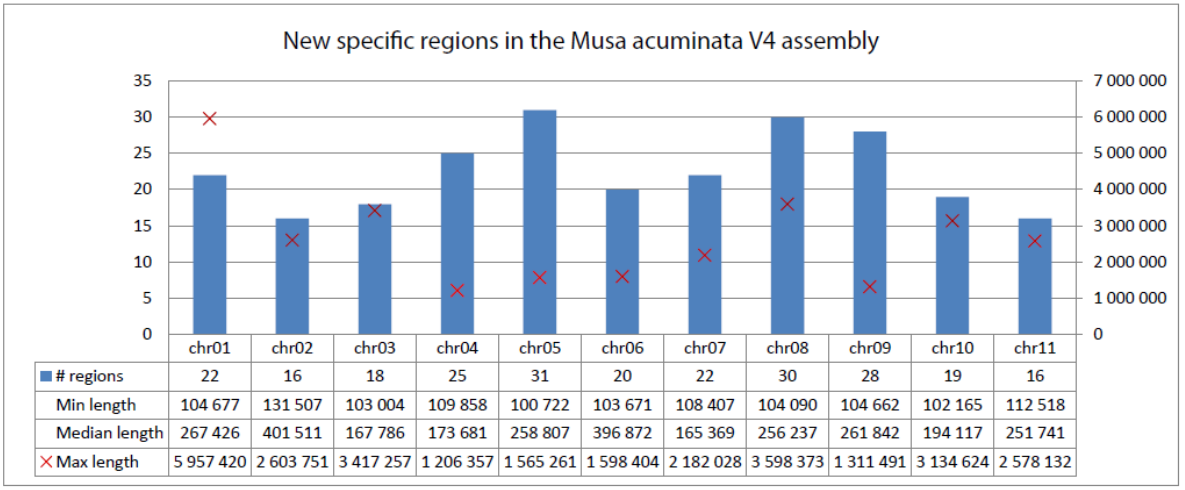
Supplementary Figure 2: Dot plot of each V4 chromosome against its relative V2 chromosome. The x axis represents the chromosome in V2. The y axis represents the chromosome in V4. The number of the chromosome is printed in brackets.



Supplementary Figure 3: Contig flagged as conflictual with the optical maps. The contigs NGS41 (380,641bp) contains in tandem repetitive elements. The contig was split at the position 299,945b in two contigs of 299,945 et 80,696bp.

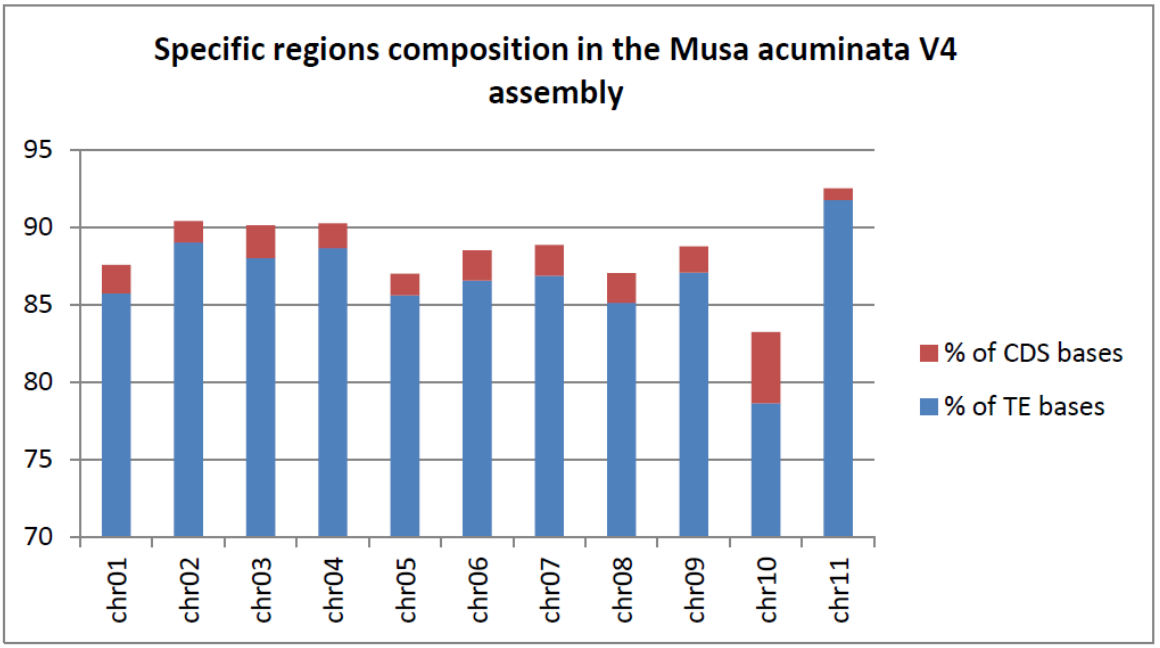


Supplementary Figure 4: Characterization of specific regions of the *Musa acuminata* V4 assembly. Blue bars represent the number of new regions for each chromosome and red crosses represent the maximum size on each chromosome.

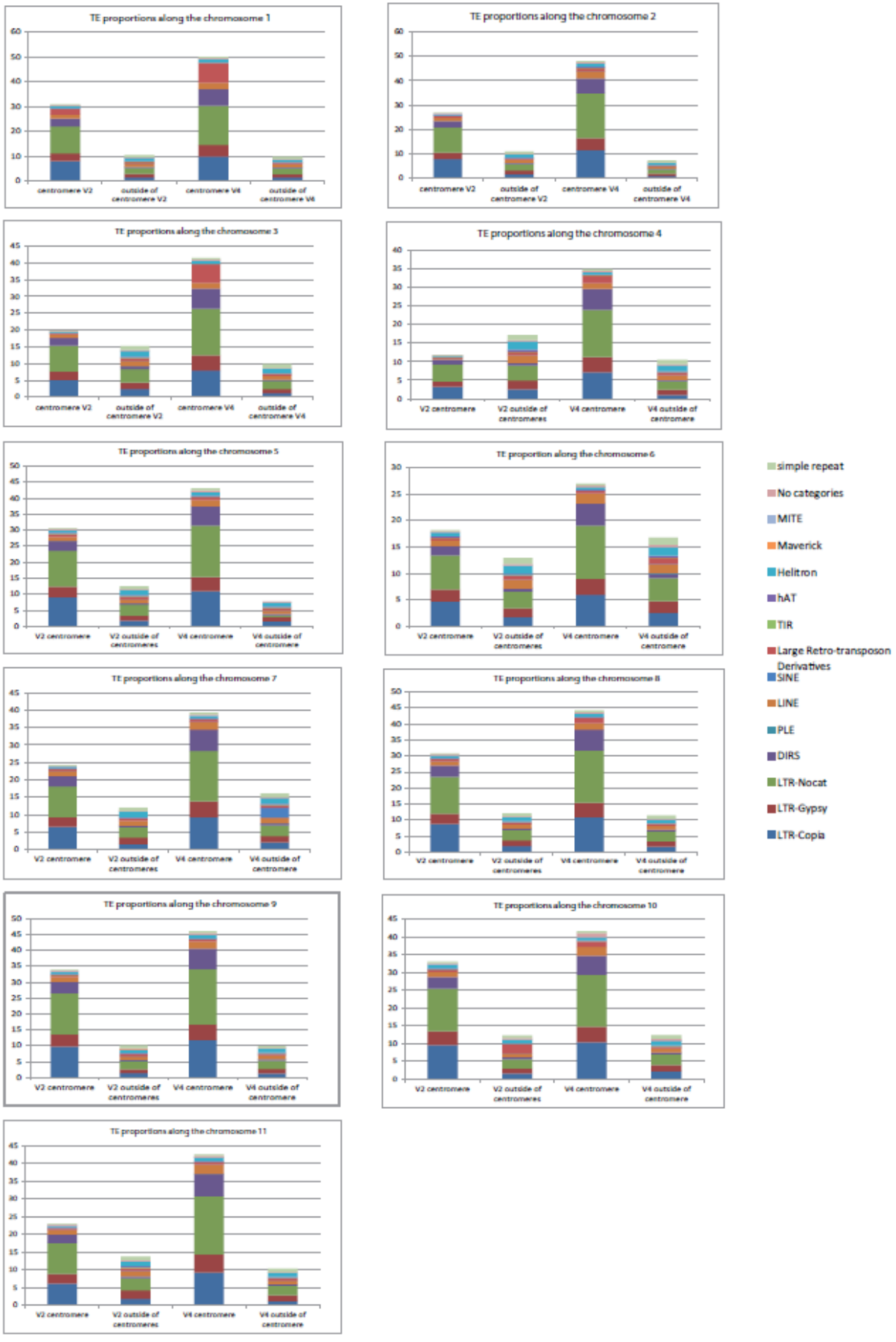


	chr01	chr02	chr03	chr04	chr05	chr06	chr07	chr08	chr09	chr10	chr11
# regions	22	16	18	25	31	20	22	30	28	19	16
Min length	104 677	131 507	103 004	109 858	100 722	103 671	108 407	104 090	104 662	102 165	112 518
Median length	267 426	401 511	167 786	173 681	258 807	396 872	165 369	256 237	261 842	194 117	251 741
Max length	5 957 420	2 603 751	3 417 257	1 206 357	1 565 261	1 598 404	2 182 028	3 598 373	1 311 491	3 134 624	2 578 132

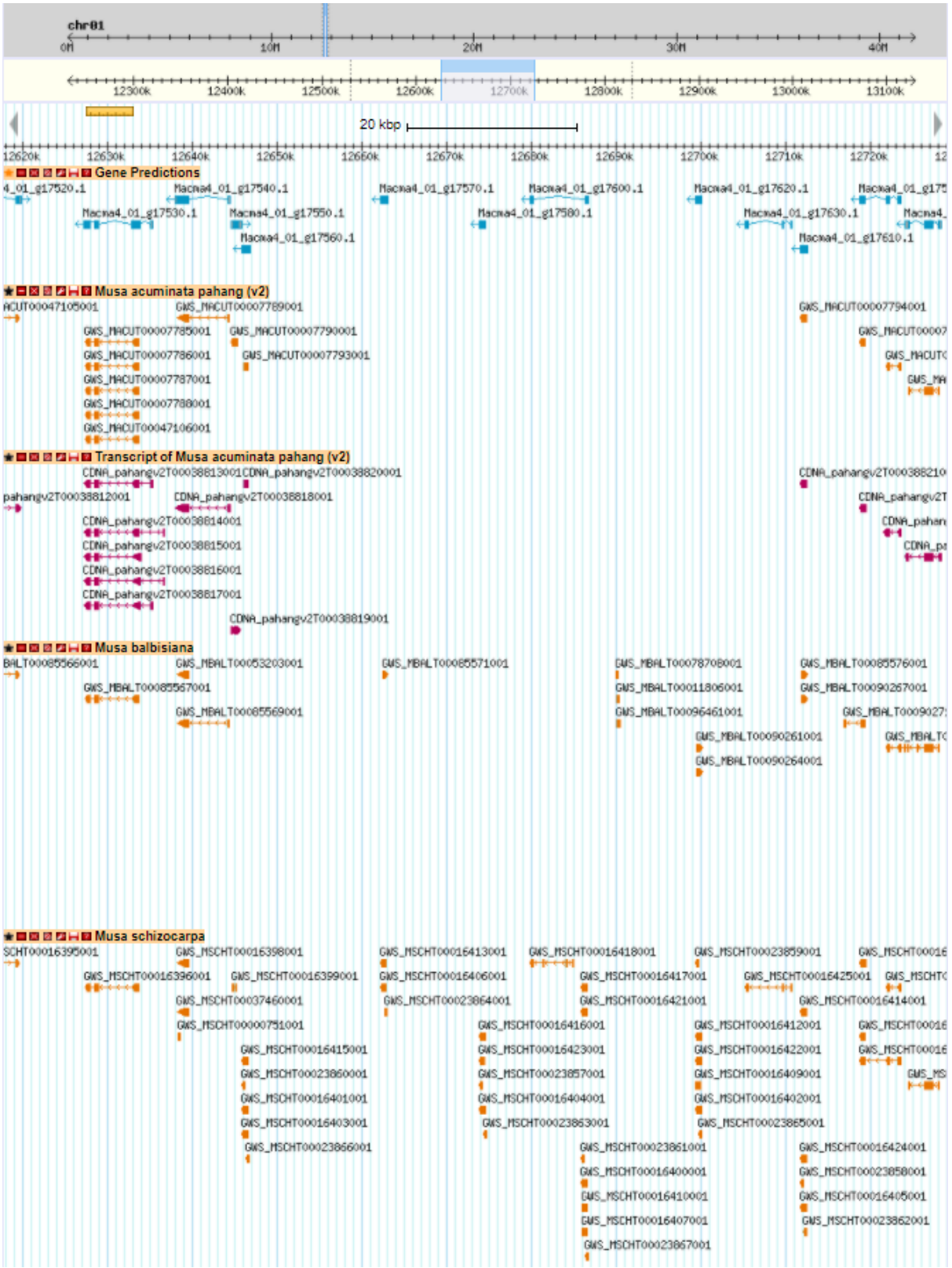
Supplementary Figure 5: Composition of specific regions of the *Musa acuminata* V4 assembly. Proportion of TEs and CDS are in blue and red respectively.



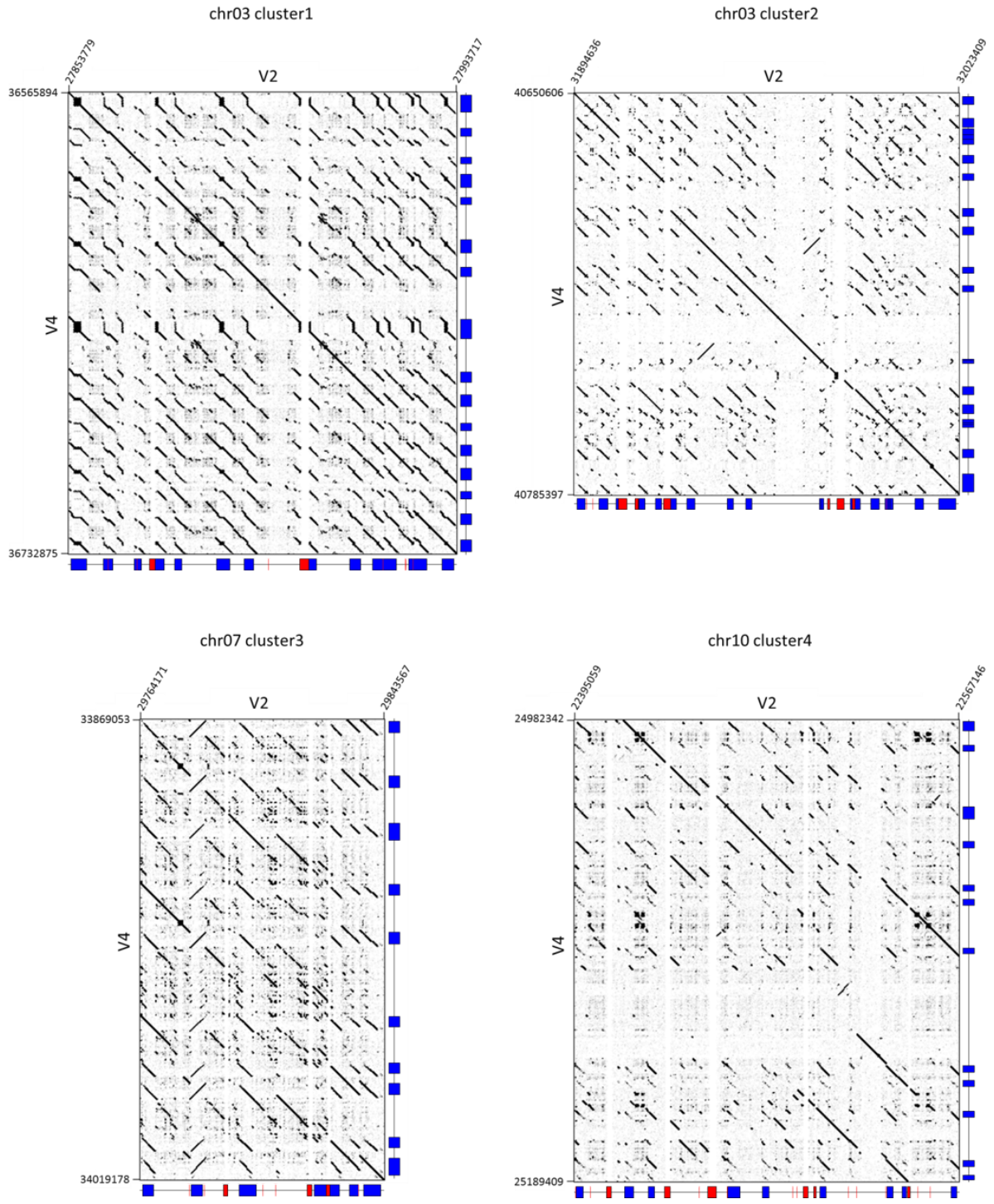
Supplementary Figure 6: Distribution of the repeats along the V4 chromosomes. Proportion of each TE category on each chromosome.



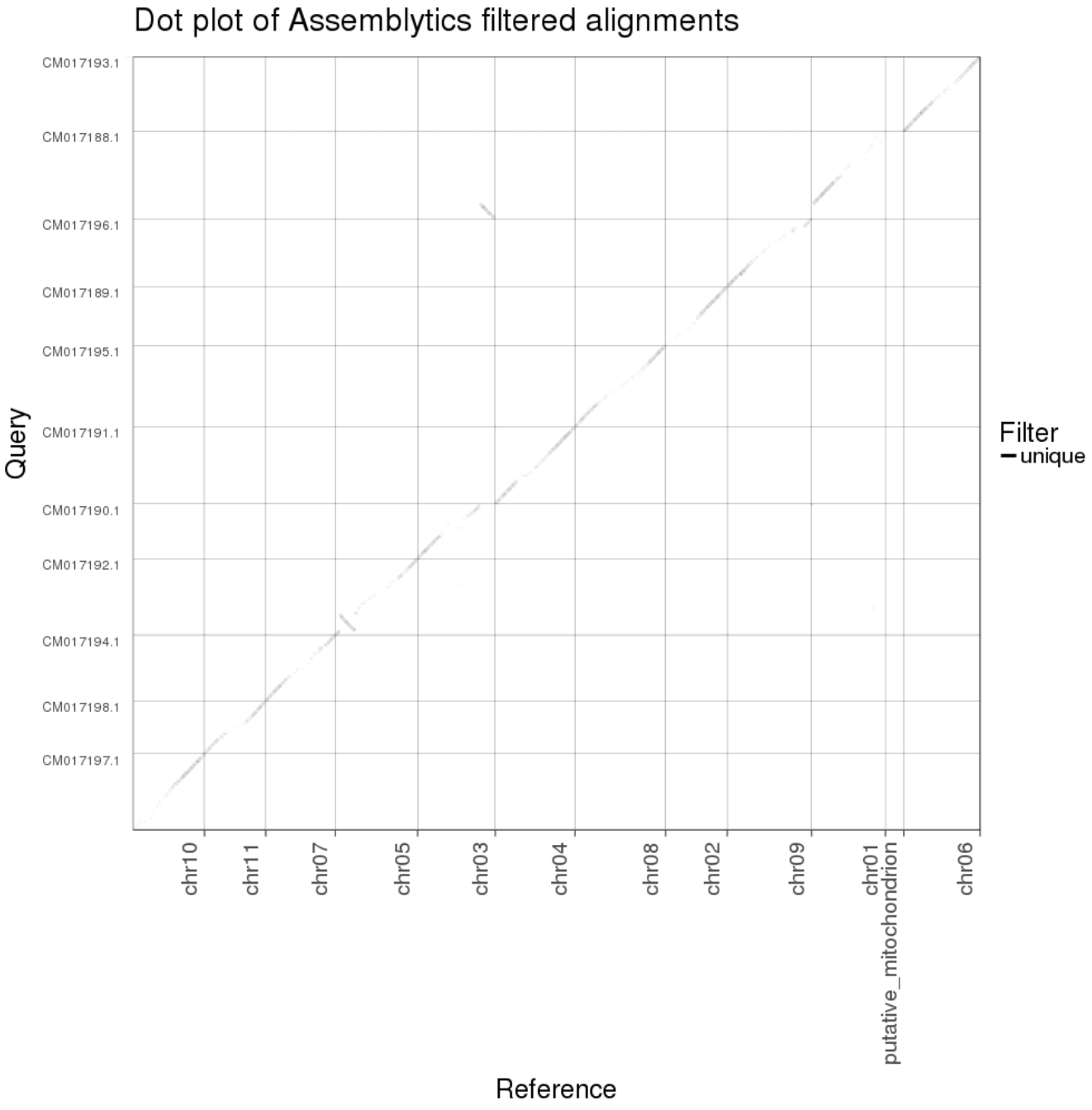
Supplementary Figure 7: Screenshot of the genome browser focusing on a *Musa acuminata* V4 chromosome 1 region. These new annotated genes in the center of track 1 (Gene Predictions) are included in TDGs cluster and were absent from the *Musa acuminata* V2 annotation.



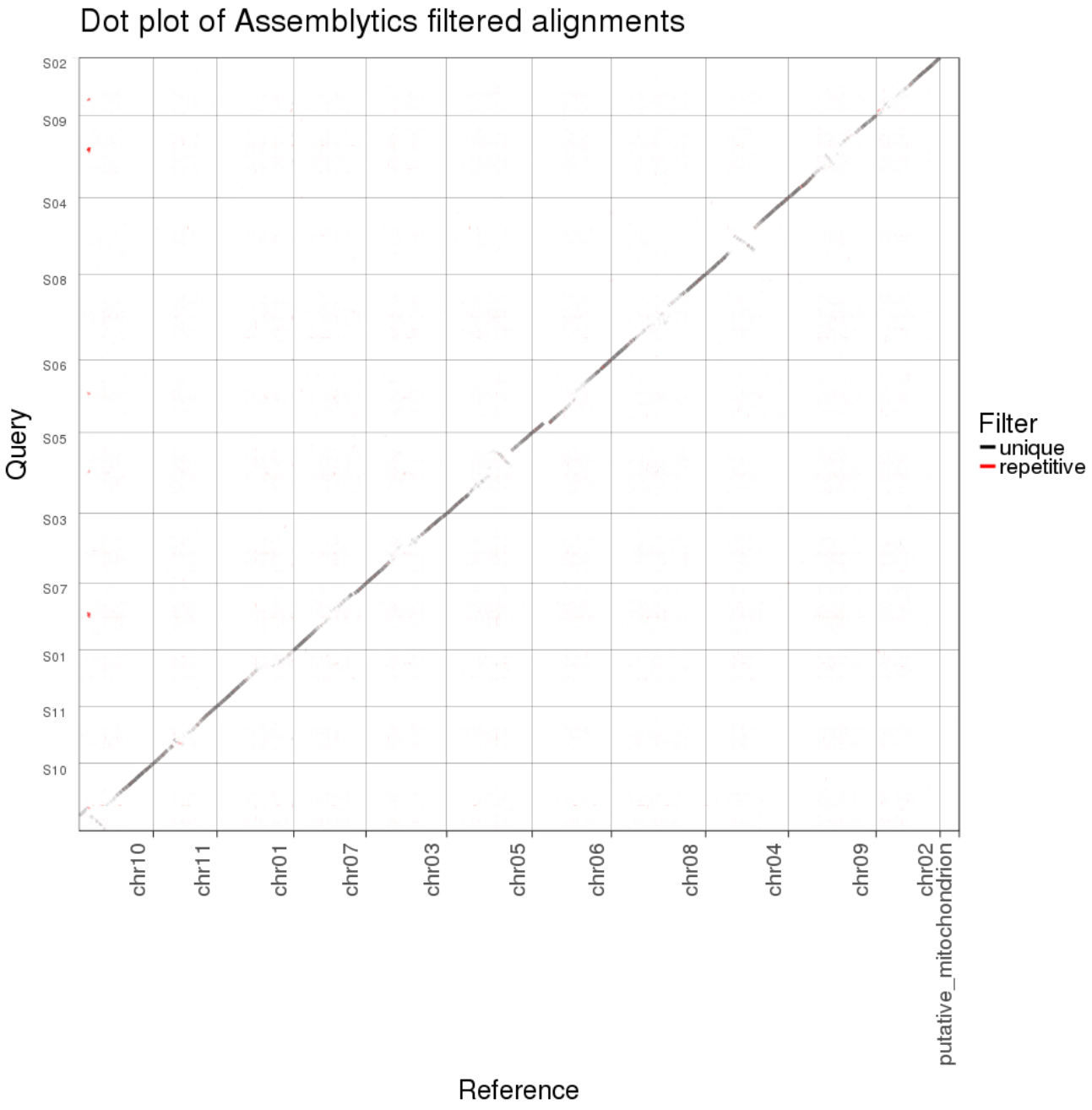
Supplementary Figure 8: Comparison of the structure of NLR loci clusters between DH-Pahang V2 and V4 assemblies. The four panels represent dot plots of NLR loci clusters on chromosomes 3, 7 and 10 as indicated on top of each panel. The predicted NLR loci for each version are represented on the right side and at the bottom of the dot plots by blue boxes. Red boxes represent regions bearing undetermined nucleotides. Region coordinates are also indicated.



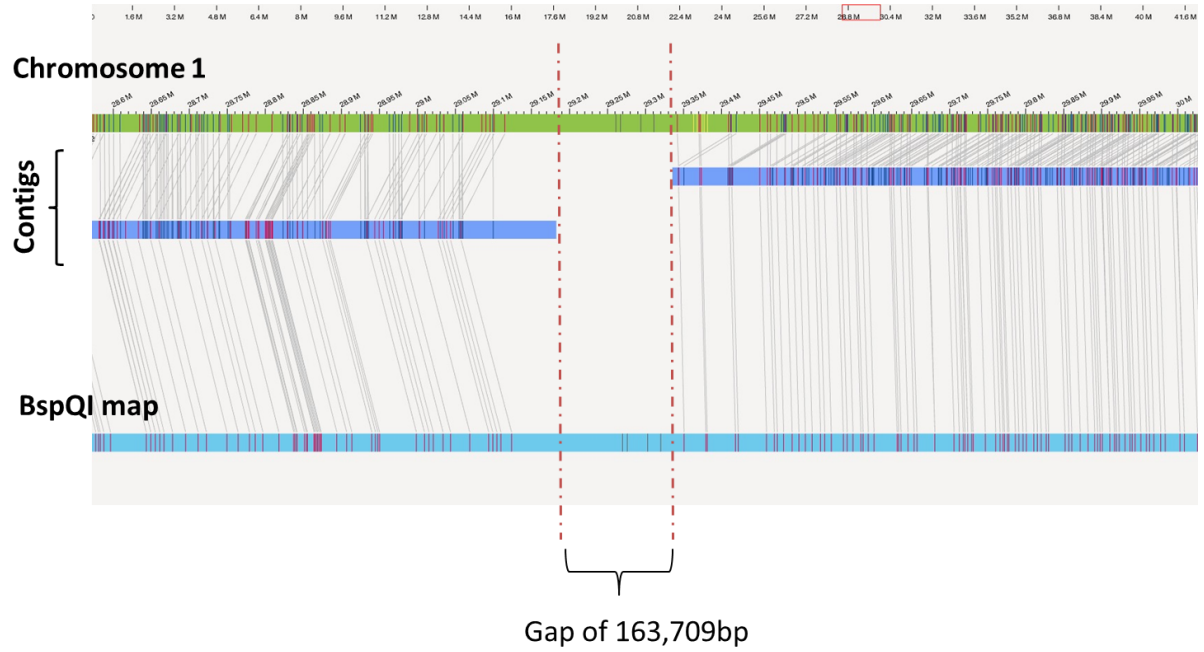
Supplementary Figure 9: Dot plot of *Musa balbisiana* assembly against *Musa acuminata* V4 assembly. *Musa balbisiana* and *acuminata* chromosomes are on the y-axis and x_axis respectively.



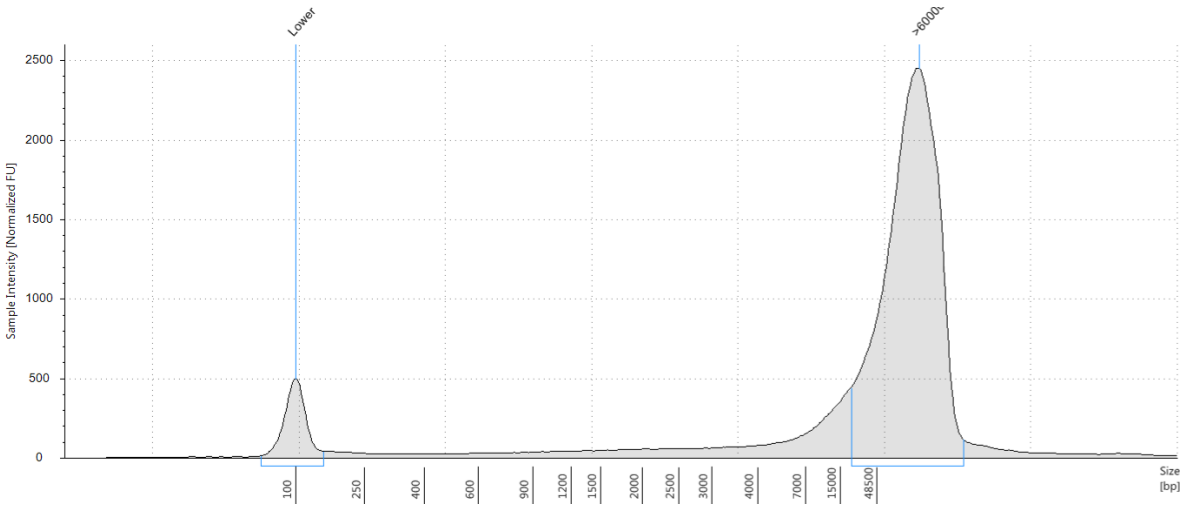
Supplementary Figure 10: Dot plot of *Musa schizocarpa* assembly against *Musa acuminata* V4 assembly. *Musa schizocarpa* and *acuminata* chromosomes are on the y-axis and x_axis respectively.

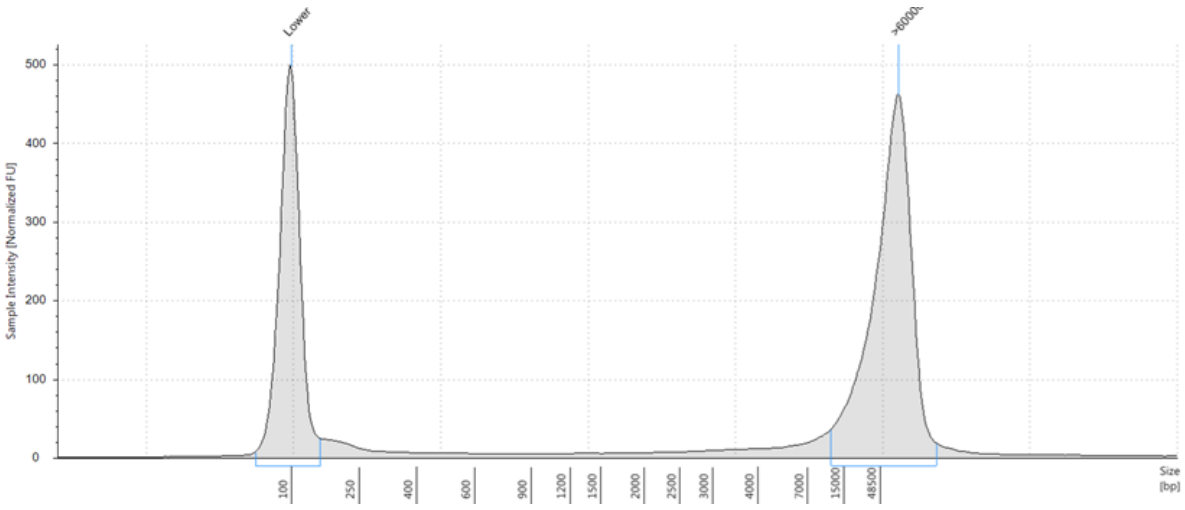


Supplementary Figure 11: Remaining gaps after negative gap resolution. Example of a remaining sized gap in chromosome 1, located between the position 29,171,567 and 29,335,274. The gap was sized thanks to the BspQI map.

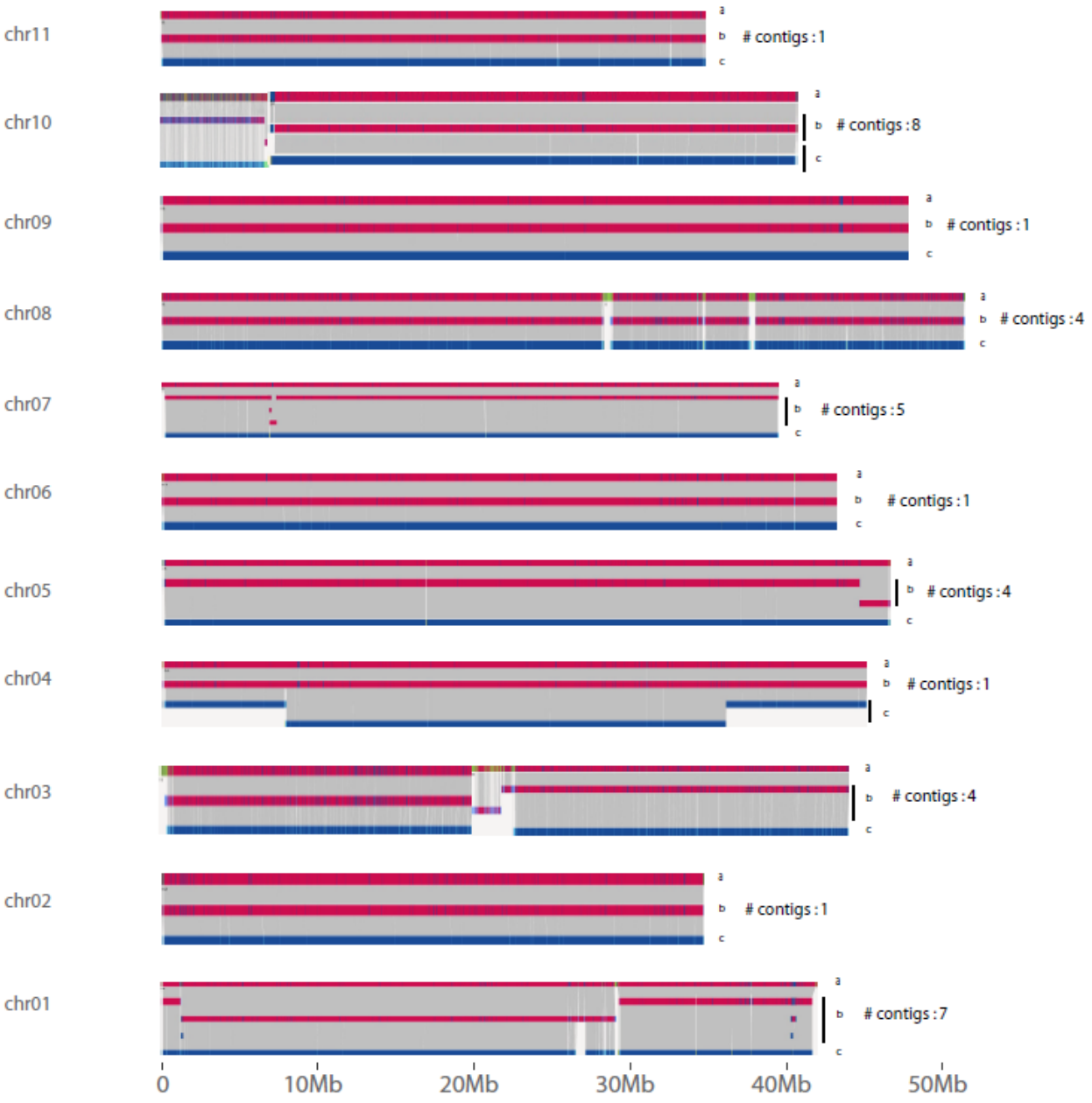


Supplementary Figure 12: gDNA extraction of DH-Pahang. DNA quality was checked on a 2200 TapeStation automated electrophoresis system (Agilent, CA, USA). Ninety seven % of the DNA fragments have a length >50Kb. (a) before removal of small DNA fragments with Short Read Eliminator XL (Circulomics, MD, USA) (b) after removal of small DNA fragments

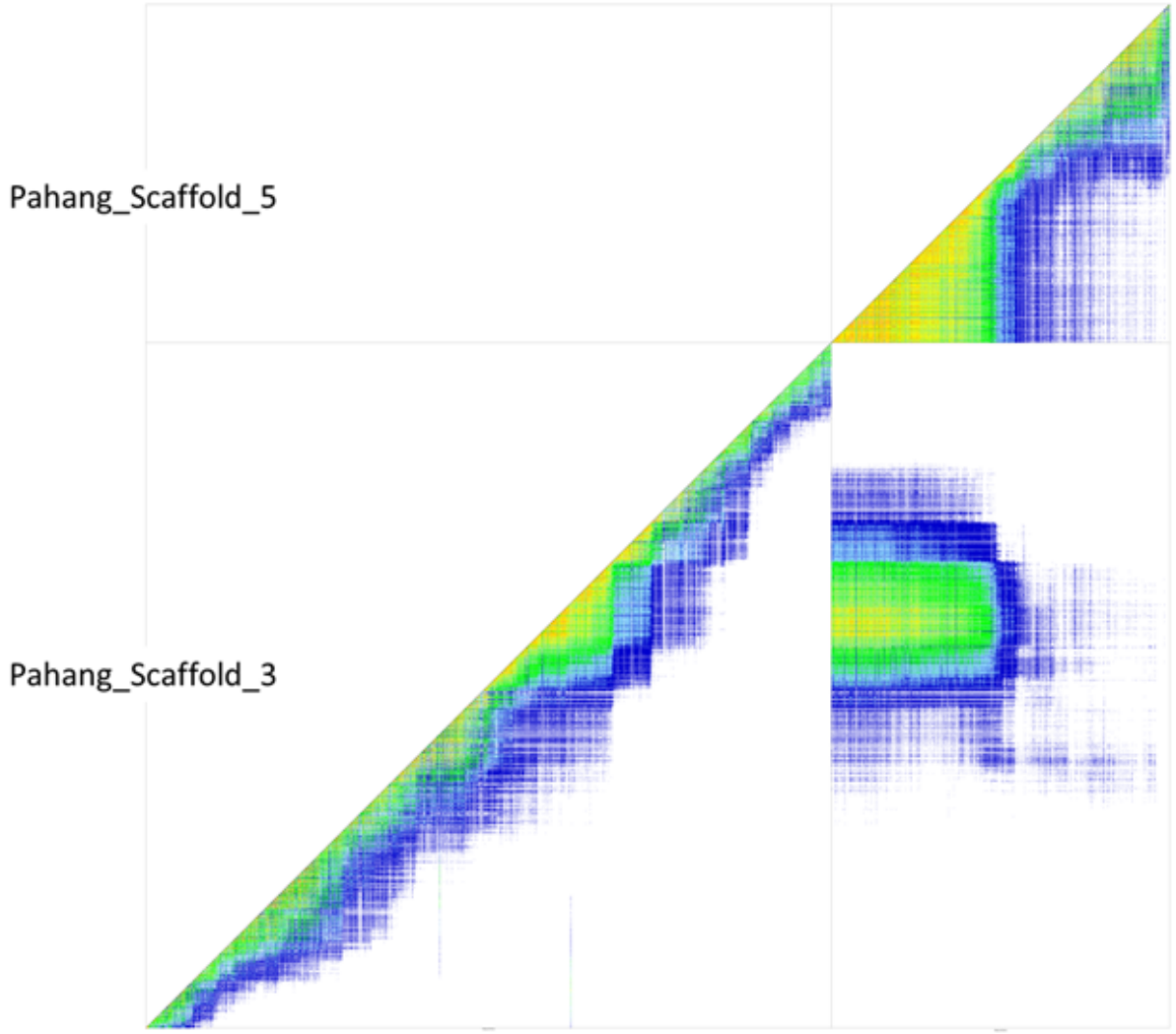




Supplementary Figure 13: Overview of the nanopore contigs in the V4 assembly. Alignment of the two optical maps (a, red) DLE-1 and (b, red) BspQI with the nanopore contigs (c, blue). The vertical grey lines between the contigs and the optical maps represent the matches between enzymatic cuts from the map and the DNA sequences. The number of contigs of each chromosome is indicated.



Supplementary Figure 14: Dot plot showing marker linkage along ordered scaffolds of linkage group 01-04. This figure showed marker linkage of linkage group 01-04 which contained markers from chromosome 01 and 04. Because of chromosomal co-segregation due to reciprocal translocation between chromosome 01 and 04, markers from these chromosomes are linked. This resulted by the linkage of markers from a region of scaffold_3 to a region of scaffold_5. Interpretation of the figure suggested that in fact scaffold_3 corresponded to one chromosome and scaffold_5 corresponded to another chromosome.



Supplementary Table 1: Contigs details before and after negative gap resolution. Gaps of 100bp are gaps of unknown length generated by the anchoring of contigs using the genetic map. Gaps of 13bp are gaps of unknown size generated by the BioNano pipeline.

DH-Pahang chromosome	# corresponding NECAT contigs	# hybrid scaffolds	# contigs after negative gap resolution	Gaps length (bp)
chr01	7	1	5	163,709 - 13 - 37,440 - 13
chr02	1	1	1	/
chr03	4	2	2	100
chr04	1	1	1	/
chr05	4	1	4	53,161 - 30,868 - 32,283
chr06	1	1	1	/
chr07	5	1	4	13 - 13 - 13
chr08	4	2	4	100 - 46,534 - 324,396
chr09	1	1	1	/
chr10	8	2	2	100
chr11	1	1	1	/

Supplementary Table 2: Assemblies quality scores. Comparison of assembly quality scores of the V1 and V4 assemblies.

Assembly	V1	V4
K-mer completion	95.7068	98.1327
qscore	49.1651	38.7878
qscore on shared region	V1 : 50.1293 V4 : 45.9284	

Supplementary Table 3: TE classes proportions in *Musa acuminata* (V2 and V4), *Musa schizocarpa* and *Musa balbisiana*. Transposable elements are classified according to their class.

		Musa acuminata V4 (%)	*Musa acuminata* V2 (%)	*Musa schizocarpa* (%)	*Musa balbisiana* (%)
Class I (retrotransposons)					
LTR	Copia	11,079	8,965	13,766	12,432
	Gypsy	5,986	4,668	5,848	5,137
	no cat	17,788	12,743	19,767	16,547
DIRS	RYX	6,323	3,252	6,094	4,157
PLE	Penelope	0,003	0,003	0,003	0,004
LINE	RIL/RIX	3,492	2,741	3,478	2,941
SINE	RSX	0,005	0,006	0,009	0,004
Large Retro-transposon Derivatives	RXX	2,678	1,347	2,590	1,597
Class II (DNA transposons)-Subclass 1					
TIR	DTX	0,163	0,172	0,180	0,185
hAT	DTA	0,506	0,538	0,525	0,637
Class II (DNA transposons)-Subclass 2					
Helitron	DHH/DHX	2,292	2,175	1,737	2,152
Maverick	DMX	0,005	0,006	0,007	0,006
MITE (miniature inverted repeat transposable elements)	DXX	0,029	0,028	0,023	0,027
No categories		0,722	0,549	1,097	0,762
simple repeat		1,549	1,190	1,221	2,767
	TOTAL	52,623	38,383	56,345	49,353

Supplementary Table 4: Comparison of 5S ribosomal gene clusters in V2 and V4 Musa assemblies. Number of bases in each assembly and on each chromosome tagged as 5S ribosomal genes.

	Base pair covered in V2 assembly (predicted genes)	Base pair covered in V4 assembly (predicted genes)
chr01	0	59,949 (1,882)
chr02	0	0
chr03	194 (3)	179,282 (4,645)
chr04	0	413 (10)
chr05	529 (21)	640 (19)
chr06	68 (1)	68 (1)
chr07	419 (7)	216 (3)
chr08	4,104 (38)	127,057 (1,135)
chr09	547 (14)	80 (1)
chr10	116 (6)	0
chr11	0	0
Un	1,605 (40)	-
Total	7,582 (130)	367,705 (7,696)

Supplementary Table 5: Proportion of new annotated genes in each *Musa acuminata* chromosome of the V4 assembly. Number of annotated genes (first column), duplicated genes (second column), new genes in the V4 assembly (third column) and new genes that are tandemly duplicated (last column).

	Nb of annotated genes	Nb of tandemly duplicated genes	Nb of new annotated genes	Nb of new tandemly duplicated genes
chr01	2,757	397 (14.40%)	369 (13.38%)	173 (46,88%)
chr02	2,680	300 (11.19%)	232 (8.66%)	88 (37,93%)
chr03	3,533	314 (8.89%)	235 (6.65%)	85 (36,17%)
chr04	4,254	343 (8.06%)	258 (6.06%)	66 (25,58%)
chr05	3,345	263 (7.86%)	251 (7.50%)	73 (29,08%)
chr06	4,076	349 (8.56%)	283 (6.94%)	96 (33,92%)
chr07	3,080	324 (10.52%)	190 (6.17%)	83 (43,68%)
chr08	3,604	299 (8.30%)	221 (6.13%)	67 (30,32%)
chr09	3,342	342 (10.23%)	243 (7.27%)	84 (34,57%)
chr10	3,552	527 (14.84%)	510 (14.36%)	250 (49,02%)
chr11	2,612	175 (6.70%)	171 (6.55%)	42 (24,56%)
Total	36,835	3,633 (9.86%)	2,963 (8.04%)	1,137 (38.37%)

Supplementary Table 6: Comparison of NLR clusters between the V2 and V4 assemblies (based on NLR-Annotator predictions). The four NLR clusters are characterized (chromosome, start and end position, size, number of unknown bases, number of genes) in the V2 (four first lines) and V4 assemblies (fout last lines).

Assembly version	Chromosome	Cluster start coordinate	Cluster end coordinate	Size (bp)	Nb. Undetermined nucleotides (N)	Nb. NLR loci detected
V2	chr03	27,854,779	27,992,717	137,938	5,255	14
V2	chr03	31,895,636	32,022,409	126,773	10,472	17
V2	chr07	29,765,171	29,842,567	77,396	4,269	6
V2	chr10	22,396,059	22,566,146	170,087	13,045	9
V4	chr03	36,566,894	36,731,875	164,981	0	16
V4	chr03	40,651,606	40,784,397	132,791	0	18
V4	chr07	33,870,053	34,018,178	148,125	0	10
V4	chr10	24,960,703	25,188,409	227,706	0	13

Supplementary Table 7: Statistics of the ONT datasets. Sequencing metrics of the nanopore long-reads: complete dataset, longest reads only and highest quality reads.

	DH-Pahang		
	Raw reads PromethION	Longest reads	Filtlong reads
Cumulative size	92,749,841,432	13,500,017,777	13,500,035,543
# of reads	5,185,398	175,937	273,413
Coverage	206x	30x	30x
N50 (bp)	31,640	74,681	52,825

Supplementary Table 8: DH-Pahang ONT assembly statistics. Metrics of the assemblies generated using different assemblers and the three nanopore dataset (all reads, longest and highest score).

	SMARTDeNovo			Redbean		
Subset of reads used	All reads	Longest	Filtlong	All reads	Longest	Filtlong
Cumulative size	482,349,453	465,668,591	431,317,546	2,021,646,269	802,627,210	606,906,240
# contigs	121	183	783	43,296	13,785	10,307
N50 (L50)	19,507,988 (9)	5,667,738 (22)	977,146 (111)	75,016 (6,675)	93,433 (1,502)	92,564 (989)
N90 (L90)	2,709,408 (28)	1,176,439 (92)	230,199 (490)	21,284 (26,326)	25,463 (7,959)	25,353 (6,103)

	Flye			NECAT
Subset of reads used	All reads	Longest	Filtlong	All reads
Cumulative size	470,287,666	471,731,899	439,685,078	485,698,222
# contigs	449	318	953	175
N50 (L50)	15,404,016 (12)	12,916,122 (13)	1,521,384 (73)	32,031,704 (7)
N90 (L90)	2,621,850 (40)	3,315,873 (37)	298,601 (333)	5,650,309 (17)

Supplementary Table 9: DH-Pahang bionano dataset. Statistics of the molecule used to generate the two optical maps.

	DLE-1 Molecule statistics	BspQI Molecule statistics
Total number of molecules	2,151,406	2,897,832
Total length (Mbp)	209,178.556	398,318,111
Average length (kbp)	97.229	137.109
Molecule N50 (kbp)	129.750	220.875
Label density (/100kb)	13.709	10.502
Number of Flow cell	2	1

Supplementary Table 10: DH-Pahang bionano genome map. Statistics of the two optical maps obtained with the DLE and BspQI enzymes.

	DLE-1 genome map	BspQI genome map
Genome map number	24	71
Total Genome Map Length (Mbp)	469.764	474.019
Genome Map N50 (Mbp)	35.022	16.002

Supplementary Table 11: DH-Pahang hybrid scaffolding and polishing. Statistics of the raw nanopore assembly (first column), the assembly obtained by combining nanopore contigs and optical maps (second, third and fourth column), and the final assembly after the resolution of negative gaps (last column).

	nanopore contigs polished	Hybrid scaffolds	contigs not scaffolded	final hybrid scaffolds (hybrid scaffolds + contigs not scaffolded)	scaffolds after negative gap resolution and polishing
number	124	16	80	96	97
N50 (L50)	32,091,274 (7)	39,508,388 (6)	249,463 (18)	39,508,388 (6)	39,373,400 (6)
N90 (L90)	5,668,018 (17)	21,536,064 (12)	87,718 (59)	21,536,064 (12)	21,536,112 (12)
maxSize	47,719,325	47,719,325	673,878	47,719,325	47,719,527
Assembly size	485,318,484	471,709,278	14,435,609	486,144,887	484,747,212
% of N	0%	0.18%	0%	0.17%	0.14%

Supplementary Table 12: **Marker number per scaffold and linkage group.** Number of markers aligned on each scaffolds (lines) for each linkage groups (columns).

scaffold	LG01-04	LG02	LG03	LG05	LG06	LG07	LG08	LG09	LG10	LG11
Pahang_Scaffold_1	0	0	0	1	0	1	0	2089	0	0
Pahang_Scaffold_2	0	0	0	1874	0	1	0	0	1	0
Pahang_Scaffold_3	2191	1	0	0	0	0	0	1	0	0
Pahang_Scaffold_4	0	0	1	0	2190	0	0	1	0	0
Pahang_Scaffold_5	1082	0	2	3	0	0	2	1	0	0
Pahang_Scaffold_6	0	0	0	1	0	1687	0	0	1	1
Pahang_Scaffold_7	0	0	0	0	1	0	0	0	1784	0
Pahang_Scaffold_8	0	1483	0	0	0	0	0	0	0	0
Pahang_Scaffold_9	0	0	0	1	0	0	1	1	0	1500
Pahang_Scaffold_10	0	0	0	1	0	0	1283	0	0	0
Pahang_Scaffold_11	0	0	1027	1	0	0	0	0	0	0
Pahang_Scaffold_12	0	0	0	0	0	0	1276	0	0	0
Pahang_Scaffold_13	0	0	850	0	0	0	1	0	1	0
Pahang_Scaffold_14	0	0	0	1	0	0	0	0	173	0
Pahang_Scaffold_59	1	0	0	0	0	0	0	0	0	0