THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Génétique et Amélioration des plantes

École Doctorale GAIA
Spécialité BIDAP - Biologie Intégrative Diversité et Amélioration des plantes

Unité de recherche DIADE - UM IRD - Diversité Adaptation et Développement des Plantes

Étude du paysage des éléments transposables sous forme d'ADN circulaire extrachromosomique et dans l'assemblage des génomes de plantes à l'aide du séquençage en lectures longues

Présentée par Panpan ZHANG
Le 30 mai 2022
Sous la direction de Alain GHESQUIERE

Devant le jury composé de :

Anne Roulin – Professeur, Université de Zurich	Rapporteur
Todd Blevins - Chargé de Recherche CNRS, IBMP, Strasbourg	Rapporteur
Magnus Nordborg – Directeur de recherche, Gregor Mendel Institute, Vienne	Examinateur
Romain Guyot – Directeur de recherche IRD, DIADE Montpellier	Examinateur
Olivier Panaud – Professeur, LGDP, Université de Perpignan	Examinateur (président du jury)
Alain Ghesquière - Directeur de recherche IRD, DIADE, Montpellier	Directeur de thèse
Marie Mirouze – Chargé de recherche IRD, LGDP, Université de Perpignan	Encadrante/Invitée
Nathalie Picault - Maitre de Conférences, LGDP, Université de Perpignan	Invitée

UNIVERSITÉ DE MONTPELLIER

À mes parents et mon fiancé

Restez affamés, restez fous

--Steve Jobs

Résumé de la thèse en 5 pages

Les éléments transposables (TE), qui désignent des éléments génétiques mobiles, sont des composants majeurs des génomes eucaryotes, représentant de ~3% chez la levure *S. cerevisae* (Kim et al., 1998) à 45% chez l'homme (Lander et al., 2001) et plus de 85% chez certaines espèces végétales, comme le blé (Wicker et al., 2018). Suivant leurs modes de transposition, les TE sont classés en rétrotransposons et en transposons à ADN (Feschotte et al., 2002 ; Wicker et al., 2007). Les TE actifs initient la formation de particules de type viral et la réverse transcription qui s'ensuit aboutit à la synthèse de l'ADN extrachromosomique (ADNe). L'ADN extrachromosomique sous forme double brin s'insère dans de nouveaux loci génomiques grâce à l'intégrase ou forme de l'ADN circulaire extrachromosomique (ADNecc) (Lanciano et al., 2017). L'ADNecc a été observé chez de nombreuses espèces eucaryotes comme la levure, la drosophile, les nématodes, les plantes et les humains (Hotta et Bassel, 1965; Hirochika et Otsuki, 1995; Sinclair et Guarente, 1997; Cohen et Méchali, 2002; Cohen et al., 2006; Kumar et al., 2017).

L'ADNecc a été découvert depuis plusieurs décennies, mais au cours des dernières années, grâce au séquençage à haut débit, son étude a connu un véritable engouement car il joue un rôle important dans l'évolution des cellules cancéreuses (Verhaak et al., 2019). Dans ces cellules, il contribue à l'évolution adaptative en favorisant les variations rapides du nombre de copies (Kim et al., 2020). Chez les plantes, l'ADNecc a été caractérisé pour la première fois par un protocole qui séquence sélectivement l'ADN circulaire enrichi suite à la digestion de l'ADN linéaire, à savoir le mobilome-seq (Lanciano et al., 2017). Cette technique, basée sur le séquençage à lectures courtes a révélé que les ADNecc proviennent notamment des TE actifs dans la plante. Avec l'arrivée des technologies de séquençage en lectures longues proposées par deux plateformes fondamentalement différentes : Pacific Biosciences (PacBio) et Oxford Nanopore Technologies (ONT), le mobilome-seq ou eccDNA-seq permet de capturer la structure des ADNecc en couvrant leur longueur totale en une seule lecture (Koche et al., 2020). Cependant, au début de ma thèse, il n'existait pas d'outil bioinformatique dédié à la détection des ADNecc à partir du séquençage en lectures longues (Figure Résumé).

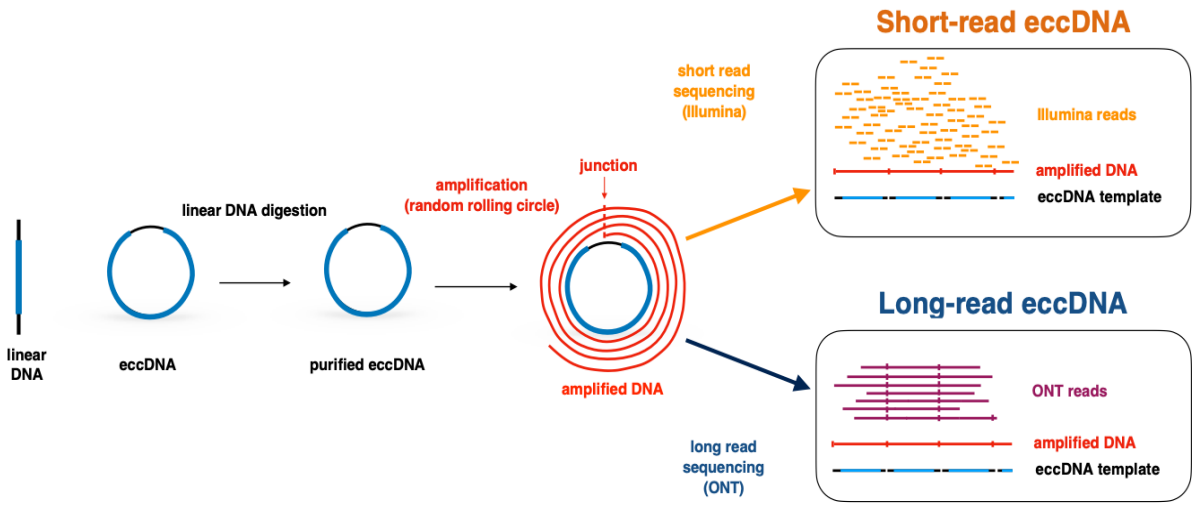


Figure Résumé. Méthode de l'eccDNA-seq ou mobilome-seq. Les eccDNA provenant de l'eccDNA-seq à lecture courte (orange) et de l'eccDNA-seq à lecture longue (bleu) sont indiqués.

Premier objectif : outil pour caractériser le paysage des ET dans les ADNecc

Au cours de ma thèse, le premier objectif était de caractériser le contenu en TE dans les ADNecc. La thèse présentera tout d'abord les propriétés des TE, les mécanismes de transposition et de « silencing » des TE, et s'étendra aux modèles proposés pour les captures de gènes par les TE. En particulier, les résultats de l'état de l'art en matière d'ADNecc tels que l'ADNecc correspondants aux TE actifs chez les plantes, ou aux oncogènes dans les cellules cancéreuses seront discutés. Mes résultats concernant le développement d'une méthode bioinformatique de détection d'ADNecc à partir de données eccDNA-seq seront présentés, ainsi que leur application chez *Arabidopsis thaliana* avec un stress thermique et chez le blé tendre *Triticum aestivum*, illustrant la puissance de calcul, la sensibilité et la précision de l'outil développé. Grâce à cet outil, l'ADNecc peut être détecté de manière robuste et précise à partir de séquençage en lecture courtes mais aussi longues, comblant ainsi les lacunes dans ce domaine.

Deuxième objectif : outil pour l'assemblage haute qualité des génomes en lectures longues

Pour annoter les polymorphismes de TE au sein d'un génome, puis pour explorer l'interaction entre l'ADNecc et le génome, il est primordial d'obtenir un assemblage de génome de haute qualité. Le processus de reconstruction du génome à partir des

millions de lectures générées par les plateformes de séquençage à haut débit est appelé assemblage *de novo* (Nagarajan et Pop, 2013). Depuis que le premier génome végétal a été déchiffré en 2000, plus de 700 génomes végétaux ont été assemblés (www.plabipd.de) (Bolger et al., 2017). La plupart des assemblages de génomes de plantes sont basés sur le séquençage d'ADN à lecture courte et composés de milliers de contigs fragmentés (Belser et al., 2018). Deux génomes de référence modèles de plantes, la dicotylédone *Arabidopsis thaliana* (The Arabidopsis Genome Project, 2000) et la monocotylédone *Oryza sativa* (International Rice Genome Sequencing Project, 2005) ont été séquencés sur la base du séquençage Sanger et de l'assemblage par une approche clone par clone et des améliorations ultérieures, et figurent parmi les meilleurs assemblages de génomes végétaux. Cependant, même les génomes de référence contiennent des lacunes. La séquence du génome d'*A. thaliana* Col-0 a été publiée en 2000, et après des décennies de recherche, le génome de référence TAIR10 est devenu l'étalon standard pour Arabidopsis. Cependant, l'enrichissement en séquences hautement répétées dans les régions centromériques, télomériques et au niveau des ADN ribosomiques a fait que ces régions présentent des erreurs d'assemblage ou ne sont pas séquencées. Ce génome contient encore 165 lacunes avec des bases inconnues (N-stretches) et environ 25 Mb de régions manquantes, principalement au niveau des centromères (Long et al., 2013). Pour obtenir le génome de télomère à télomère d'*A. thaliana* Col-0, Wang et al. (2021) ont introduit la stratégie de substitution de séquence clonée en chromosome artificiel bactérien (BAC) avec le séquençage Pacbio, résolvant les centromères et les lacunes (Wang et al., 2021). Pour évaluer les caractéristiques génétiques et épigénétiques des centromères, Naish et al. ont assemblé un génome Col-0 d'*A. thaliana* hautement contigu à l'aide de lectures ultra-longues générées par ONT, fournissant un paysage approfondi de l'évolution des centromères (Naish et al., 2021).

Les progrès réalisés en matière de longueur moyenne des lectures, d'algorithmes d'assemblage et de logiciels ont grandement contribué à l'intégrité et à la qualité de l'assemblage des génomes. Les difficultés liées au comblement des lacunes, à la caractérisation des haplotypes et à la construction de génomes gigantesques sont en cours de résolution (Marx, 2021 ; Sun et al., 2021). Au cours des deux dernières années, l'analyse comparative de génomes ou de plusieurs individus d'une même espèce a montré qu'un seul génome de référence ne suffit pas à rendre compte de la diversité génétique d'une espèce (Bayer et al., 2020). De nombreux facteurs ont

conjointement favorisé la construction et la recherche sur les pan-génomes végétaux et animaux. Le pan-génome est un terme général désignant tous les gènes d'une espèce, où les gènes totaux sont distincts de ceux du génome individuel (Tettelin et al., 2005). Ainsi, le séquençage à long terme à l'échelle de la population a progressivement commencé à se développer dans la recherche en génomique évolutive et fonctionnelle et dans la recherche sur la sélection des plantes cultivées. Des pan-génomes ont été réalisés sur diverses plantes modèles et cultivées, notamment la tomate (Gao et al., 2019 ; Alonge et al., 2020), le riz (Qin et al., 2021), le blé (Walkowiak et al., 2020), le maïs (Hufford et al., 2021), etc.

Dans ce manuscrit de thèse, les percées clés pour l'assemblage des génomes des plantes seront présentées, notamment le comblement des lacunes, la mise en phase des haplotypes, la construction de très gros génomes et le pan-génome mentionné ci-dessus. Cependant, dans la pratique, le défi de l'assemblage du génome demeure. Différents assemblages produits par différents assembleurs ou le même assembleur avec différents paramètres ont des performances différentes, et le meilleur assemblage ayant à la fois une haute résolution en termes de contiguïté et de répétition ne peut pas être obtenu dans un seul assemblage. Mes résultats sur le développement d'un outil de méta-assemblage (SASAR) pour réconcilier le résultat de différents assemblages à partir de données de séquençage en lectures longues seront présentés. Les résultats obtenus sur les génomes d'*A. thaliana* Col-0 et du riz *Oryza sativa ssp. japonica* cv. Nipponbare seront discutés. Grâce à SASAR, l'assemblage du génome sera construit de manière robuste avec une grande contiguïté, permettant de détecter les variants structuraux avec une précision accrue.

Troisième objectif : quel impact des ADNecc sur la stabilité du génome et les variations structurales ?

Les variations structurales (SV) font référence à l'altération de fragments chromosomiques qui sont différents du génome de référence, les fragments variants étant généralement plus grands que 50 pb. Les principaux types de SV sont l'insertion, la délétion, la duplication, l'inversion et la translocation (Stankiewicz et Lupski, 2010). Un grand nombre d'études ont montré que les SV jouent un rôle clé dans des caractéristiques agronomiques importantes, telles que la résistance aux stress biotiques et abiotiques, le temps de floraison, l'architecture de la plante, le rendement,

la qualité des grains ou des fruits (Tao et al., 2019). De nombreux SV détectés dans 100 variétés de tomates ont un impact sur le dosage et les niveaux d'expression des gènes, entraînant des changements dans le goût, la taille et le rendement (Alonge et al., 2020). Zhou et al. (2021) ont identifié une inversion chromosomique de 1,67 Mb dans le génome de la pêche plate, responsable du passage de la pêche ronde à la pêche plate. En outre, plusieurs études ont montré que les SV peuvent aider à résoudre la structure de la population et fournir des informations supplémentaires valables pour mieux comprendre les processus de domestication des plantes (Alonge et al., 2020 ; Hufford et al., 2021 ; Qin et al., 2021).

L'insertion de TE dans le génome peut provoquer des changements spectaculaires dans la structure des chromosomes, à la fois par l'intégration de fragments de gènes et par l'induction de SV (Feschotte et Pritham, 2007). Dans le génome du riz, par exemple, le nombre de copies de TE et la distribution des inversions et des délétions contribuent à la variation au sein du genre *Oryza* (Piegu et al., 2006 ; Hurwitz et al., 2010). Dans le génome du maïs, les translocations des éléments Ac peuvent entraîner des délétions, des inversions, des translocations ou d'autres réarrangements (Yu et al., 2012). En outre, les TE à capacité de capture, tels que Pack-MULE chez le riz et Pack-TIR chez 100 espèces animales, ont été décrits comme favorisant l'évolution adaptative en formant de nouveaux gènes (Talbert et Chandler, 1988 ; Jiang et al. 2004, 2011 ; Tan et al. 2021). De plus, les fusions entre les transposons d'ADN et les gènes codant pour les protéines dans tous les génomes de tétrapodes démontrent que les TE constituent un réservoir pour façonner de nouvelles structures protéiques (Cosby et al., 2021). Cependant, on sait encore peu de choses sur les fusions entre TE et gène dans l'ADNecc en raison du faible nombre d'exemples mis en évidence chez les plantes.

Dans ce manuscrit de thèse, mes résultats sur les SV associés aux TE et SV dans le génome de mutants épigénétiques seront décrits. L'accent sera mis sur l'algorithme et la validation visuelle des SVs dans le développement d'outils, favorisant l'étude de l'interaction entre l'ADNecc et le génome. Pour cette partie de ma thèse, le matériel végétal choisi sera un mutant hypométhylé d'*A. thaliana*, qui possède un fort taux d'ADNecc généré par des TEs actifs. Des plantes combinant des mutations dans la méthylation de l'ADN associée à DDM1 (Decrease DNA Methylation 1), le silencing post-transcriptionnel et la méthylation de l'ADN dirigée par l'ARN (triple mutants *ddm1 rdr6 pol4*) ont été étudiées au niveau ADNecc et génome en utilisant le séquençage en

lectures longues. D’après ces résultats sur la dynamique de l’ADNecc et des SV dans des mutants épigénétique d’*A. thaliana*, je montrerai que les voies épigénétiques contrôlent la stabilité du génome au-delà de la mobilité des TE. Les réarrangements chaotiques du génome et le chimérisme des gènes mis en évidence dans cette étude renforcent le concept d'une évolution du génome à deux vitesses chez *A. thaliana*, guidée par l'épigénome.

Résumé

Les éléments transposables (TEs) sont des séquences d'ADN répétitives avec la capacité intrinsèque de se déplacer et de s'amplifier dans les génomes. La transposition active des TEs est liée à la formation d'ADN circulaire extrachromosomique (ADNecc). Cependant, le paysage complet de ce compartiment d'ADNecc ainsi que ses interactions avec le génome n'étaient pas bien définies. De plus, il n'existait au début de ma thèse aucun outil bioinformatique permettant d'identifier les ADNecc à partir de données de séquençage en lectures longues.

Pour répondre à ces questions au cours de mon doctorat, nous avons tout d'abord développé un outil, appelé ecc_finder, pour automatiser la détection d'ADNecc à partir de séquences en lectures longues et optimisé la détection à partir de séquences de lecture courte pour caractériser la mobilité des TE. En appliquant ecc_finder aux données eccDNA-seq d'Arabidopsis, de l'homme et du blé (avec des tailles de génome allant de 120 Mb à 17 Gb), nous avons documenté l'applicabilité d'ecc_finder ainsi que l'optimisation du temps de calcul, de la sensibilité et de la précision.

Dans le deuxième projet, nous avons développé un outil de méta-assemblage appelé SASAR pour réconcilier les résultats de différents assemblages de génomes à partir de données de séquençage en lectures longues. Pour différentes espèces de plantes, SASAR a obtenu des assemblages de génome de haute qualité en un temps raisonnable et a permis de détecter les variations structurales causées par les TE.

Dans le dernier projet, nous avons utilisé le génome assemblé par SASAR et l'ADNecc détecté par ecc_finder pour caractériser les interactions entre les ADNecc et le génome. Dans les mutants épigénétiques hypométhylés d'*Arabidopsis thaliana*, nous avons mis en évidence le rôle de l'épigénome dans la protection de la stabilité du génome non seulement contre la mobilité des TE mais aussi envers les réarrangements génomiques et le chimérisme des gènes. Globalement, nos découvertes sur l'ADNecc, la stabilité du génome et leurs interactions réciproques, ainsi que le développement d'outils, offrent de nouvelles perspectives pour comprendre le rôle des TE dans l'évolution adaptative des plantes à un changement rapide de l'environnement.

Mots clés: ADN circulaire extrachromosomique, assemblage du génome, élément transposable, séquençage en lectures longues.

Abstract

Transposable elements (TEs) are repetitive DNA sequences with the intrinsic ability to move and amplify in genomes. Active transposition of TEs is linked to the formation of extrachromosomal circular DNA (eccDNA). However, the complete landscape of this eccDNA compartment and its interactions with the genome are not well defined. In addition, at the beginning of my thesis, there were no bioinformatics tools available to identify eccDNAs from long-read sequencing data.

To address these questions during my PhD, we first developed a tool, called ecc_finder, to automate eccDNA detection from long-read sequencing and optimized detection from short-read sequences to characterize TE mobility. By applying ecc_finder to Arabidopsis, human and wheat eccDNA-seq data (with genome sizes ranging from 120 Mb to 17 Gb), we documented the broad applicability of ecc_finder as well as optimization of its computational time, sensitivity and accuracy.

In the second project, we developed a meta-assembly tool called SASAR to reconcile the results of different genome assemblies from long-read sequencing data. For different plant species, SASAR obtained high quality genome assemblies in an efficient time and resolved structural variations caused by TEs.

In the last project, we used SASAR-assembled genome and ecc_finder-detected eccDNA to characterize eccDNA-genome interactions. In *Arabidopsis thaliana* hypomethylated epigenetic mutants, we highlighted the role of the epigenome in protecting genome stability not only from TE mobility but also from genomic rearrangements and gene chimerism. Overall, our findings on eccDNA, genome stability and their interactions, as well as the development of tools, offer new insights into the role of TEs in the adaptive evolution of plants to rapid environmental change.

Keywords: extrachromosomal circular DNA, genome assembly, transposable element, long read sequencing.

Acknowledgments

A PhD thesis is a long journey that no one travels alone. I would like to thank all the people who accompanied me throughout these 4 long years.

First of all, I would like to thank Marie Mirouze and Alain Ghesquière for welcoming me in their institute. I thank the members of the jury, my rapportrice **Anne Roulin** and rapporteur **Todd Blevins** as well as my examiners, **Nathalie Picault, Romain Guyot** and **Magnus Nordborg** who accepted to evaluate my work. I would also like to thank Romain Guyot, Moaine El Baidouri, Olivier Panaud and Hajk-Georg Drost for having accepted to be part of my thesis committee.

A big thank you to my thesis director **Marie Mirouze**. Thank you for being so amazing, so nice, so cool, for letting me do this thesis under the best conditions a PhD student could dream of, and for supporting me through the years. Thank you for welcoming me as a non-French speaker at the beginning, you were very kind to help me with many uncountable tasks including administration, apartment and basic weekend activities. You have always praised me without hesitation and this has kept me motivated. From Montpellier to Versailles, from Tubingen to Leipzig, from Amsterdam to Udine… you offered me so many opportunities, freedom in my research. When I had difficulty expressing in English, you never lost patience and helped me improve my presentation skills. Thank you for the tremendous support you have given me especially under difficulties of the pandemic, no matter what I write, words won't be enough to express the immense pleasure that it was to do my PhD with you. I hope I would not have made you too desperate, especially with my procrastination in writing. I hope you keep having great PhD students, because you deserve it. If one day I become a researcher, I hope I can be like you.

I am fortunate to be part of the MANGO team of bioinformaticians and technicians. I would like to thank **Olivier Panaud** who was always available for scientific or not conversations, listened to me, taught me and advised me despite his busy. I would also like to thank **Christel Llauro** for her great help in the wet lab, without her I would have no data to analyze. Thank you for all your efforts to manage the sequencing perfectly. I would like to thank **Moaine El Baidouri**, not only for having shared so many times your experience and advice on research, but especially for your help in helping me to get

acquainted and integrate into the French environment. I would like to thank **Eric Lasserre** for your huge help in solving different difficulties in data analysis, without which I would have been lost. Thanks to **Joris Bertrand** for constantly providing so many fascinating orchid images that open the door to outdoor exploration.

It is my pleasure to work with all the PhD students of the team: **Marie Christine Carpentier**, thank you for having took us different academic and cultural activities, and for having made us laugh. **Emilie Aubin**, thanks for those days spent together talking about everything and for the countless times we experienced laughter and sadness together. **Assane Mbodj** and **Abirami Soundiramourtty**, even though I haven't known you for a long time. Thank you for your good cheer and for your kindness in any situation.

I also want to thank you **Alain Ghesquière** for following my thesis work, for all your advice, encouragement and support. Thank you for your patience.

The life of a PhD is a roller coaster, with ups and downs, and Covid brings a cliff drop. However, when I look back on four years, I realize that there were many more ups than downs and this thanks to all the people in LGDP. Especially, I would like to thank **Frédéric Pontvianne** with whom shared an office at the beginning of my PhD. Thank you for the welcome and every scientific discussion. I would also like to thank **Guillaume Moissiard** for organizing every journal club, inviting amazing speakers and many international congresses. Thanks to **Elisabeth Goetschy** for managing all the administrative questions. My words are pale, thank you all for countless effort to make this friendly and enjoyable environment. I also want to thank the best officemate **Avilien Dard** and **Jean Loup Zitoun**. We made it through those four years, not without pain, questioning and muscle building, but we made it. I am very happy to have shared this experience with you, so thank you. And to **Eduardo Muñoz**, you know that I will miss your hugs a lot! I sincerely want to thank you for the moments we spent together, our passionate and angry. Thanks for all that and for all the other stuff.

A big thank you also to my research consortium, Epidiverse, for providing multiple travel fellowships for secondments and organizing five summer schools to enrich the interaction between cross-disciplinary research.

I would like to thank **Detlef Weigel** for hosting me in his open, friendly and well-organized research group. Thank you for deepening my knowledge and helping me tremendously in the direction of my research. To **Haik Georgi Drost**, thank you for every brainstorming conversation during the lunch and for availability during my numerous questions. To **Benjamin Buchfink**, thank you for those days spent together talking about every little thing that later mattered and huge help in bioinformatics. To **Angel Wibowo**, thank you for caring and looking after me like a big brother. To **Adrian Contreras**, thank you for your perfect hosting and for the compliments on my work. I would also like to thank **Peter Stadler** for welcoming me on secondment to his group. Thank you, Adam Nunn, for great help in developing tools and for having a great time in Germany. I would also like to thank **Etienne Bucher** for continuous discussion and cooperation.

Thank you all the remaining members of Epidiverse: Dr. **Koen Verhoeven**, programme manager **Margreet Bruins** and their students: **Morgane van Antro** and **Cristián Peña** from NIOO-KNAW, Netherland; Dr. **Noe Fernandez Pozo**, Dr. **Katrin Heer**, Dr. **Lars Opgenoorth**, Dr. **Stefan Rensing** and their students: **Nilay Can** and **Bárbara Díez Rodríguez** from Philipps University Marburg, Germany; Dr. **Oliver Bossdorf**, Dr. **Niek Scheepens** and their student **Dario Galanti** from University Tübingen, Germany; Dr. **Vit Latzel** and his student **Iris Sammarco** from Botanicky Ustav AVCR, Czechia; Dr. Etienne Bucher's student **Maria Estefanía Lopéz** from University of Geneva, Switzerland; Dr. **Claude Becker** and his student **Daniela Ramos Cruz** from LMU Munich, Germany; Dr. **David Langenberger** from ecSeq , Germany; Dr. **Conchita Alonso** and her student **Anupoma Niloya Troyee** from CSIC, Spain; Dr. **Emanuele DePaoli** and his student **Bhumika Dubay** and **Paloma Perez Bello Gil**; Dr. **Ivo Grosse** and his student **Samar Fatma** from Martin Luther University, Germany. What a great journey to have 15 PhD students and 18 principal Investigators to work on a project, words can't express my thanks and love enough!

A big thank you also to all my friends who supported me during these 4 years of thesis or since much longer, **Abraham**, **Sandrine**, **Benjamin**, **Dadi**, **Yuanyuan**, **Jiayue**, **Jiaqi** and I surely forget some...

Finally, I would like to express my gratitude to my **parents**. I'm sorry you couldn't come to my defense because long flights are being cancelled by the terrible war, but I thank

you for your none-stop support so that I could achieve my dreams without worry. To my fiancé **Shiva**, thank you for your efforts to maintain a long-distance relationship, without your support and optimism I would not be as happy as I am now. Thank you, **Daksh**, my youngest and sweetest nephew, your constant laughter is the best thing in the world.

Carl Sagan said: “Extinction is the rule. Survival is the exception”. When our life, when the world, feels like a never-ending emergency. Sometimes, just making it through the day can be a struggle. But we collect scars, physical and mental reminders of what we've been through. The journey itself is all that life has to offer.

I sincerely thank everyone I have met and I wish you all happiness every day. Stay curious and optimistic about the days ahead. Thank you, thank you, and thank you !!!

Abbreviations

Activator	Ac
Argonaute RISC Component 4	AGO4
Aspartic protease	AP
Cysteine-rich RECEPTOR-like protein kinase 19	CRK19
Decrease in DNA Methylation I	DDM1
Dissociator	Ds
Duplication	DUP
Epigenetic Recombinant Inbred Lines	EpiRILs
Extrachromosomal circular DNA	eccDNA
Recognition of Peronospora Parasitica 5	RPP5
Fork Stalling, Template Switching and Transposition	FoSTeST
Group-specific antigen	GAG
Homologous Recombination	HR
Insertion	INS
Integrase	IN
Inversion	INV
Long interspread nuclear element	LINE
Long terminal repeat	SINE
Lysine 9 methylation on histone H3	H3K9me
Megabases / Gigabases	Mb / Gb
Miniature inverted-repeat transposable element	MITE
Next generation sequencing	NGS
Non-Homologous End-Joining	NHEJ
Nucleolar organizing regions	NORs
Oxford Nanopore Technologies	ONT
Pacific Biosciences	PacBio
Polyprotein	POL
Reverse-transcriptase	RT
Ribonuclease H	RH
RNA-dependent RNA polymerase 2	RDR2
RNA-dependent RNA polymerase 6	RDR6
RNA-directed DNA methylation	RdDM
Short interspread nuclear element	SINE
Single nucleotide polymorphism	SNP

Structural variant	SV
Terminal inverted repeat	TIR
Transposable element	TE
Virus-like particule	VLP

Table of Contents

List of Figures *19*

1. Introduction *21*

1.1. Tracking transposable elements mobility in plants **23**

1.1.1 Characterization of Transposable Elements (TEs) 23

1.1.2 TE transposition and gene capture 27

1.1.3 Mechanisms of TE silencing 31

1.1.4 Genome instability mediated by TEs 33

1.1.5 TEs in the form of extrachromosomal circular DNA (eccDNA) 34

1.2 Obtaining a high-quality genome assembly prior to TE annotation **36**

1.2.1 What is a genome assembly 36

1.2.2 The revolution of long read sequencing technologies 37

1.2.3 Assembling centromeres: first « gap free » plant genomes 38

1.2.4 The challenge of assembling very large plant genomes 40

1.2.5 Resolving haplotypes in plant genomes 41

1.2.6 Pan-genomes: One genome is not enough 43

1.3 Structural variants in the plant genome **48**

1.3.1 Why to detect structural variants (SVs)? 48

1.3.2 Algorithms of SV detection 48

1.3.3 Visual validation for SV prediction 50

1.4 Objectives of the thesis work and main achievements **52**

2. Methods and Results *53*

2.1 Ecc_finder: Developing a new tool for eccDNA detection **55**

2.1.1 My contribution to ecc_finder 55

2.1.2 ecc_finder manuscript (Frontiers in Plant Science, 2021) 57

2.1.3 Update on ecc_finder 66

2.2 SASAR: a new tool for meta-assembling plant genomes with long reads **69**

2.2.1 My contribution to SASAR 69

2.2.2 SASAR manuscript (bioRxiv, 2022) 70

2.3 Tracking TE mobility and genome instability with long reads **83**

2.3.1 My contribution to the discovery of genome instability in Arabidopsis epigenetic mutants 83

2.3.2 « Chimera » manuscript (bioRxiv, 2022) 84

3. General discussion and perspectives 113

3.1 Future trends to understand the role of eccDNAs 115

3.1.1 Remaining questions on eccDNA inheritance and the emergence of new genes 115

3.1.2 Towards eccDNA detection directly from genomic data 116

3.2 Future trends to obtain high-quality genome assembly 117

3.2.1 Choice of long read sequencing technologies 117

3.2.2 Emerging tools to characterize SVs in pan-genomes 118

3.2.3 Having a pan-genome alternative 119

3.3 Future trends to uncover the relationship between TEs, eccDNAs and SVs 120

3.3.1 On the role of VLP in fast TE and TE-gene chimerism evolution 120

3.3.2 Interactions between eccDNA and the genome in the context of the 3D genome 121

4. Bibliography 123

5. Appendix 141

A short summary of my 3 contributions 143

Picart-Picolo et al., Genome research 2020 144

Lanciano et al., Plant Transposable Elements 2021 154

Nunn et al., Plant biotechnology journal 2021 161

List of Figures

1. Introduction 21

1.1. Tracking transposable elements mobility in plants 23

Figure I.1. Classification and structural features of transposable elements

Figure I.2. Detailed structure of the *Gypsy* and *Copia* LTR retrotransposons.

Figure I.3. TE transposition mechanisms and formation of eccDNA.

Figure I.4. Proposed models through which TE capture host sequences.

Figure I.5. Main epigenetic actors involved in DNA methylation maintenance and *de novo* mechanisms in *Arabidopsis thaliana*.

1.2 Obtaining a high-quality genome assembly prior to TE annotation 36

Figure I.6. Sequencing technologies used for plant genome assemblies until September 2021.

Figure I.7. An overview of genomic population-scale studies in plant using long-read sequencing.

1.3 Structural variants in the plant genome 48

Figure I.8. Detecting structural variants using *de novo* assembly and read mapping modes.

2. Methods and Results 53

2.1 Ecc_finder: Developing a new tool for eccDNA detection 55

Figure II.1. Reanalysis of the characteristics of eccDNAs from leaf, flower, stem and root tissues of *Arabidopsis thaliana* in Wang et al. (2021).

2.2 SASAR: a new tool for meta-assembling plant genomes with long reads 69

Figure II.2. Schematic pipeline of SASAR (Super ASsembly from Assembly Reconciliation).

Figure II.3. SASAR performance in assembling the centromeres in *A. thaliana* Col-0.

Figure II.4. SASAR performance in a 1kb insertion in the non-centromeric region of *A. thaliana* Col-0.

Figure II.5. SASAR improved the assembly of ONSEN/ATCOPIA78 in *A. thaliana* Col-0.

2.3 Tracking TE mobility and genome instability with long reads 83

Figure II.6. The eccDNA repertoire in *A. thaliana* epigenetic mutants contains active full length and truncated TEs.

Figure II.7. TE insertion polymorphisms in *A. thaliana* epigenetic mutants revealed by ONT genome sequencing.
Figure II.8. Chimeric eccDNA containing a truncated EVD-gene fusion corresponding to a chimeric genomic integration in *A. thaliana ddm1 rdr6 pol4* mutant.
Figure II.9. ATCOPIA21 mobility in the *A. thaliana ddm1 rdr6 pol4* triple mutant is associated with *RPP5* locus duplication.
Figure II.10. Detection of a 55kb duplication on chromosome 1 in the *A. thaliana ddm1-2* mutant genome.
Figure II.11. Overall genomic instability detected in this study in *A. thaliana* epigenetic mutants.
Supplementary Figure 1. A large inversion in the *A. thaliana ddm1*-2 mutant genome.

1. Introduction

1.1. Tracking transposable elements mobility in plants

1.1.1 Characterization of Transposable Elements (TEs)

TEs are repetitive DNA sequences with the intrinsic ability to move within the genome by a mechanism called transposition. They replicate and expand in the genome like a virus, are usually ranging from 100 to 10,000 bp in length, but sometimes far larger (Arkhipova and Yushenova, 2019). In recent years, genome sequencing of many species has been able to demonstrate that TEs and their relics are major components of eukaryotic genomes ranging from ~3% in the yeast *S. cerevisiae* (Kim et al., 1998) and up to 45% in humans (Lander et al., 2001) and >85% in some plants, such as maize (Schnable et al., 2009) and wheat (Wicker et al., 2018). TEs are very diverse in nature and number. According to the replication mode, TEs can be classified into class I retrotransposons and class II DNA transposons (Figure I.1) (Wicker et al., 2007).

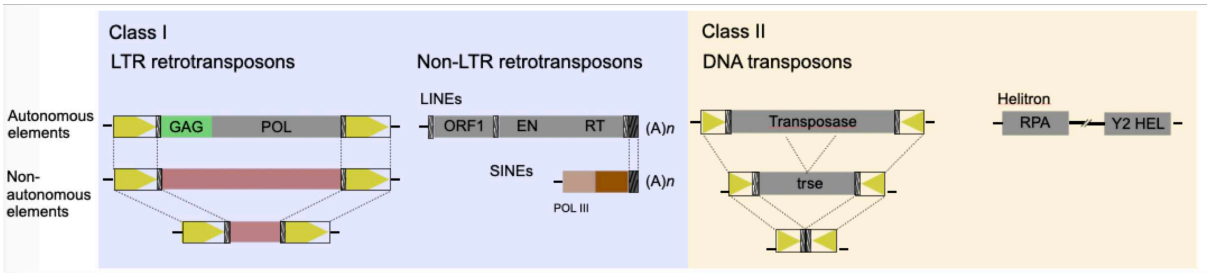


Figure I.1. Classification and structural features of transposable elements. Two classes of TEs, class I retrotransposons and class II DNA transposons, have autonomous and non-autonomous elements, respectively. Gag is highlighted in green, terminal repeats are colored in yellow (edited from Feschotte et al., 2002; Wicker et al., 2007)

Class I retrotransposons

Class I elements or retrotransposons are the most abundant and widespread in eukaryotes. This is due to their "copy-and-paste" transposition mechanism that allows the generation of a large number of copies from a single DNA sequence. LTR retrotransposons are characterized by one or two Open Reading Frames (ORFs) flanked by two LTRs, usually starting with TG in the 5' and ending with CA in the 3'. LTR sequences range in length from a few hundreds to over a thousand nucleotides. They contain promoter and regulatory regions separated into 3 functional domains (U3, R and U5). U3 domain harbors trans-activator binding sites, while U5 domain marks the

start of transcription but also indicates the end of transcription and the signal for polyadenylation. The ORFs typically code for Group-specific antigen (GAG), a capsid polyprotein participating in the formation of a virus-like particle (VLP), and polyprotein (POL) cleaved into 4 active functional domains for: RT, a reverse-transcriptase, RH, a RNase H, AP, an aspartic protease, and IN, an integrase (Figure I.2) (Havecker et al., 2004; Sabot and Schulman, 2006).

LTR retrotransposons are grouped into 2 superfamilies that differ in the relative position of the POL gene-encoded enzyme domains (Figure I.2). TEs of the *Gypsy* superfamily sometimes possess an ORF called putative env, similar to the env encoding the retrovirus envelope glycoprotein. Indeed, the identification of common protein motifs between the sequences of integrases, reverse transcriptases and env proteins of retrotransposons and retroviruses indicates an evolutionary linkage (McClure, 1991; Capy et al., 1996; Lerat and Capy, 1999).

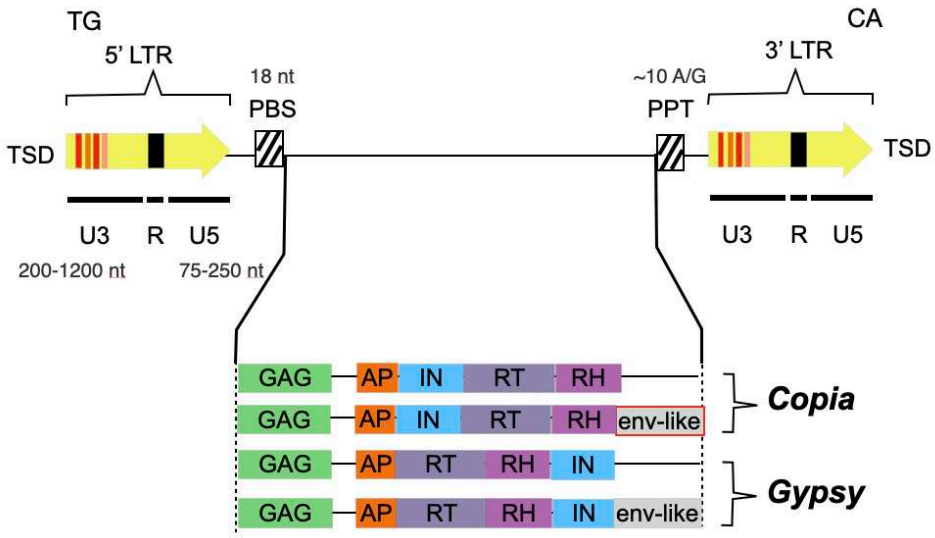


Figure I.2. Detailed structure of the *Gypsy* and *Copia* LTR retrotransposons. These two TE families differ in the organization of the POL polyprotein domains. The PBS (primer binding site) and PPT (polypurine tract) sites are involved in reverse transcription of the element. (Modified from Havecker et al., 2004; Sabot and Schulman, 2006).

Retrotransposons without LTRs, are divided into two classes: LINEs and SINEs (Long and Short Interspread Nuclear Elements). Both types of elements terminate at the 3' end with a polyA sequence of variable length. LINEs have coding regions that include: ORF1, a gag-like protein; EN, an endonuclease; and RT a reverse transcriptase. LINEs

represent up to 20% of the human genome (Lander et al., 2001) whereas in plants LINEs appear to be rare. SINEs do not encode a protein and are non-autonomous elements that depend on LINEs to transpose. The internal region of SINEs is highly variable and depends on the family of the element. The most studied element belonging to the SINE class is the *Alu* element, which alone represents 11% of the human genome (Lander et al., 2001). Although the 3' half of these elements is of unknown origin, the 3' end shows similarities with LINE sequences indicating that SINE elements may parasitize the transposition machinery of LINE elements (Ogiwara et al., 1999). SINEs are thought to result from reverse transcripts of short RNAs (usually tRNAs in plants) followed by their integration into the genome (Deragon and Zhang, 2006).

Unlike LTR retrotransposons and despite the presence of an ORF encoding a gag-like protein, non-LTR retrotransposons do not produce a DNA copy of their RNA in the cytoplasm. Indeed, the transposition of these elements is achieved by reverse transcription at the integration site (Cost et al., 2002; Kazazian, 2004). Finally, it has been shown that most LINE sequences in genomes are truncated at a priori random 5' region by a mechanism called "5' truncation" (Kazazian, 2004).

Class II DNA transposons

Class II TEs, or DNA transposons, do not require RNA as a mediator in the transposition process. They move by a mechanism of excision from one genomic position and then integration at another position by a transposase. This "cut and paste" mechanism usually does not lead to an increase in copy number. However, a gap repair mechanism can allow the restoration of the sequence of the element at the donor site.

DNA transposons are characterized by the presence of terminal inverted repeats (TIRs) at both ends, with TIR ranging from 14-500 bp in length. Autonomous elements encoding a transposase are divided into different superfamilies such as Ac/Ds, CACTA, MULE. Like class I TEs, class II elements can become non-autonomous via accumulating mutations in their coding region or by complete deletion of the coding region. Finally, some non-autonomous elements called MITE ("Miniature Inverted-repeat Transposable Elements") are composed only of TIR sequences. This type of non-autonomous transposons can rapidly increase in copy number until they exceed the copy number of autonomous elements from which they originate. For example, in

the rice genome, MITEs are the elements with the highest copy number (about 90,000 copies in some varieties, Jiang et al., 2004).

A special case of DNA transposon that transpose using a rolling circle, called Helitrons was discovered more recently in eukaryotic genomes with typical 5'TC and 3'CTRR (R as A or G) termini and a stem-loop structure about 15-20 bp upstream of the 3' terminus (Kapitonov and Jurka, 2001). These transposons have the specificity of not possessing terminal repeats and of not generating target site duplications (TSDs). Their transposition would be done by a still hypothetical mechanism called "rolling-circle" close to the replication mechanism of certain bacteriophages and plasmids. Helitrons transposons, after transposition, are usually inserted into AT-rich regions of the AT target site (Kapitonov and Jurka, 2007).

Each class of TEs has both autonomous and non-autonomous elements. Autonomous elements encode the enzymes required for their transposition in contrast to the mobility of non-autonomous elements, which depends on the enzymes produced by autonomous elements of the same or related families. For example, in the *Activator* (*Ac*) / *Dissociator* (*Ds*) system, *Ac* is the autonomous type and *Ds* is the non-autonomous type. Without *Ac*, *Ds* cannot function by itself. Non-autonomous family members are usually derived from an autonomous family member by internal deletion (McClintock, 1950; A Howard and S Dennis, 1984).

1.1.2 TE transposition and gene capture

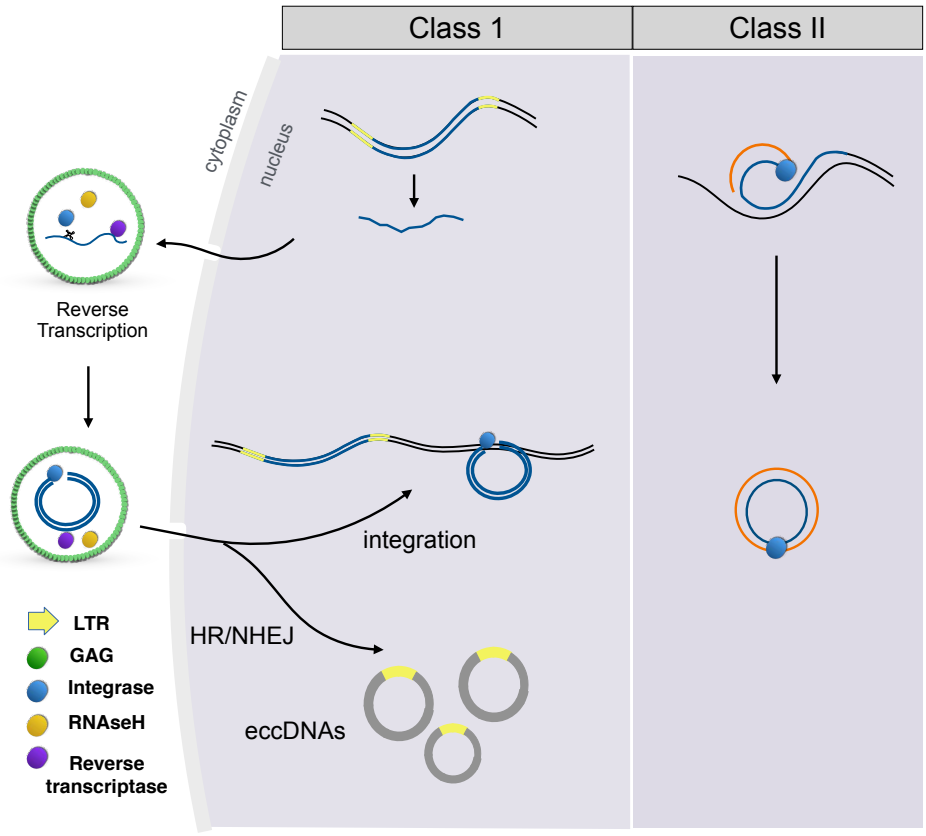


Figure I.3. TE transposition mechanisms and formation of eccDNA. Active retrotransposons initiate transcription and are translated into proteins and form the VLP. After undergoing reverse transposition, double-stranded extrachromosomal linear DNA enters the nucleus and inserts into new genomic loci thanks to the integrase or form eccDNA through homologous recombination (HR) or non-homologous end-joining (NHEJ). Active DNA transposons can also lead to the formation of eccDNA, for instance here an helitron transposing through a rolling circle mechanism. Edited from Lanciano et al., 2017; Wells and Feschotte, 2020.

Since TEs were first discovered in maize by Barbara McClintock in 1948 (McClintock, 1948) and have since been found in all animals and plants, as well as in various eukaryotes. The transposition is the hallmark feature of TEs, and the integration of a TE from a donor site to a target site is known as a complete transposition process.

The mechanism of an LTR retrotransposon transposition is close to that of retroviruses. This mechanism involves a complete transcription of the element from the 5' R region of the 3’ LTR to the 3' R region of the 5’ LTR. These transcripts are translated in the cytoplasm where they are used as a template for translation and also as a template for reverse transcription into DNA double strands (Schulman, 2013). In the cytoplasm, the polyprotein is processed by retrotransposon-encoded proteases into reverse

transcriptase, RNAseH, and integrase (Figure I.3). These, along with two transcripts forming a kissing loop structure, are specifically packaged into GAG-derived virus-like particles (VLPs). Subsequent reverse transcription involves two transfers of the DNA strand, resulting in the synthesis of complete copies of the retrotransposon with two identical LTRs in the form of extrachromosomal linear DNA (eclDNA). eclDNA in double-stranded form enters the nucleus by an unknown mechanism, and inserts into new genomic loci mediated by the integrase. Nevertheless, eclDNA can also be recognized by DNA repair mechanisms before its reinsertion into the genome and be captured by the Homologous Recombination (HR) or Non-Homologous End-Joining (NHEJ) pathway inducing the formation of extrachromosomal circular DNA (eccDNA). The model for the formation of these eccDNAs has been established from work on retroviruses (Li et al., 2001; Kilzer et al., 2003; Lanciano et al., 2017).

Over time, the newly integrated copies undergo mutations. Based on the divergence between the two LTRs, insertion age can be estimated. However, it should be noted that Sanchez et al. observed that new insertions of the *Arabidopsis thaliana* LTR retrotransposon *ATCOPIA78/ONSEN,* a heat-induced retrotransposon family, corresponded to high-frequency recombination between old and recent copies (Sanchez et al., 2017) suggesting that the widespread involvement of young autonomous copies may revive 'older relatives' (Sanchez et al., 2017; Drost and Sanchez, 2019).

DNA transposons have a transposition cycle that appears to be relatively short compared to class I elements and takes place only in the nucleus (Figure I.3). Transposase enzymes recognize and bind to terminal inverted repeats (TIRs) at both ends of the element. The element is excised from its locus and inserted at a new locus (Muñoz-López and García-Pérez, 2010). However, prior to reinsertion, the eclDNA can be recognized by DNA repair mechanisms (NHEJ or HR) inducing the formation of eccDNA (Sundaresan and Freeling, 1987; Li et al., 2001).

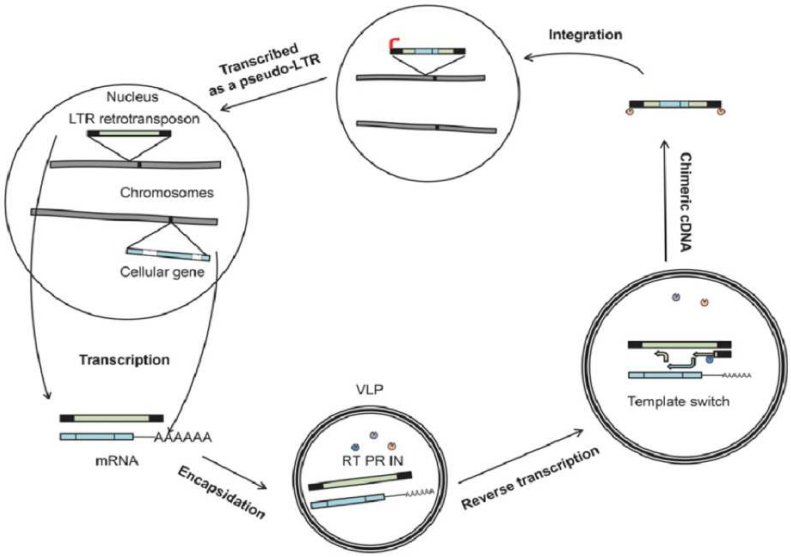
DNA transposons can take along, in addition to its sequence, a potentially coding genomic segment: this is the mobilization of endogenous elements. The ability of DNA transposons to mediate gene duplication has been revealed in plants, where MULE elements (Mutator Like transposable Element) have captured 1500 parental genes in rice, for instance, forming a Pack-MULE chimeric structure (Talbert and Chandler, 1988;

Jiang et al., 2004, 2011; Cerbin and Jiang, 2018). What's more, a recent study identified 370 Pack-TIRs mediated gene duplications in 100 animal reference genomes. This study demonstrates that Pack-TIRs prefer to capture exon sequences and most exons are fused to genes with transcriptional signals, and thus remodel gene structure and generate new genes (Tan et al., 2021).

Proposed models for TE capturing additional coding sequences have been proposed (Figure I.4): (A) the template switch model through which LTR TEs capture host sequences. In this model, a transcript originating from a host gene is encapsidated into the VLP. The template switching between a LTR retrotransposon transcript and this gene transcript occurs in the VLP, and thus generates a chimeric eclDNA. After integrating into host genome, the chimeric sequences act like pseudo-LTR retrotransposon and can be transcribed to enter a new cycle of retroposition (Tan et al., 2016). (B) the gap-filling model for the capture sequence of TEs by TIR DNA transposons. Double-strand breaks (DSBs) occur within TEs due to fragile sites (i) or excision of active TEs (ii). The 5' end is excised by exonucleases and gap repaired normally using the TE as template. The repaired strand may switch to a non-TE sequence. (C) The FoSTeST model (Fork Stalling, Template Switching, Transposition) through which TIR TEs capture sequences. 1. replication fork stalls at the transposon and a DSB occurs; 2. transposon and parent sequence are spatially close, leading to template jumping during repair to produce a chimeric fragment; 3. transposase recognizes the chimeric fragment and cleaves the insertion to another position 4 (Figure I.4) (Tan et al., 2021). Similar processes have been widely reported in human genetics and cancer genomics (summarized from Tan et al., 2016, 2021).

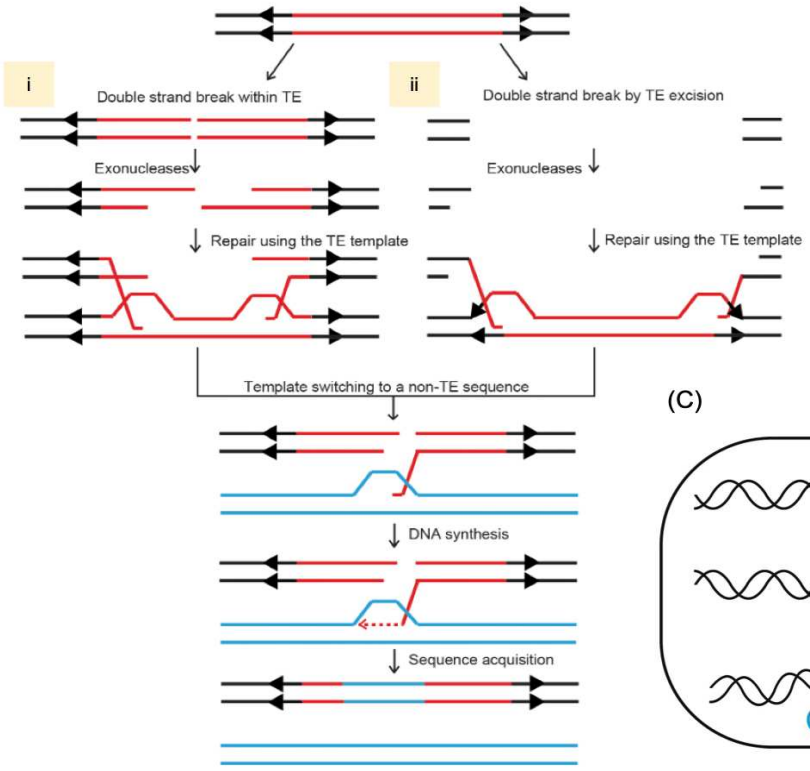
Retrotransposons

(A)



DNA transposons

(B)



(C)

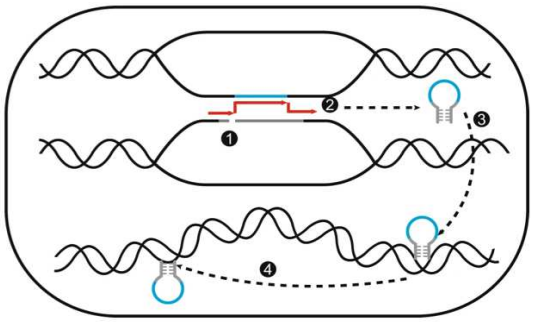


Figure I.4. Proposed models through which TE capture host sequences. (A) The template switch model through which LTR TEs capture sequences in the VLP, and then integrate into host genome (Tan et al., 2016). (B) Gap-filling model for the capture sequence of TIR TEs. Double-strand breaks (DSBs) occur within TEs due to fragile sites (i) or excision of active TEs (ii) and then gap repair. (C) FoSTeST model through which TIR TEs capture sequences (Tan et al., 2021).

1.1.3 Mechanisms of TE silencing

To suppress the activity of TEs, the host genome has evolved mechanisms triggering and maintaining the silencing of TEs through DNA methylation, repressive histone modifications, small RNA and chromatin pathways (Fultz et al., 2015).

DNA methylation corresponds to the addition of a methyl group on certain nucleotides. In eukaryotes this modification affects almost exclusively cytosines. In plants, unlike mammals, methylation is not restricted to cytosines in a CG context but can also be observed at cytosines in a CHG and CHH context (where H can be any nucleotide except G). Cytosine methylation is mainly detected at pericentromeric regions rich in repeated sequences and poor in genes (Inagaki et al., 2017). This methylation is also strongly correlated with the presence of a heterochromatin-specific mark, histone 3 lysine 9 (H3K9me2) dimethylation, and with the presence of siRNAs (Kasschau et al., 2007; Roudier et al., 2009; Inagaki, 2021). TEs are methylated in all 3 cytosine contexts. Their methylation is thought to be associated with their repression, since loss of this methylation leads to transcriptional reactivation of TEs (Lippman et al., 2004) and an increase in their mobilization (Kato et al., 2003; Mirouze et al., 2009; Tsukahara et al., 2009). Maintenance of these methylation patterns over generations is mediated by specific methyltransferases such as MET1 (METHYLTRANSFERASE 1), CMT2 and CMT3 (CHROMOMETHYLASES 2 and 3) (Figure I.5). But while most methylations are passed down through generations, they can also occur *de novo*, in any context, through a mechanism of RNA-directed DNA methylation (RdDM) (Wassenegger et al., 1994; Law and Jacobsen, 2010; Lloyd and Lister, 2022).

Additional proteins, such as the chromatin remodeler protein DDM1 (DECREASE IN DNA METHYLATION 1) protein, related to the SW12/SNF2 family of ATP-dependent chromatin remodeling factors, are essential for the maintenance of DNA methylation in different cytosine contexts (Vongs et al., 1993; Lippman et al., 2004). DDM1 appears to primarily control the silencing of TEs (approximately 40% of TEs in *Arabidopsis thaliana*), and in particular of long TEs localized in heterochromatin, thus preventing their reactivation and transcription (Lyons and Zilberman, 2017). Recently, Berger et al. showed that DDM1 is involved in depositing the histone variant H2A.W to silence TEs (Bourguet et al., 2021; Osakabe et al., 2021). The *ddm1* mutants have been widely observed as hypomethylated in all cytosine contexts. Some phenotypes revealed in a

ddm1 context are related to alterations in genome structure (Tsukahara et al., 2009), but some others are associated with epigenetic modifications that influence gene expression and generate stable epialleles (Kinoshita et al., 2007). To assess the stability of DNA methylation perturbations and their consequences, epiRIL (epigenetic Recombinant Inbred Lines) populations were generated from *ddm1* or *met1* mutants (Johannes et al., 2009; Reinders et al., 2009). The principle of these populations is to maximize epigenetic variability by minimizing nucleotide changes. The *ddm1* mutant-derived epiRIL population was obtained by a cross between the *ddm1* mutant and a wild-type (WT) plant followed by a back-cross between the F1 and a WT plant. The resulting F2 plants were genotyped to select *DDM1* homozygous individuals. Genotypes and epigenotypes were then fixed by self-fertilization for 6 generations (Cortijo et al., 2014a). The *met1* epiRIL population was obtained using a similar design except that the F1 was selfed to obtain the F2 material (Reinders et al., 2009).

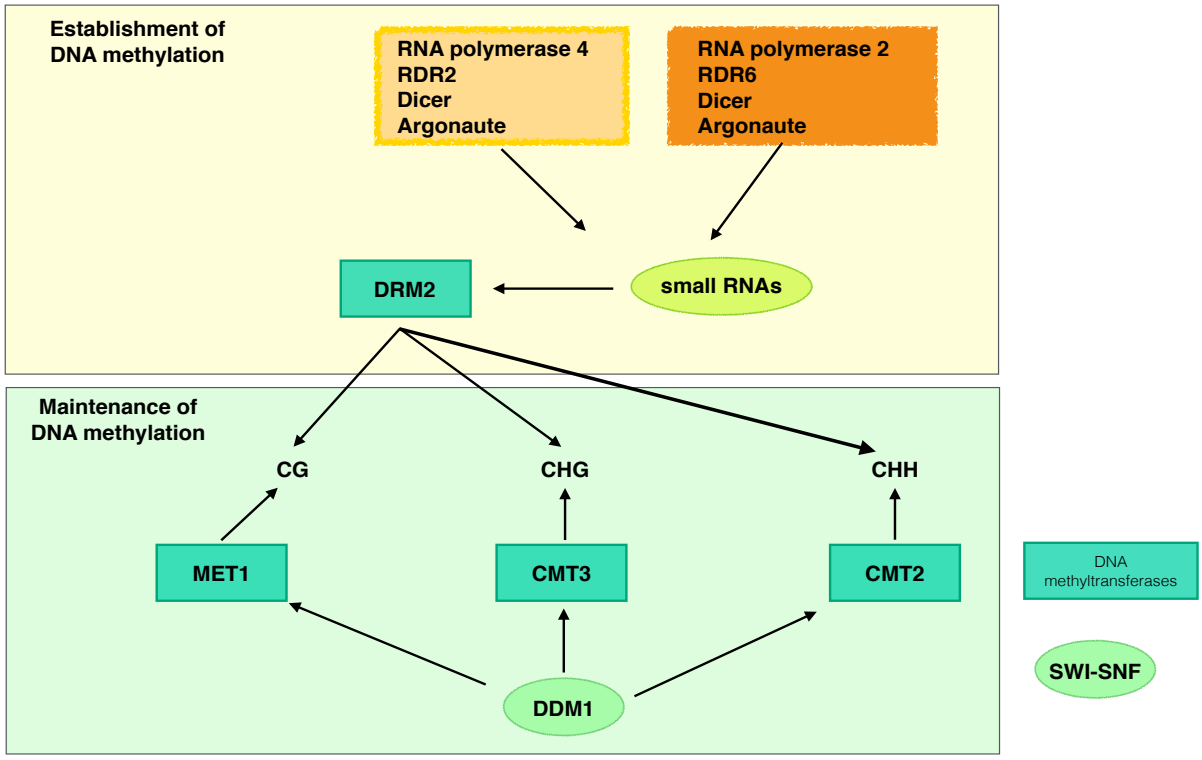


Figure I.5. Main epigenetic actors involved in DNA methylation maintenance and *de novo* mechanisms in *Arabidopsis thaliana*. DNA methyltransferases (green boxes) and chromatin remodeler (green bubble) are highlighted. See text for details. Adapted from Castel and Martienssen, 2013.

The RdDM pathway mediated by RNA polymerases Pol IV and Pol V acts on TEs that already contain methylation modifications to enhance the silent state of TE (Panda et al., 2016). Pol-IV allows the synthesis of transcripts matured into siRNAs via RDR2 (an

RNA-dependent RNA polymerase) that polymerizes the complementary strand and DCL3 (Dicer protein) that cleaves the duplex into 24-nt small RNAs. These siRNAs are then modified and loaded the AGO4 Argonaute protein. Pol-V dependent transcripts serve as a template for the pairing of the siRNAs carried by AGO4. Finally, DNA methylation is catalyzed by DRM2, allowing the deposition of repressive chromatin marks (H3K9me2) (Castel and Martienssen, 2013). In contrast, the silencing of transcribed or unmethylated TEs is mediated by RNA-dependent RNA polymerase 6 (RDR6) that produces 21-22 nt siRNAs triggering RNAi and *ab initio* DNA methylation, in a process known as RDR6-RdDM (Nuthikattu et al., 2013). RDR6-RdDM is essential for triggering silencing of active transposons. Previous studies have shown that initial cleavage of mRNA is a critical prerequisite for RDR6 recognition and that siRNA production is confined in the cell space to siRNA vesicles co-localized with RDR6 and SGS3. To investigate how RDR6 specifically recognizes transposon transcripts and selectively process siRNA, a recent study demonstrated that plant transposon RNAs contain non-optimal codons leading to a common ribosomal arrest during translation. That ribosomal arrest subsequently induces RNA truncation and localization to cytoplasmic siRNA vesicles (Kim et al., 2021).

1.1.4 Genome instability mediated by TEs

A major structural effect of TE insertion on the genome is that it can cause dramatic changes in chromosome structure both through the embedding of genes or gene fragments upon recombination and through the induction of chromosome rearrangements (Feschotte and Pritham, 2007). Chromosomal rearrangements may involve a number of mechanisms that affect DNA structure and play an important role in genome evolution. For instance, TE copy number contribute to variation in genome size within the genus *Oryza* (Piegu et al., 2006; Hurwitz et al., 2010); In the maize genome, translocations of *Ac* elements can lead to deletions, inversions, translocations or other rearrangements (Yu et al., 2012); This type of movement, can allow the insertion and duplication of genes or gene fragments into new chromosomal environments, sometimes altering their regulation, which may then lead to the emergence of new phenotypic features (Liu et al., 2016). These rearrangements may also lead to the formation of "island", such as a set of sequences with both genes and TEs, facilitating local adaptation (Turner et al., 2021).

1.1.5 TEs in the form of extrachromosomal circular DNA (eccDNA)

As the genetic material of life, DNA can be divided into linear form, such as genomic chromosomal DNA, and circular form. Circular DNA includes organelles (mitochondria and chloroplasts), most bacteria, and some viral genomic DNA. Extrachromosomal circular DNA (eccDNA) refers to extrachromosomal, non-organelle and circular structural DNA found in eukaryotes. It has been observed in many eukaryotic species for decades, including yeast, Drosophila, nematodes, plants and humans (Hotta and Bassel, 1965; Hirochika and Otsuki, 1995; Sinclair and Guarente, 1997; Cohen and Méchali, 2002; Cohen et al., 2006; Kumar et al., 2017). In plants, eccDNA is found to originate from tandem repeats (such as ribosome DNA copies or telomeric repeats) but also active TEs (Hotta and Bassel, 1965; Lanciano et al., 2017).

The discovery of eccDNA greatly predated the completion of the Human Genome Project, and no sequence analysis was performed, until recently. In the past few years, with the prevalence of high-throughput sequencing, eccDNA has been studied in a spurt and its role, notably in cancer cells, was revealed (Verhaak et al., 2019). Because of the huge variation in size: ranging from hundreds of base pairs to hundreds of kilobase pairs, extrachromosomal circular DNA has been classified into: 1) microDNA, which mainly refers to circular DNA within 400 bp (Shibata et al., 2012); 2) ecDNA (extrachromosomal DNA), which describes extrachromosomal DNA found in cancer cells that is hundreds of kb in size or more, and is large enough to comprise full-length genes and DNA replication initiation sites. EcDNA can replicate and amplify autonomously, and thus is associated with oncogene amplification and cancer development (Turner et al., 2017; Zhu et al., 2021); 3) In contrast, eccDNA refers to all extrachromosomal circular DNAs smaller than ecDNA.

Our laboratory previously developed eccDNA-seq (or mobilome-seq) to selectively sequence eccDNA from any plant or animal tissue (Lanciano et al., 2017, 2021). The method involves first digesting linear DNA using an ATP-dependent DNase and then enriching circular DNA by rolling cycle amplification. Circle-seq established in yeast (Møller et al., 2015) and CIDER-seq established in plants and virus (Mehta, 2020) are similar methods, increasingly coming to be used in the cancer field.

In cancer research, eccDNA is a key feature which can encode a variety of genes that promote tumor development (Kumar et al., 2017) and drug resistance (Yan et al., 2020). eccDNA is also able to promote the transcription of oncogenes through highly open chromatin and ultra-long-range regulation (Wu et al., 2019). Furthermore, oncogenes and their adjacent enhancers can be amplified as eccDNAs, suggesting that eccDNA plays a central role in accelerating tumor evolution. Notably, eccDNA-seq in neuroblastoma not only identified a variety of unidentified eccDNA, but also revealed that eccDNA is a major source of genomic rearrangement in somatic cells, revealing that eccDNA leads to oncogenic gene rearrangement through chimeric cyclization and reintegration into the linear genome (Koche et al., 2020). The fetal eccDNA detected in pregnant women's plasma can be used as a novel non-invasive molecular marker for prenatal testing, suggesting that eccDNA does not only play a significant role in tumor development, but may also be a highly promising molecular marker (Sin et al., 2020).

Given that a large number of TEs in the genome are activated during early embryonic development and are capable to reinsert and destabilize the genome, it is an important scientific question to understand how the embryo can avoid the damage caused by activated TEs. Wang et al., proposed that the activated TEs end up as eccDNA thus preventing their reinsertion in the genome (Wang et al., 2021c). With full-length sequence and genomic origin location information of more than 1.6 million eccDNA extracted from mouse embryonic stem cells, they proved that (1) eccDNA is randomly derived from chromosomal genomic DNA with no apparent location or sequence specificity; (2) eccDNA is a cyclization product of Lig3-mediated apoptotic DNA fragments; (3) and eccDNA has a superb ability to stimulate innate immune responses. (Wang et al., 2021c).

1.2 Obtaining a high-quality genome assembly prior to TE annotation

Note that the following part will serve as a basis for a review on genome assembly and structural variation in plants.

1.2.1 What is a genome assembly

In order to annotate the full picture of TEs in a given genome, it is necessary to obtain a high-quality genome assembly. The process of reconstructing the original genome from the millions of reads generated by high-throughput sequencing platforms from scratch is named *de novo* assembly. Some basic terms are involved: a contig refers to a long fragment formed by the assembly of multiple reads; a scaffold is a longer fragment formed by joining multiple contig sequences. Since the orientation and order of these contigs have been determined, the linkage between contigs is generally denoted by NNNN (Nagarajan and Pop, 2013; Compeau et al., 2011; Miller et al., 2011; Nagarajan and Pop, 2013; Sohn and Nam, 2018). Note that the overlap between reads is the core of the assembly algorithm. A graph which refers to a network of nodes (points) connected by edges (bridges), represents overlapping reads (Compeau et al., 2011; Rizzi et al., 2019).

Three criteria, namely completeness, correctness and contiguity are assessed to measure the quality of genome assembly (Gurevich et al., 2013; Mikheenko et al., 2018; Seppey et al., 2019). Completeness requires that the total length of the assembled sequence be as large as possible in proportion to the length of the genomic sequence; correctness requires that the assembled sequence conform as closely as possible to the true sequence; contiguity requires that the length of the sequences obtained by assembly is as long as possible, measured by the N50. N50 reflects the smallest contig length so that 50% of the entire assembly is contained in contigs equal to or larger than this value. If all contigs are sorted from longest to shortest (e.g., Contig 1, Contig 2, Contig 3,,, Contig 25), the added contig for which the total length of the contigs reaches 50% of the entire assembly, gives the N50. The higher the value of N50, the longer the contiguity of the assembly.

1.2.2 The revolution of long read sequencing technologies

Due to the fact that plant genomes are large, complex, and have a large number of repetitive regions, it is difficult to obtain high-quality genome assemblies (Michael and VanBuren, 2020). In particular, genome assembly is limited by short sequencing technologies and related assembly algorithms. The main challenges rely on (1) read length being very short compared to genome length, producing a puzzle with millions to billions of pieces; (2) lots of overlaps between reads due to short length making assembly ambiguous; (3) highly repeated regions causing difficulties in genome assembly. In addition, the insufficient sequencing depth, which refers to the ratio of the total number of bases obtained by sequencing to the size of the genome to be sequenced, will leave gaps in assembly algorithms (Alkan et al., 2011; Sohn and Nam, 2018).

In the last four years, long-read sequencing technologies offered by two fundamentally different platforms: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have emerged as a strong player in the genomics field. PacBio SMRT (single molecule real time sequencing) technology applies the idea of sequencing while synthesizing and uses the SMRT chip as the sequencing vector. In the base pairing stage, the addition of different bases (4 bases will be 4-color fluorescence labeled), will emit different light. According to the wavelength and peak of light, the type of bases entered can be determined (Eid et al., 2009). Of all the long reads generated by Pacbio sequencing, including continuous long reads (CLR) and cyclic consensus sequencing reads (CCS), as well as the latest high-fidelity (HiFi) reads, HiFi reads have the highest accuracy rate (99%). Nanopore sequencing technology is based on a special nanopore with covalent binding for sequencing. When DNA bases pass through the nanopore, they cause a change in charge that transiently affects the strength of the current flowing through the pore, and sensitive electronics detect these changes to identify the bases being passed (Ashton et al., 2015).

Both sequencing technologies generate single molecule reads longer than 10kb, exceeding the simplest repeat lengths in many genomes, enabling highly contiguous genome assembly. As a result, an increasing number of high-quality genomes of different species are being sequenced and assembled using long reads (Figure I.6).

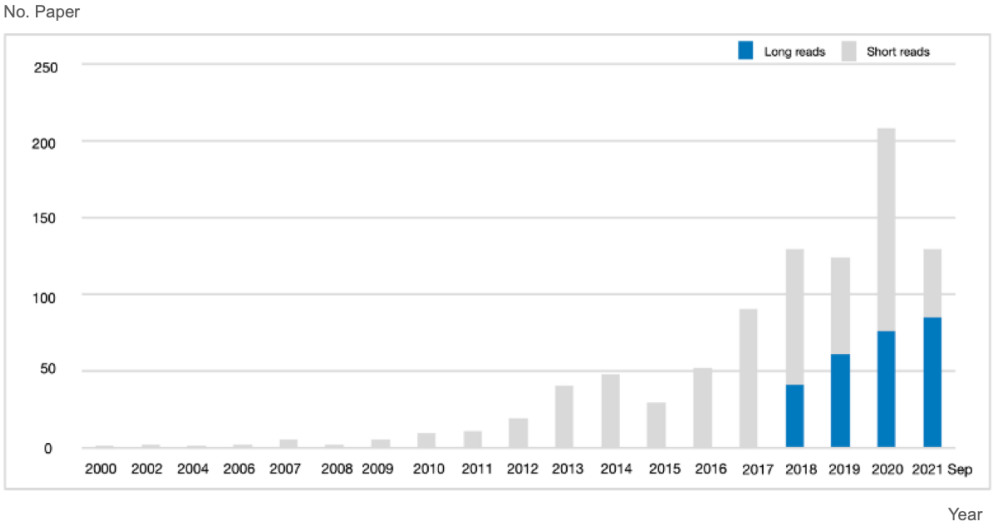


Figure I.6. Sequencing technologies used for plant genome assemblies until September 2021. The proportion of plant genome sequenced with long reads is colored in blue.

Advances in average read length, optimized assembly algorithms, and software have greatly contributed to the integrity and quality of genome assemblies. Gap filling, haplotype phasing, construction of very long genomes, and pangenome remain key breakthroughs for future plant genome assembly (Marx, 2021; Sun et al., 2021b) that will be introduced below.

1.2.3 Assembling centromeres: first « gap free » plant genomes

Gap free (also known as genome completion map) is the highest standard for genome assembly, and the construction of a gap free genome not only provides the most comprehensive reference genome information for population genetic studies and gene function localization, but also allows for structural and functional analysis of the centromeric and telomeric regions.

The *Arabidopsis thaliana* (ecotype Columbia-0 or Col-0) genome sequence was published in 2000, and after decades of research, the reference genome has become the "gold standard". However, the enrichment of highly repetitive sequence units in centromeric, telomeric and nucleolar organizing regions (NORs) has left these regions either with assembly errors or not sequenced. To obtain the telomere-to-telomere *A. thaliana* Col-0 genome, Wang et al. (2021) introduced the bacterial artificial chromosome (BAC)-anchored sequence substitution strategy into the Col-XJTU

genome assembly strategy, resolving the complete centromeric sequence of chromosomes 3, 4, and 5 and partially centromeric sequence of chromosomes 1 and 2. The *A. thaliana* Col-XJTU genome was assembled with high accuracy, and the sequencing quality score was significantly higher than that of TAIR10.1 (Wang et al., 2021a). To assess simultaneously genetic and epigenetic features of the centromeres, Naish et al. assembled a highly contiguous *A. thaliana* Col-0 genome using ultra-long reads generated by ONT, filling the gap of centromeres and providing an in-depth landscape of centromere evolution. The resulting Col-CEN assembly reveals the detailed architecture of the *A. thaliana* centromeres, i.e., the retrotransposon *ATHILA* interfering with the *CEN180* satellite arrays and DNA methylation inhibiting the meiotic DNA double stand breaks within centromeres. Thus, *A. thaliana* centromeres evolved under the opposing forces of satellite homogenization and retrotransposon interference (Naish et al., 2021).

In the construction of two gap-free rice genomes, Song et al. (2021) used high-depth Pacbio sequencing to assemble 0 gap genomes of ZS97 and MH63 (genome size 391.56Mb and 395.77Mb, respectively). Based on these gap free reference genomes, they investigated the structure and function of the centromeric region on 12 rice chromosomes in detail and found that the length of the core region of the centromeric region differed 10-fold on different chromosomes. In the ZS97 and MH63 centromeric regions, a total of 395 and 539 non-TE genes were identified respectively, but their transcriptional activity and the percentage of specific expression were low, and most of the actively transcribed genes were located in the peri-centromeric region. Additionally, they found that the similarity of CentO, which refers to a 155-bp satellite repeat (Dong et al., 1998), in the same chromosome was higher than that across chromosomes; the length of CentO satellite repeat sequences in the core region of the same chromosome differed significantly between varieties in the same subspecies (or natural population) of Asian rice (Song et al., 2021).

In general, due to the complex structure and large number of repetitive regions in the centromeric regions, the road to constructing a plant genome completion map is extremely winding, and often requires a combination of different sequencing platforms, different sequencing modes, and different assembly softwares. Using PacBio HiFi sequencing to take advantage of the results of different assembly softwares and adding

manual error correction may make the journey of constructing a plant genome completion map a little easier (Wang et al., 2021; Song et al., 2021).

1.2.4 The challenge of assembling very large plant genomes

Very large genomes usually have highly repetitive sequences, high heterozygous segments, and it is challenging to decipher their complete genome sequences.

For instance, garlic (*Allium sativum*) has a unique smell and high economic value, but its high heterozygosity and large genome size had hampered the characterization of its genome sequence. Combining PacBio, Nanopore, Illumina, 10X Genomics and Hi-C technology, Sun et al (2020) have constructed a chromosome-level reference genome of garlic with a genome size 16.24Gb. Comparative genome analysis showed that the root cause of the huge garlic genome is multiple whole genome duplication (WGD) events and rapid expansion of repeat sequences. Combined with the transcriptome data, they also established the allicin biosynthesis pathway and identified 4 genes related to the accumulation of garlic alliinase (Sun et al., 2020).

As a typical relict species, *Ginkgo biloba* is the only extant member of the Ginkgo family. The assembled genome assembled is 9.88 Gb in size (contig N50=1.58 Mb), and 27,832 protein-coding genes have been annotated. The intron length is the largest among the plant species studied so far, further suggesting that repetitive sequences not only facilitate genome expansion but also increase the size and complexity of protein-coding genes. Both genome and transcriptome studies helped understanding some of the ginkgo specific phenotypes, such as the preserved sperm flagellum, unformed flowers, and fan-shaped leaves, important features for environmental adaptations and gymnosperm evolution in Ginkgo (Liu et al., 2021).

Conifers are also known for their gigantic genomes. Xiong et al. (2021) constructed a reference genome at the chromosome level of southern *Taxus (Taxus chinensis* Rehd. *var*. *mairei*) with a genome size of 10.23 Gb (contig N50=2.44 Mb), using DNA extracted from endosperm call containing haploid chromosomes. Comparative genomic analysis showed that a WGD event occurred in the *Taxus* genus, and the unique families of *Gypsy* and *Copia* retrotransposons have expanded in this genus. *Taxus* conifers are used for the production of paclitaxel (Taxol for the

commercial name), a well-known anti-cancer drug. Combining genomic, transcriptomic and metabolomic data, the authors showed that the paclitaxel synthesis-related genes are organized and co-expressed as clusters. They further identified a gene cluster consisting of six genes in tandem that is responsible for the first two steps of paclitaxel biosynthesis (Xiong et al., 2021).

1.2.5 Resolving haplotypes in plant genomes

Genome haplotyping aims at reflecting the differences in allele composition between homologous chromosomes and is critical for genomic analyses for many plant and animal models, notably for polyploid organisms.

The pineapple strawberry (*Fragaria* × *ananassa*), which is widely grown around the world, is an octoploid complex genome formed by crossing the wild Virginia strawberry (*Fragaria virginiana*) with the Chilean strawberry (*Fragaria chiloensis*). The parents are wild-type octoploids and are derived from four diploid ancestral species. Edger et al. (2019) have used second- and third-generation sequencing techniques, combined with 10X Genomics and Hi-C, to construct a nearly complete genome (805.5 Mb) of the bromeliad strawberry, combined with transcriptome data, to provide a new basis for the evolutionary history of the origin of the octoploid strawberry. Transcriptome sequencing of 31 RNAseq datasets from four diploid strawberries and phylogenetic analysis of octoploid strawberries in combination with geographic distribution and historical evolution showed that octoploid strawberries originated in North America, and also indicated that *Fragaria iinumae, Fragaria nipponica, Fragaria viridis* and *Fragaria vesca* are the ancestral species of the octoploid strawberry. In addition, the evolutionary dynamics analysis of strawberry disease resistance genes (R genes) showed that TEs are closely related to R gene expression (e.g., *Fragaria vesca* subgenome has increased gene expression and its TE density is the lowest compared to the other three diploid strawberry species), leading to the identification of a dominant subgenome in strawberry (Edger et al., 2019).

Zhou et al. (2020) resolved the haplotyped genome of an heterozygous diploid potato using a complex strategy based on long read sequencing and genetic mapping (assembly size 1,67 Gb, N50=2Mb). Briefly, they produced two assemblies using (1) ONT data and 10xGenomics and (2) PacBio HiFi reads, respectively. For each

assembly, the scaffolds were assigned to the 24 genetic groups (2n=24) thanks to the resequencing data of an F2 population. Then the authors used Hi-C data to perform a scaffolding step combining the two assemblies into a final one. Comparative analysis of the two sets of haplotyped sequences revealed the presence of more than 20,000 deleterious mutations in the diploid potato, with 16.6% of alleles differentially expressed and 30.8% differentially methylated. These mutations are dispersed in the genome. The authors located on the phased genome several loci involved in inbreeding depression and displaying segregation distortion. For instance, the seedling albino gene (white seedling *WS1*) and the plant architecture gene (*PA1*) are two linked genes on chromosome 1. The deleterious genotypes of these two genes (*ws1* and *pa1*) are located on two different haplotypes and closely linked to the « normal » genotypes (*WS1* and *PA1*) with very low segregation probability in the offspring (2 recombinants out of 1200 screened F2 plants). The phased genome thus offers the possibility to improve breeding in this clonally propagated plant. Importantly, the authors highlight the fact that despite the use of long-reads and HiC data, the availability of genetic data was instrumental in the determination of this large haplotype-resolved genome (Zhou et al., 2020).

From this knowledge, Zhang et al. (2021) established a genomics-assisted breeding design for hybrid potato that includes four steps: (1) selecting two heterozygous lines of phenotypic interest and breaking self-incompatibility; (2) analyzing the genetic composition of segregating progeny; (3) producing inbred lines with favorable haplotypes, counter-selecting haplotypes with deleterious mutations; (4) crossing the obtained homozygous lines to obtain F1 plants with hybrid vigor (Zhang et al., 2021a).

Despite the two above-mentioned genome assemblies of diploid potato, the homozygous tetraploid genome of cultivated potato had not been assembled and genomic haplotyping of the tetraploid potato remained a challenge. In a recent BiorXiv study, inbred homozygous tetraploids were successfully haplotyped. In this analysis, the haplotigs (contigs representing only one haplotype) were determined using a large dataset of diploid gamete sequences (717 short-reads low-coverage single-cell sequencing of pollen cells), reasoning that from this large number the underlying sets of haplotypes could be uncovered. Then HiFi reads were assigned to each haplotig and HiC data was used for scaffolding. The authors noted that when assembling the long reads first, around one third of the haplotypes were collapsed and could not be

subsequently untangled. This analysis revealed that 50% of the tetraploid genome is fragmentally identical in at least two haplotypes. This high level of inbreeding contrasts with the extreme structural rearrangements found in about 20% of the genome and enriched for retrotransposons. In addition, 148,577 gene were annotated, of which only 54% were present in all four haplotypes, with an average of 3.2 copies for each gene (Sun et al., 2021a), reinforcing the importance of obtaining phased genomes.

Today, the improvement of technology and algorithms has greatly improved our capacities to address the challenge of obtaining phased plant genomes with a high sequence resolution. In this context, HiFi reads have proved to be more promising for haplotype-resolved assembly than ONT reads due to their accuracy. Indeed, the combination of high read length and improved base accuracy is a game changer. One current way to obtain phased information through HiFi reads is: (1) to use HiFi reads to sequence a single individual; (2) to use diploid genome assembly softwares, such as hifiasm (Cheng et al., 2021) or HiCanu (Nurk et al., 2020) for genome assembly; (3) to use softwares including Google DeepVariant (Poplin et al., 2018) for mutation detection, and use WhatHap (Patterson et al., 2015) for haplotype phasing; (4) to combine HiFi data with other technologies such as Hi-C or Strand-seq to extend haplotype phasing to chromosomes, enabling phase analysis of the entire genome. Eventually, if there are three samples available from both parents and offspring, before the genome is assembled, the short read data from the parents can be used to phase the HiFi data into the respective parental data (Sun et al., 2020; Garg, 2021; Sun et al., 2021a; Zhang et al., 2021a).

1.2.6 Pan-genomes: One genome is not enough

In the last two years, the comparative analysis of genomes or genome fragments of multiple individuals of the same species has shown that a single reference genome is not enough to capture the genetic diversity of a species (Bayer et al., 2020). These findings indicate that the genome within a species may differ in more significant ways, including the diversity of structural variants (SV), and these variants may contain one or more genes. A large number of studies have shown that SVs play a key role in important agronomic traits, such as resistance to biotic and abiotic stress, flowering time, plant architecture, yield, grain or fruit quality (for a review see Tao et al., 2019). These results imply that the functional gene content of a species is more variable than previously thought. Therefore, for a species, if only a single reference genome is used

for the study of genetic domestication and/or selection, a lot of meaningful genetic information may be lost. The above factors have jointly promoted the construction and research on plant and animal pan-genomes.

Pan-genome is a general term for all genes of a species, where the whole genes are distinct from those of the individual genome (Tettelin et al., 2005). The genome of an individual is not representative of the genome of the species. Therefore, the analysis of a genome as a reference does not provide a complete picture of the genetic information of a species at the gene level, especially when studying different subspecies or variants of the same species, where the differences in such unique segments are often more important than those in the shared segments. The analysis of core and non-core genes is fundamental to the study of within-species variation from the perspective of unique gene sequences (Bayer et al., 2020).

Since a single reference genome cannot represent the entire sequence diversity within a species, population-scale long-read sequencing has gradually begun to flourish in evolutionary and functional genomics research and crop breeding research (Figure I.7). It has been conducted in various model and crop plants including *A. thaliana* (Jiao and Schneeberger, 2020), tomato (Gao et al., 2019; Alonge et al., 2020), rice (Qin et al., 2021), soybean (Liu et al., 2020b), rapeseed (Song et al., 2020; Chawla et al., 2021), wheat (Walkowiak et al., 2020), barley (Jayakodi et al., 2020) and maize (Hufford et al., 2021)

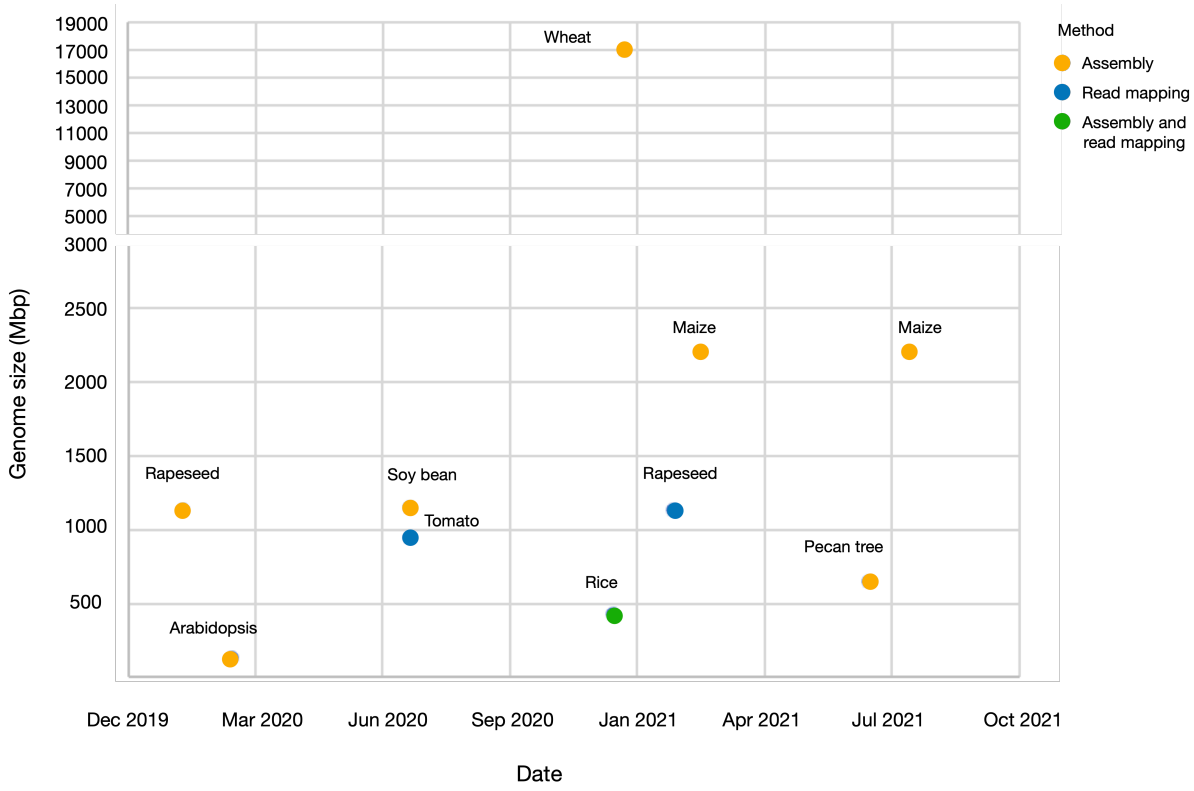


Figure I.7. An overview of genomic population-scale studies in plant using long-read sequencing. Construction strategies based on assembly comparison and/or read mapping are color-coded as indicated. See text for details.

Although 1001 *Arabidopsis* genomes have been sequenced since 2016 (Alonso-Blanco et al., 2016), the degree of genomic variation within this species is still poorly understood due to the small number of chromosome-level assemblies. Jiao and Schneeberger's study provided chromosome-level reference assemblies of seven *Arabidopsis* germplasms, selected from across the globe. In each genome rearrangements of 13-17 Mb in length were detected, as well as 5-6 Mb of non-reference sequences, causing copy number variations (CNVs) in approximately 5,000 genes (including approximately 1,900 non-reference genome-containing genes). Quantifying the variability between genomes revealed approximately 350 autosomal regions where tandem duplications had occurred. Interestingly, these rearrangement hotspot regions are enriched in genes associated with biotic stress response and display reduced meiotic recombination in hybrids. This suggests that the rearrangement hotspots have undergone differential evolutionary dynamics compared to the rest of the genome, largely based on the accumulation of new variants rather than on recombination of existing variants, allowing for a rapid response to biotic stresses (Jiao and Schneeberger, 2020).

The rapeseed pan-genome of 1.8 Gb was constructed based on the genome sequences of 8 *Brassica napus* germplasms assembled on PacBio, HiC, BioNano platforms, containing about 150,000 genes (Song et al., 2020). Genome-wide association studies (GWAS) based on presence and absence variations (PAVs) identified previously undiscovered differences in traits caused by TE insertion, and differential expression of genes related to different traits caused by PAV among different ecotypes. The study focused on three types of SVs in the *FLC* gene, involved in flowering time and vernalization. PAV-GWAS peaks corresponded to insertions of a hAT transposon in a *BnaA02.FLC* exon and the *BnaA10.FLC* promoter that correlated with early and late flowering, respectively. The genome assemblies further revealed four TE insertions polymorphisms at the *BnaC02.FLC* locus. Interestingly, the haplotypes of the TEs were more consistent with ecotype information and flowering time than the haplotypes of the SNPs. A LINE insertion in the first exon of *BnaA10.FLC* in spring oilseed rape leads to a loss of function, and spring rape needs weak or no vernalization to flower. In contrast, a MITE insertion in the *BnaA10.FLC* promoter up-regulates *FLC* expression, and winter rape requires stronger vernalization to bloom. In conclusion, this study revealed the molecular basis of winter and spring flowering regulation through pan-genome and PAV-GWAS approaches (Song et al., 2020).

The rice pan-genome was constructed based on 31 high-quality genome assemblies and two published rice genomes, Nipponbare and Shuhui 498 (Qin et al., 2021), identifying 171,072 SVs and 25,549 gene copy number variations (gCNVs). Detailed studies on the mechanism of SV formation, the effect of SV on gene expression, and the distribution of SV among subpopulations were conducted, demonstrating how SVs and gCNVs affect environmental adaptation and domestication in rice. Especially most of genomic variants had not been found in previous studies that used traditional short read methods, but they play an important role in the regulation of important agronomic traits. For example, the tandem duplication of the *OsMADS18* gene in *Koshihikari* variety was identified by long reads. Considering that increased expression of *OsMADS18* has been shown to cause early flowering, it can be inferred that the duplication of *OsMADS18* may be the cause of the early flowering phenotype of this variety. In addition, the graphical pangenome-based SV-GWAS identified many phenotype-related genetic variants that could not be detected when using only SNPs and single reference combinations.

Maize is the most widely grown crop in the world and an important model system for studying gene function with high genetic diversity. A pan-genomic analysis was performed on maize nested association mapping (NAM) populations. A total of 26 lines from 25 NAM populations and B73 were selected to construct a pan-genome of maize containing 100,000 genes with only one-third present in all 26 maize lines (Hufford et al., 2021). GWAS analysis based on SNP and SV showed that 93.05% of SNP and SV loci overlapped with each other. However, SV-GWAS, but not SNP-GWAS, identified an association locus for blast disease on chromosome 10, indicating that combining SNP-GWAS and SV-GWAS could improve the accuracy of trait-gene association (Hufford et al., 2021).

1.3 Structural variants in the plant genome

1.3.1 Why to detect structural variants (SVs)?

Variants are the most important factors leading to genomic differences and can be specifically classified as single base pair variants, small insertions or deletions, and structural variants (Mérot et al., 2020). Single nucleotide variants, often called single nucleotide polymorphisms (SNPs), are differences in individual DNA bases. Small indels (short for insertion and deletion) refer to the insertion or deletion of a small fragment of sequence, usually under 50 bp in length. Structural variants (SVs) refer to the alteration of chromosomal fragments different from the reference genome, usually larger than 50bp. The main types of SVs include insertion deletion, duplication, inversion and translocation (Carvalho and Lupski, 2016). Inversion is when a large segment of DNA is reversed compared to the reference genome. Translocation is when a large segment of DNA is moved out of one place and inserted into another (Stankiewicz and Lupski, 2010).

Importantly, SVs are closely related to a number of key agronomic or breeding-related traits and the basis for crop improvement and domestication. Multiple SVs detected in 100 tomatoes were able to alter gene dosage and expression levels, resulting in changes in taste, size, and yield traits (Alonge et al., 2020). Zhou et al. (2021) identified a chromosomal inversion of 1.67 Mb in the flat peach genome located approximately 3 Kb downstream of the *PpOFP1* gene stop codon (a gene of the *OVATE* family involved in transcriptional repression). This chromosomal inversion, which is not present in common peaches, is responsible for the change from round to flat peaches (Zhou et al., 2021). In addition, several studies have shown that structural variation can better resolve population structure and provide further valid information to gain insight into plant domestication processes (Alonge et al., 2020; Hufford et al., 2021; Qin et al., 2021).

1.3.2 Algorithms of SV detection

With the rise of second and third-generation sequencing, the throughput of SV detection has started to improve. The SV detection algorithms rely on read depth, short-read pairs, long read alignment or *de novo* assembly, and these four types of algorithms

have some differences in detection accuracy and the range of detectable structural variation (Figure I.8) (Mahmoud et al., 2019; Mérot et al., 2020).

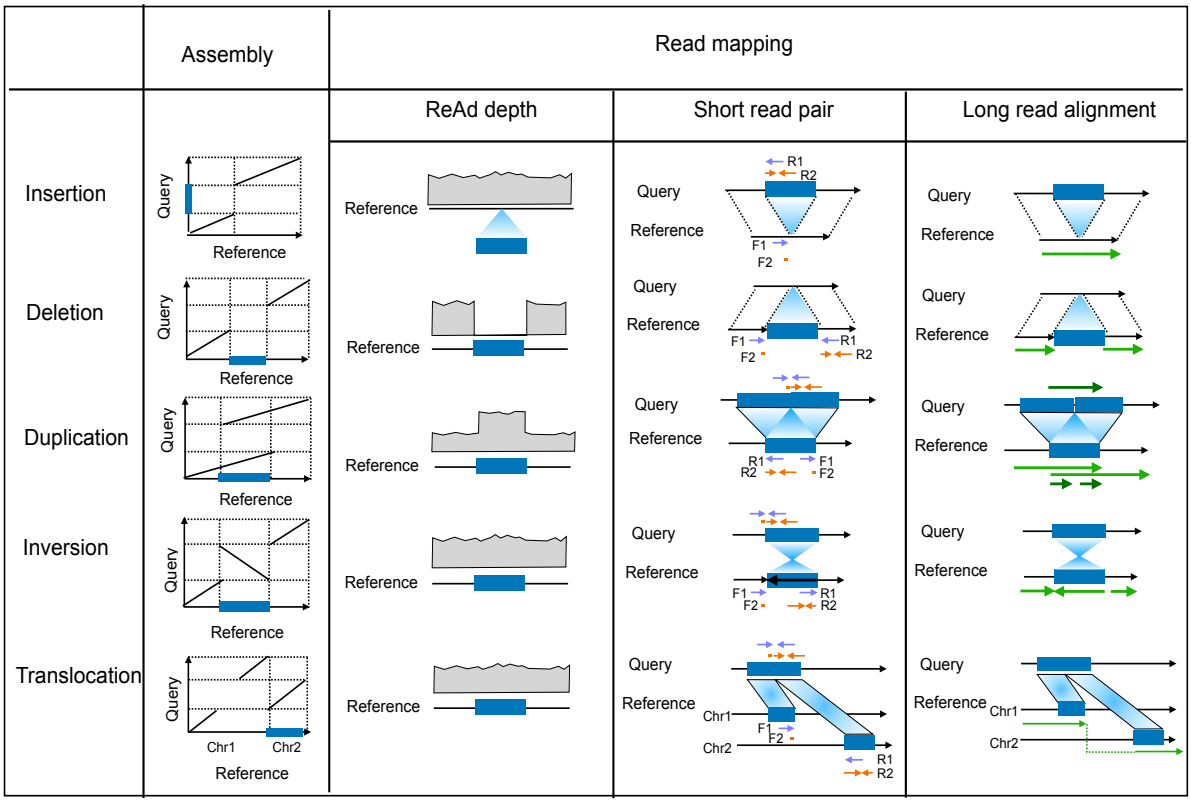


Figure I.8. Detecting structural variants using *de novo* assembly and read mapping modes. In *de novo* assembly mode, segment positions in the dot plot of query and reference sequence comparison indicate the type and size of the SV. In read mapping mode, paired reads (orchid) and split reads (orange) from short read alignment patterns are typically used to detect different types of SV, as indicated. For long reads (green), alignment patterns across junctions are typically used to detect different types of SVs. In addition, read depths showing coverage aberrations can be used to improve the accuracy of deletion and duplication detection (modified from Mahmoud et al., 2019; Mérot et al., 2020).

Read depth method assumes that sequencing reads are randomly distributed (e.g., Poisson distribution). Duplicated and missing intervals are calculated primarily by the longitudinal coverage of sequencing reads within a specified region. The duplicated intervals have a higher depth, while the missing intervals have a lower sequencing depth. This method is commonly used for the detection of genomic copy number variation (CNV) as read depth of a genomic region is correlated with its copy number (Miller et al., 2011).

The short read-pair method is based on the distance between the paired-end reads and their direction. Paired-end reads are mapped to the reference genome to identify pairs of abnormal reads, and then the abnormal regions are extracted. Based on the location, size, and number of abnormal reads, the corresponding SVs are determined. Split read pair allow to uncover sequences that may come from different intervals of the genome (Cameron et al., 2019; Mahmoud et al., 2019).

This principle was applied to characterize new TE insertion polymorphisms (TIPs). On the base of the case where one of the two paired-end reads can match to a certain TE normally, but the other cannot, and the unmatched read maps to the reference genome, many tools have been developed to detect TE insertions from short-read sequencing, such as ngs_te_mapper (Linheiro and Bergman, 2012), RelocaTE (Robb et al., 2013), PoPoolationTE2 (Kofler et al., 2016, 2), TRACKPOSON (Carpentier et al., 2019) and TEMP2 (Yu et al., 2021).

However, all the methods mentioned above are actually limited by the fact that the reads are too short. Because the reads are too short, they cannot span genomic repeat regions during comparison; nor can they capture many large insertion sequences. To overcome the limitations of tools developed with short sequencing reads, many tools emerged from the use of long-read sequencing data, such as cuteSV (Jiang et al., 2020), Sniffles2 (Sedlazeck et al., 2018; Smolka et al., 2022) and pbsv (https://github.com/PacificBiosciences/pbsv). However, introducing long sequences requires considering the relatively high sequencing price, hammering its wide application to population genomics. Ideally, long-read based *de novo* assembly should be the most efficient method for genomic SV detection, as it can detect all types of SVs (Mahmoud et al., 2019; Mérot et al., 2020).

1.3.3 Visual validation for SV prediction

Visual validation is an important step in reducing false positives for structural SV prediction. To visualize SVs from read mapping, Belyeu et al. have developed Samplot, a tool that visualizes read depth and sequence alignment pairs for predicting so-called SV across regions (Belyeu et al., 2021). Samplot is applicable to many biological problems such as SV prioritization in disease research, genetic variation analysis or ab initio SV prediction. It includes a machine learning package that significantly reduces

the number of false positives without eye review. Tools such as the Integrative Genomics Viewer (IGV) are also useful for relatively small regions, allowing accurate SV detection at the base pair level (Robinson et al., 2011). However, IGV is very limited in displaying the alignment of large regions, slowing down the process of visual verification.

Dot plot is a gold standard tool to detect SVs between genomes, including D-GENIES (Cabanettes and Klopp, 2018), shinyChromosome (Yu et al., 2019) and ggplot2 mummerplot (https://jmonlong.github.io/Hippocamplus/2017/09/19/mummerplots-with-ggplot2/). Another popular tool is Syri (Goel et al., 2019). Methods used in the rice pan-genome provide a good example (Qin et al., 2021). All 32 rice assemblies were matched against the IRGSP1.0 reference genome using Mummer4 (Marçais et al., 2018). The raw match results were further filtered using delta-filter and then subjected to detect SV using Syri. Then, based on the results of Syri, Qin et al. further subdivided the SVs into three major categories of resultant variants: presence/absence variants, inversions, and translocations.

JBrowse 2 (Buels et al., 2016) provides a superior visual review of SVs in read mapping and genome assembly, including modes for linear and circular genome view. It is able to visualize and analyze files in different formats using synteny analysis, dot plot, and SV inspector mode, which greatly reduces false positive SVs. There is no doubt that there are and will be many tools to visual validate SV detection. A user-friendly usage and high-quality figures produced are always welcome and the tools above are highly recommended.

1.4 Objectives of the thesis work and main achievements

The general objective of my work was to characterize the interplay between the eccDNA compartment and the genome. For this I focused on three aspects and technological developments.

First, despite the rapid development of high-throughput sequencing and the continuous updating of eccDNA-seq, the landscape and dynamics of eccDNA are poorly understood due to the lack of dedicated computational tools. To reach this objective, my first project was to fill the gap for tools dedicated to eccDNA detection, notably using long-read sequencing data. I thus developed ecc_finder, a new tool specifically for eccDNA detection from long-read data of eccDNA-seq. In addition, I optimized the algorithm for eccDNA detection from short-read data, previously developed in the group, to improve its computational performance. We tested the pipeline using *Arabidopsis thaliana* and wheat (genome sizes of 125Mb and 17Gb, respectively). To further evaluate the accuracy of the developed tool, datasets of experimentally validated eccDNA were be used.

Second, to understand the impact of TEs on the genome, e.g. TE polymorphisms, a high-quality of genome assembly is required. However, different assemblies produced by different assemblers or the same assembler with different parameters have different performances. The best assembly with high contiguity and high repetition resolution cannot be achieved in one single assembly. My second project consisted in developing SASAR, a meta-assembly tool to reconcile the result of different long read assemblies. In order to reconcile multiple assemblies and to resolve structural variants with accuracy, I used the strategy of assembly graph.

Finally, thanks to the ability to detect the dynamics of eccDNA and to obtain high quality genome assemblies with long read sequencing, I addressed the question of the impact of a high load of eccDNA on the genome structure. My third project was thus to explore the structural variants in *A. thaliana* epigenetic mutants with a high load of eccDNA. For this I detected eccDNAs in long-read eccDNA-seq and SVs in long-read assembled genomes, respectively.

2. Methods and Results

2.1 Ecc_finder: Developing a new tool for eccDNA detection

2.1.1 My contribution to ecc_finder

Our laboratory previously developed eccDNA-seq (or mobilome-seq) to selectively sequence the eccDNA form of active TEs in order to characterize the mobility of TEs in any plant or animal tissue (Lanciano et al., 2017a). To get familiar with this type of data, I started training on a dataset of *Arabidopsis thaliana* treated with heat stress. The method was based on a bash script developed by the former PhD student in the lab who developed eccDNA-seq. After examining the results generated at each step, I understood that this bash script was based on screening read coverage of short read mapping.

However, its subsequent steps were not standardized: firstly, the peak calling method was not precise and did not yield the exact boundaries of the eccDNA loci. Secondly, it is worth noting that although our experimental step was to obtain circular DNA enrichment by digesting linear DNA, high coverage of genomic loci was not always caused by circular DNA, and duplicated regions could also have a high coverage. Therefore, it took me a long time to manually filter the output of this bash script and make it accurate. Considering that our laboratory collaborates with different institutes to provide eccDNA-seq sequencing for many plants and animals, there was a need to automate the eccDNA detection process and achieve a greater capacity, for example at the population level.

Therefore, I set out to develop a new detection algorithm specifically for circular loci. Lacking relevant experience in software development, I deepened my understanding by interning in a lab that specialized in developing short-read aligners in 2019 (lab of Peter Stadler, who developed segemehl). Through careful screening of the read alignment and positive controls validated by the wet lab experiments, I finally understood that there were specific types of split and discordant reads.

I gradually wrote the entire pipeline in Python after being back from my internship. Although it eventually worked, it still needed improvements to make it more fluid. What's more, I had noticed that sequencing eccDNA using long reads had started to emerge from 2020 (Koche et al., 2020), and had become a hot topic in cancer research. On the basis of my experience writing the pipeline for short reads, I further detected the

characteristics of reads originating from circular loci from long read alignment on a dataset of *Arabidopsis thaliana*.

In order to produce a comprehensive tool that could be used for both short-read and long-read eccDNA-seq, I reasoned that more datasets were needed to support the normalization of the tool. I proposed a collaboration with the laboratory of Etienne Bucher (member of the EpiDiverse consortium) that was performing eccDNA-seq on the very large genome (17Gb) of wheat. Knowing that the larger the genome, the stronger the computational memory needed, I added an option to the aligner to decrease the computation time. A complete framework for the detection tool, which I called ecc_finder, was finally developed.

I contributed to the manuscript by writing a draft on the Results and Discussion sections and describing the usage and commands of ecc_finder on the wiki page on github. Our collaborator Haoran Peng provided us with the long-read sequencing data of wheat that I analyzed with ecc_finder. During the reviewing process I completely rewrote the Methods section and updated the wiki page on github and the ecc_finder manual accordingly. I further uploaded 5 videos of relevant commands on YouTube for the public without bioinformatics background. Finally, the software ecc_finder was published in Frontiers in Plant Science 2021.

2.1.2 ecc_finder manuscript (Frontiers in Plant Science, 2021)

frontiers in Plant Science

METHODS
published: 01 December 2021
doi: 10.3389/fpls.2021.743742

ecc_finder: A Robust and Accurate Tool for Detecting Extrachromosomal Circular DNA From Sequencing Data

Panpan Zhang[1,2], Haoran Peng[3,4], Christel Llauro[2,5], Etienne Bucher[3] and Marie Mirouze[1,2]**

[1]Institut de Recherche pour le Développement (IRD), Montpellier, France, [2]Laboratory of Plant Genome and Development, University of Perpignan, Perpignan, France, [3]Crop Genome Dynamics Group, Agroscope Changins, Nyon, Switzerland, [4]Department of Botany and Plant Biology, Section of Biology, Faculty of Science, University of Geneva, Geneva, Switzerland, [5]Laboratory of Plant Genome and Development, Centre National de la Recherche Scientifique (CNRS), Perpignan, France

Extrachromosomal circular DNA (eccDNA) has been observed in different species for decades, and more and more evidence shows that this specific type of DNA molecules may play an important role in rapid adaptation. Therefore, characterizing the full landscape of eccDNA has become critical, and there are several protocols for enriching eccDNAs and performing short-read or long-read sequencing. However, there is currently no available bioinformatic tool to identify eccDNAs from Nanopore reads. More importantly, the current tools based on Illumina short reads lack an efficient standardized pipeline notably to identify eccDNA originating from repeated loci and cannot be applied to very large genomes. Here, we introduce a comprehensive tool to solve both of these two issues.[1] Applying ecc_finder to eccDNA-seq data (either mobilome-seq, Circle-Seq and CIDER-seq) from *Arabidopsis*, human, and wheat (with genome sizes ranging from 120 Mb to 17 Gb), we document the improvement of computational time, sensitivity, and accuracy and demonstrate ecc_finder wide applicability and functionality.

Keywords: eccDNA, nanopore amplicon sequencing, mobilome, *Arabidopsis*, wheat

OPEN ACCESS

Edited by:
Agnieszka Zmienko, Institute of Bioorganic Chemistry (PAS), Poland

Reviewed by:
Vladimir A. Trifonov, Institute of Molecular and Cellular Biology (RAS), Russia
Xiukai Cao, Yangzhou University, China

****Correspondence:***
Panpan Zhang
panpan.zhang@ird.fr
Marie Mirouze
marie.mirouze@ird.fr

Specialty section:
This article was submitted to Plant Systems and Synthetic Biology, a section of the journal Frontiers in Plant Science

Received: *19 July 2021*
Accepted: *25 October 2021*
Published: *01 December 2021*

Citation:
Zhang P, Peng H, Llauro C, Bucher E and Mirouze M (2021) ecc_finder: A Robust and Accurate Tool for Detecting Extrachromosomal Circular DNA From Sequencing Data. Front. Plant Sci. 12:743742. doi: 10.3389/fpls.2021.743742

INTRODUCTION

Circular DNA is a ubiquitous form of biological DNA molecules. Indeed, it can be found as bacterial, viral, mitochondrial, and chloroplastic genomes and plasmids, but also as extrachromosomal circular DNA (eccDNA) in eukaryotes (Hotta and Bassel, 1965). eccDNA has been described for decades in yeasts (Sinclair and Guarente, 1997), *Drosophila* (Cohen et al., 2003), mammals (Cohen et al., 2006; Kumar et al., 2017), and plants (Hirochika and Otsuki, 1995). Recently, the role of eccDNA as an important genomic feature of cancer cells has been revealed. Indeed, in cancer cells, eccDNA molecules arise from chromosomal oncogenes inducing their overexpression (Kumar et al., 2017) and are associated with poor prognosis (Verhaak et al., 2019; Kim et al., 2020; Wang et al., 2021) and drug resistance (Yan et al., 2020). In plants, genes located in eccDNA molecules can be overexpressed leading to herbicide resistance (Koo et al., 2018).

[1]https://github.com/njaupan/ecc_finder

Besides genes, eccDNA can arise from repetitive genomic sequences, such as telomeric DNA (Cohen and Méchali, 2002; Zellinger et al., 2007; Mazzucco et al., 2020), satellites (Navrátilová et al., 2008), or ribosomal RNA genes (rDNA; Sinclair and Guarente, 1997) through homologous recombination. Moreover, eccDNA is part of the life cycle of certain types of active transposable elements (TEs; Hirochika and Otsuki, 1995; Lanciano et al., 2017). The presence of eccDNA thus generally reflects genome plasticity. We previously developed Mobilome-seq (Lanciano et al., 2017, 2021) as a method to selectively sequence eccDNA purified from plants or animal tissue. The method is based on two main steps: (1) linear DNA digestion using an ATP-dependent DNase followed by (2) eccDNA enrichment by random rolling circle amplification. Several similar methods have been established to enrich and detect eccDNA molecules, such as Circle-Seq (Møller et al., 2015) and CIDER-seq (Mehta et al., 2020). With the arrival of single-molecule real-time sequencing by Pacific Biosciences and nanopore sequencing by Oxford Nanopore Technologies (ONT), eccDNA sequencing with long reads allows capturing comprehensive eccDNA content by spanning the full length of eccDNA in one read (Koche et al., 2020). However, following short- or long-read sequencing, only a handful of bioinformatic tools has been developed for the downstream analysis of eccDNA data. CIDER-Seq2 is the only tool based on PacBio long reads alone. AmpliconArchitect (Deshpande et al., 2019), Circle-Map (Prada-Luengo et al., 2019), Circle_finder (Kumar et al., 2017), and ECCsplorer[2] are tools based on Illumina short reads. Moreover, except for ECCsplorer, all software packages require a reference genome, thus limiting the analyses to model species (**Figure 1**).

Here, we developed a new tool called ecc_finder dedicated to the detection of eccDNA from both Illumina and Nanopore eccDNA sequencing data. We demonstrate its suitability and sensitivity when applied on eccDNA data sets originating from small (*Arabidopsis thaliana*, 120 Mb) and very large genomes (wheat *Triticum aestivum*, 17 Gb) for detecting eccDNAs.

MATERIALS AND METHODS

Description of the ecc_finder Algorithm and Validation Metrics

The complete ecc_finder source code and documentation are available on GitHub at https://github.com/njaupan/ecc_finder. ecc_finder is written in Python3 and relies on two mapping tools: minimap2 (Li, 2018) for ONT long reads and BWA (Li and Durbin, 2009) for Illumina short reads. It also bundles TideHunter (Gao et al., 2019) for discovering tandem repeat patterns and generating high-quality consensus sequences and Genrich[3] for peak calling. ecc_finder mainly uses the PAF format generated by paftools.js to realize format conversion and alignment filtering.

Long-Read Pipeline Algorithm Overview

Thanks to the Phi29 rolling circle amplification of eccDNAs, the matrix for long-read sequencing comprises tandem repeats of the original eccDNA sequence. Therefore, reads originating from circular DNA will display two or more sub-read alignments to the reference in the same direction. ecc_finder thus uses a tandem repeat pattern detection from read alignments to identify candidate loci. First, to exclude long reads originating from linear genomic repeats (such as satellites), ecc_finder performs standard alignment to extract alignment block length using minimap2. By default, ecc_finder will remove any alignment shorter than

[2]https://github.com/crimBubble/ECCsplorer

[3]https://github.com/jsh58/Genrich

	Amplicon Architect	Circle-map	Circle_finder	CIDER-seq2	ECCsplorer	ecc_finder
Suitable for short reads	✓	✓	✓	✗	✓	✓
Suitable for long reads	✗	✗	✗	✓	✗	✓
Handle giant genome	✗	✗	✓	✓	✗	✓
Reference free	✗	✗	✗	✗	✓	✓
Consider repeated loci	✗	✓	✗	✗	✗	✓

FIGURE 1 | Summary of the characteristics of up-to-date eccDNA detection tools.

200 bp (**Supplementary Figure 4**). ecc_finder then uses TideHunter to identify candidate reads with a tandem repeat pattern and divide each read into repeat units (**Supplementary Figure 4**). Any read that do not have two or more repeat units or in which the divergence rate between repeat units exceeds 25% will be discarded. ecc_finder then uses minimap2 to map these selected reads to a reference genome. Only loci displaying more than two reads coverage are selected. The *value of p* for each base of the genome is calculated assuming a null model with a log-normal distribution given by Genrich. ecc_finder sets the enriched genomic locus as the reference boundary and applies bedtools groupby to calculate the number of tandemly repeated reads, sub-alignments, and boundary coverage (**Supplementary Figure 1**). Loci covered by a minimum of three reads are kept. Finally, detected loci that are covered for at least 80% of their length are retained.

Short-Read Pipeline Algorithm Overview

Ecc_finder uses a standard method based on discordant pairs and split reads at the junction to detect reads originating from circular DNA in short-read sequencing data (**Supplementary Figure 2**). ecc_finder uses BWA-MEM as the default mapping software for short reads because it is more accurate at the basic level alignment, but users can still choose minimap2 with the short-read parameter "sr" to speed up the alignment for large genomes. By grouping by chromosome and read ID, the read alignments are sorted and merged to remove the overlapping reads between a pair, and ecc_finder extracts the read pair information and read pair direction. Properly mapped read pairs with inward-facing tags ("–>, <–") or single mapped reads ("–> / <–") are discarded, while discordant read pairs with outward-facing tags ("<–, –>") are kept. For split reads, ecc_finder then selects read pairs with 3 unique hits on the same chromosome, with orientations suggesting a circular template such as ("–>, <–, –>") and ("<–, –>, <–"). The enriched genomic sites extracted with Genrich are set as the reference boundaries to group split reads and discordant reads. Only the reads spanning the same boundaries are kept. Loci covered by a minimum of two split reads and one discordant read pair are kept. Finally, the *bona fide* eccDNA-producing loci are defined as regions displaying an even distribution of split and discordant reads (**Supplementary Figure 2**). In addition, ecc_finder benchmarked BWA-MEM and segemehl to access the accuracy of different short-read aligners. Compared with BWA, segemehl requires more computing time for indexing and aligning, as well as high storage capacity, especially for large genomes (**Supplementary Figure 3**). In the *Arabidopsis* samples, ecc_finder did not find any difference for eccDNA detection using either segemehl or BWA aligner (**Supplementary Figure 3**).

Confidence Score

For short reads, the confidence score of each eccDNA locus is calculated from the number of discordant and splits reads at the locus and the coverage at the locus boundaries. For long reads, the confidence score of each eccDNA locus is calculated from the number of repeat units in each mapping read and the total coverage at the locus boundaries. Users can adjust all parameters to customize confidence score calculation.

Plant Material and Growth Conditions

Seeds from *Arabidopsis thaliana* WT ecotype Columbia-0 were surface sterilized and sown on 1/2 MS medium [1% sucrose, 0.5% Phytagel (Sigma), pH 5.8], stratified for 2 days at 4°C, and grown in a controlled chamber (Percival, United States) at 21°C under long-day conditions (16-h light). Leaf material from 12 individuals was harvested after 2 weeks. For heat shock, the *in vitro* plates were exposed at 6°C for 12 h and 37°C for 24 h and material was extracted after a 24 h recovery at 21°C. Swiss winter wheat (*Triticum aestivum* cv. *Arina*) seeds originate from the Agroscope GenBank. Wheat seeds were presoaked in sterilized water overnight and sterilized by a 10 min 50°C heat shock. Seedlings were germinated and grown under controlled conditions in a Sanyo MLR-350 growth chamber under long-day conditions 16 h (light) at 20°C (day) and 18°C (night) for 4 days.

DNA Extraction

For Arabidopsis, seedlings were collected into one tube immediately snap-frozen in liquid nitrogen and stored at −80°C until DNA extraction, in duplicate. For wheat, three individual seedlings were collected into one tube immediately snap-frozen in liquid nitrogen and stored at −80°C until DNA extraction. For both species, the total DNA was extracted using the CTAB method. Total DNA quantity was measured with a Qubit Fluorometer (Thermo Fisher Scientific).

eccDNA Enrichment

Genomic DNA (2 µg) of each sample was treated with Plasmid-Safe™ ATP-Dependent DNase (Epicentre) according to the manufacturer's instructions overnight. Following digestion, DNA was precipitated with 0.1 volume of 3 M sodium acetate (pH 5.2), 2.5 volumes of ethanol, and 1 µl of GlycoBlue™ Coprecipitant (Ambion) overnight at −20°C. After centrifugation at 4°C for 1 h and washing with 70% ethanol, 100 ng of precipitated circular DNA was directly resuspended in the Illustra TempliPhi Sample Buffer and then amplified by random rolling circle amplification using the Illustra™ TempliPhi Amplification Kit (GE Healthcare) according to the manufacturer's instructions. The enriched amplification product was precipitated and debranched using the NEB T7 Endonuclease following the manufacturer's instructions. For both *Arabidopsis* and wheat samples, 1 ng of amplified DNA was used to prepare libraries for Miseq sequencing as in Lanciano et al., 2017. For the wheat samples, after final precipitation, 400 ng of DNA was used to prepare an ONT library using the Nanopore Rapid Barcoding Sequencing Kit (SQK-RBK004). DNA was sequenced on a MinION.

Data and Code Availability

All high-throughput sequencing data generated in this study have been deposited to the European Nucleotide Archive[4] under the

[4]https://www.ebi.ac.uk/ena

PRJEB46420 project. Source code and test samples for the ecc_finder pipeline are available at https://github.com/njaupan/ecc_finder.

RESULTS

Overview of eccDNA Detection Using ecc_finder

ecc_finder is designed to analyze eccDNA data generated from eccDNA-seq using Illumina paired-end shorts reads or ONT long reads (**Figure 2**). Two modes of analysis are proposed: either a mapping mode guided by a reference genome or a *de novo* assembly mode that is reference-free. Both modes can be used on the same data set (hybrid mode). For the Illumina short reads in the mapping mode, ecc_finder first uses BWA (Li and Durbin, 2009; default aligner) to map eccDNA data to a reference genome to detect loci enriched for eccDNA signals (**Figure 2A**). ecc_finder then detects eccDNA-producing loci based on discordant and split read pairs at junction and then filter by confidence score (Methods). For long reads in the mapping mode, ecc_finder detects circular long reads based on sub-read alignment from tandemly repeated reads (**Figure 2B**) and further filter by confidence score (Methods). The output of the mapping mode using either Illumina short reads or ONT long reads or a hybrid of both results in a list of candidate loci (**Figure 2C**).

Taking into account the high similarity of eccDNA-producing repeated loci, ecc_finder then calculates the read distribution of each candidate locus to filter out false positives (**Figure 2D**). In the end, ecc_finder produces bed files of the coordinates of each eccDNA-producing locus and the corresponding eccDNA sequence (**Figure 2E**). In addition, for comparative analysis, the bed output of all samples is further normalized to easily implement multiple samples into a final report (**Figure 2K**).

In the assembly mode, ecc_finder uses the k-mer assembler Spades (Prjibelski et al., 2020) to assemble short reads and the repeat unit recognition and consensus calling tool Tidehunter (Gao et al., 2019) to assemble long reads (**Figures 2G,H**). Instead of examining the performance of different assemblers, ecc_finder constructs a representative set by clustering the assembled contigs. ecc_finder then uses CD-hit (Li and Godzik, 2006) to self-align contigs to contigs and filter out redundant

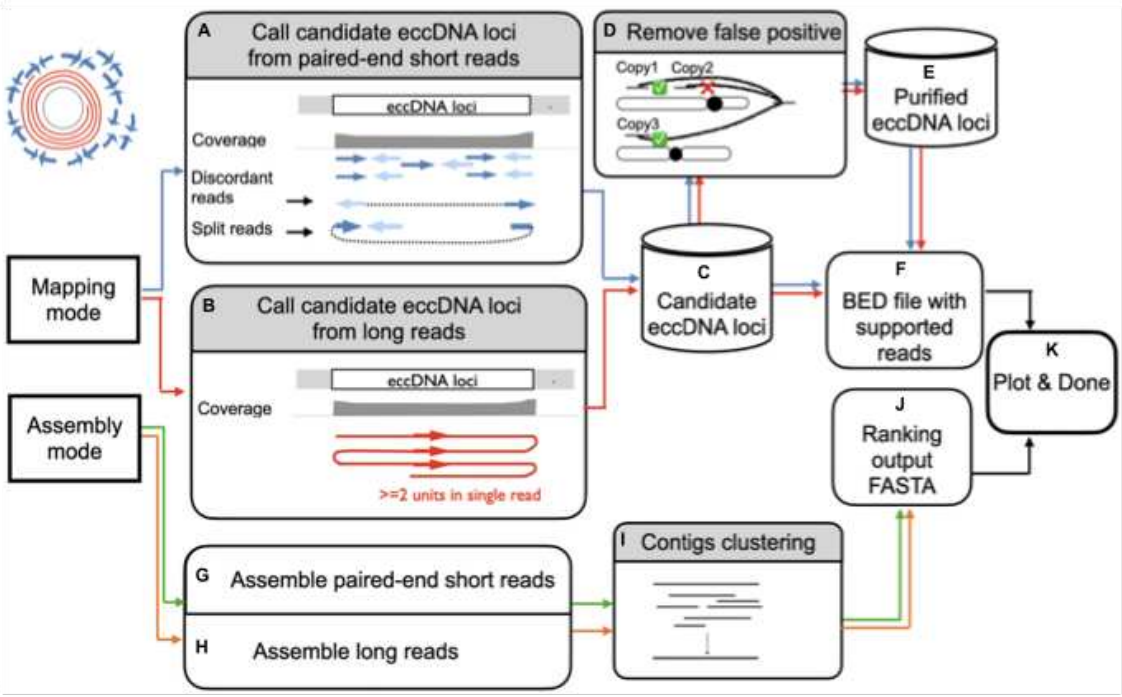


FIGURE 2 | Overview of eccDNA detection using ecc_finder. ecc_finder identifies eccDNA loci from illumina paired -end short reads (SR, *blue/green*) or Nanopore long reads (LR, red/orange) with or without a reference genome. In mapping mode, ecc_finder filters discordant and split reads detected from SR (**A**, *blue*), and filters for more than 2 junctions in a single read from LR (**B**, *red*), the reference genome being provided by the user. Using these filtered reads together with the covergae information, ecc_finder establishes a list of eccDNA candidates **(C)**. ecc_finder further detects false positive eccDNAs originating from repeated loci **(D,E)**. The bed output of a sample is further normalized **(F)** to append multiple samples to create a heat map **(K)**. In assembly mode, ecc_finder assembles SR with a k-mer assembler (**G**, *green*), and/or assembles LR using a repeat unit recognition algorithm (**H**, *orange*). The contigs are then clustered based on highly similarity **(I)**. The output is a fasta file indicating the number of supported reads for each contig **(J)**. Both mapping and assembling mode can be run in parallel to generate a heat map **(K)**.

contigs with 80% similarity (**Figure 2I**). Finally, the output of ecc_finder is a FASTA sequence file of all contigs ranked by the number of supporting reads (**Figure 2J**).

Benchmarking eccDNA Detection Tools Based on Short Reads

In order to evaluate the sensitivity, accuracy, and computational requirements of different eccDNA detection tools, we used public eccDNA data from *Homo sapiens* (NA12878, Møller et al., 2018), and we produced heat-stressed *Arabidopsis thaliana* and common wheat (*Triticum aestivum*) eccDNA-seq data. We selected these species for their diverse genome sizes and for the presence of previously described eccDNAs in the case of *Arabidopsis*. The initial step of eccDNA detection tools corresponds to genome indexing, necessary to speed up the mapping algorithms. Circle-Map, Circle_finder, and ecc_finder (default mode) use BWA to map eccDNA data on the corresponding reference genome, whereas ECCsplorer requires segemehl (Hoffmann et al., 2009). ECCsplorer spent 2.4 h indexing the human genome and 16.3 h for the wheat genome, which is twice the time compared to BWA. The following comparisons thus excluded indexing and mapping steps, in order to only account for the eccDNA detection step. Among all tools, ecc_finder greatly improved the computational time and performed faster on all data sets, followed by Circle_finder, and ECCsplorer (**Figure 3A**). The Circle-Realign step of Circle-map is the most time-consuming one, and it cannot process index files created by BWA for very large genomes (such as wheat). ECCsplorer is the only up-to-date automated pipeline that can detect eccDNA and establish consensus sequences for non-model species. Unfortunately, it failed to process our wheat eccDNA data because the tools it implemented ran out of memory and disk storage on a cluster with 96 CPUs and 496 GB RAM (segemehl produced 200G for indexing the wheat genome). Therefore, when considering computation performance, ecc_finder is one of two options to solve the problem of eccDNA analysis in large genomes.

We then evaluated the eccDNA detection accuracy of all tools. ecc_finder filters eccDNA-producing loci not only by the number of split and discordant reads, their alignment, and orientation but also by genomic enrichments to remove noisy signals and optimize redundancy. In heat-stressed *Arabidopsis* samples, ecc_finder detected 4 eccDNA generating loci corresponding to the active copies of the *ONSEN/ATCOPIA78* TE (**Figure 3B**; Sanchez et al., 2017). ecc_finder further identified known eccDNAs originating from repeats: a 9321 bp region located on chromosome 2 (Chr2:1029–10,350) and a 18433 bp region located on chromosome 3 (Chr3:14190444–14207658) corresponding to rDNA (Cloix et al., 2000; Abou-Ellail et al., 2011). As expected, ecc_finder also detected a region encompassing 245.8 kb on chromosome 2 (Chr2:3234927–3294252, Chr2:3297349–3,401,635, Chr2:3424305–3,453,213, and Chr2:3456196–3509451) and corresponding to mitochondrial DNA integrated into the nuclear genome (Saccone et al., 2000).

All detected eccDNAs had previously been validated, showing the accuracy of ecc_finder in identifying eccDNA-producing loci. The clear eccDNA sequence boundaries of ecc_finder output are also a specificity of this tool (**Figures 3C,D**). By comparison, ECCsplorer detected 7 *ONSEN* producing loci, 2 of 7 being incomplete and corresponding to false positives, Circle_finder detected 6 *ONSEN* producing loci, 2 of them being false positive, while Circle-map did not detect any *ONSEN* eccDNA (**Figure 3E; Supplementary Figures 4, 5; Supplementary Table 1**).

We then tested ecc_finder accuracy in detecting eccDNAs in human and wheat eccDNA-seq data. By default, ecc_finder removes circles smaller than 100 bp to reduce the noise coming from satellites. The eccDNA size distribution in the human data set indicated that ecc_finder detected a smaller set of eccDNA, but remained similar to the size found in the original study (**Figure 3F**). ECCsplorer and Circle_finder also detected a smaller set, while Circle-map gave similar circle size ranges (**Supplementary Figure 6**). For the wheat data set, given that over 80% of the wheat genome contains TEs (Wicker et al., 2018), eccDNA detection is challenging. ecc_finder detected 600 eccDNA-producing loci in the wheat eccDNA-seq (Illumina data) filtering by at least 10 split reads and 5 discordant reads, with 95.5% of these loci also being identified by Circle_finder.

Detection of eccDNA in Wheat Using Nanopore Long Reads

We then tested the performance of ecc_finder on wheat ONT eccDNA-seq data. ecc_finder detected 161 eccDNA-producing sites in two replicates (**Figure 4A**). These loci were distributed over all the 21 chromosomes, with eccDNA sizes ranging from 100 bp (detection threshold) to 40.1 kb (**Figure 4B**). We further characterized the output of ecc_finder and confirmed that the eccDNA-producing loci corresponded to rDNA, chloroplast DNA, and repetitive sequences. For example, the largest eccDNA-producing locus on chromosome 1B is 40.1 kb long and covers an 18S rDNA gene cluster (**Figure 4C**). The second and third largest loci (26.2 kb on chromosome 1D and 21.6 kb on chromosome 7B) are 99% identical to the chloroplast genome (**Figure 4D**). Overall, these findings are consistent with the observations we made in *Arabidopsis*.

Detection of eccDNA in Short-Read Genomic Data Without eccDNA Enrichment

EccDNAs have recently been identified from genomic and/or ATAC-seq data in mammal samples (Turner et al., 2017; Wu et al., 2019; Kumar et al., 2020), without prior enrichment for circular DNA. We have tested ecc_finder on genomic data using the tumor sample GBM39 sequenced by low coverage whole-genome sequencing (50 bp paired-end reads; Turner et al., 2017). ecc_finder successfully detected 95 discordant reads at the junction of the 1.29 Mb eccDNA (data not shown), which is also consistent with the results of Wu et al., (2019). However, ecc_finder was unable to

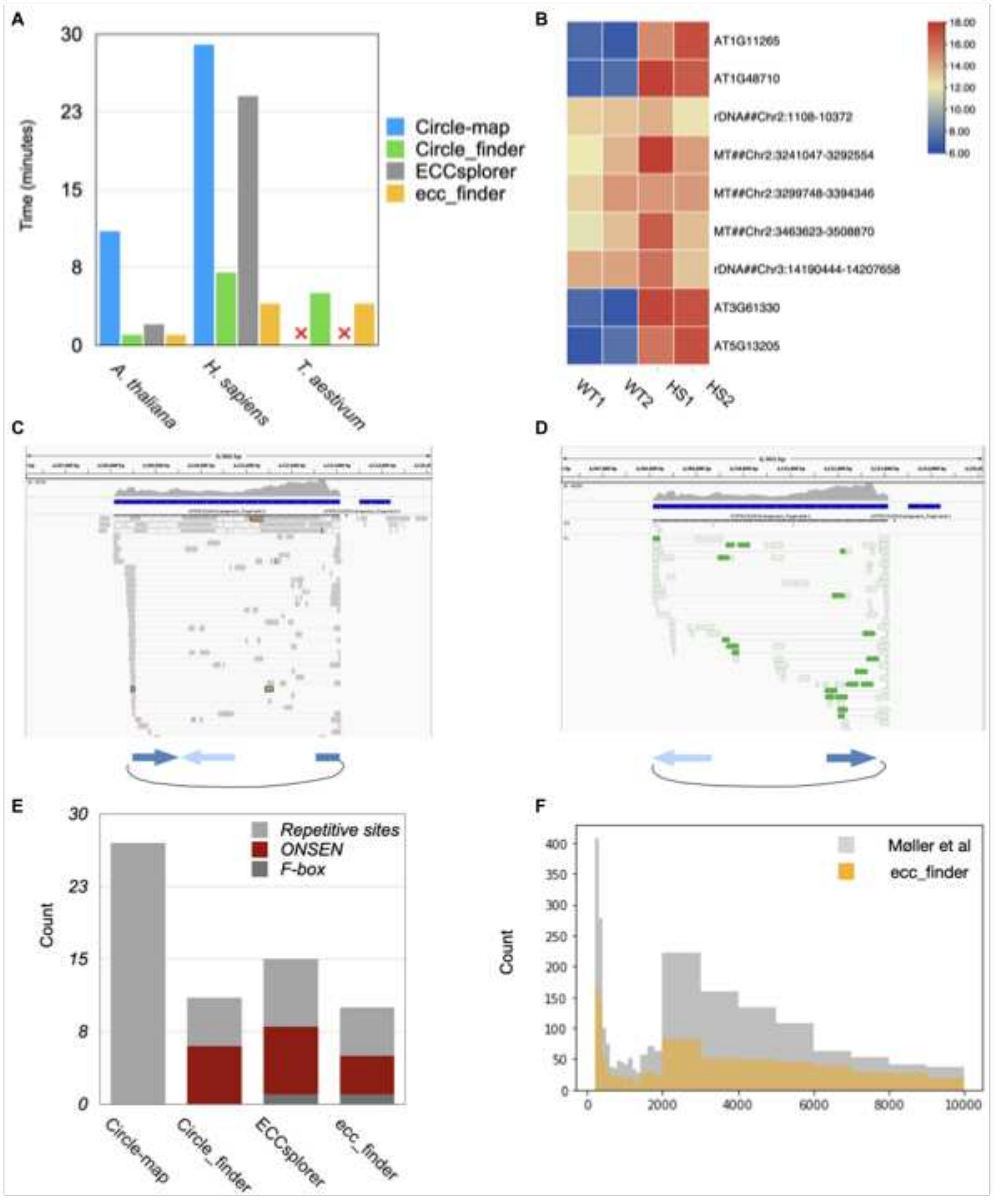


FIGURE 3 | Performance of different eccDNA detection tools using Illumina short reads. **(A)** Real consumed time on three different species using different tools. Circle-map and ECCsplorer failed to process the ecc-DNA-seq data of *T. aestivum* because (1) it ran out of RAM, (2) the csi file produced from samtools index for large genome cannot be used for the next step. **(B)** Heat map of 8 eccDNA sites detected using ecc_finder in the heat-stressed *Arabidopsis* (HS1, HS2) and wild-type Col-0 (WT1, WT2) with two replicates, respectively. Among them, 4 out of 9 correspond to *ONSEN/ATCOPIA78* loci. **(C,D)** Examples of the clear boundaries of split **(C)** and discordant reads **(D)** detected by ecc_finder, supporting the eccDNA form of *ONSEN/ATCOPIA78* in the *Arabidopsis* mobilome-seq data. **(E)** Number of classified genomic sites forming eccDNA detected from different tools in the heat-stressed *A. thaliana* eccDNA-seq data. **(F)** Size distribution of detected eccDNA using ecc_finder compared to original study using the same circular-DNA-enriched dataset (Møller et al., 2018).

construct the final structure of this large eccDNA because no split read could be detected. Therefore, in its mapping mode, ecc_finder will output the bed file containing the numbers of split and discordant reads for any genomic data. However, the peak calling will not be effective because of the lack of enrichment. We have not tested ecc_finder on ATAC-seq data but a recent method for this specific type of data has been described (Kumar et al., 2020).

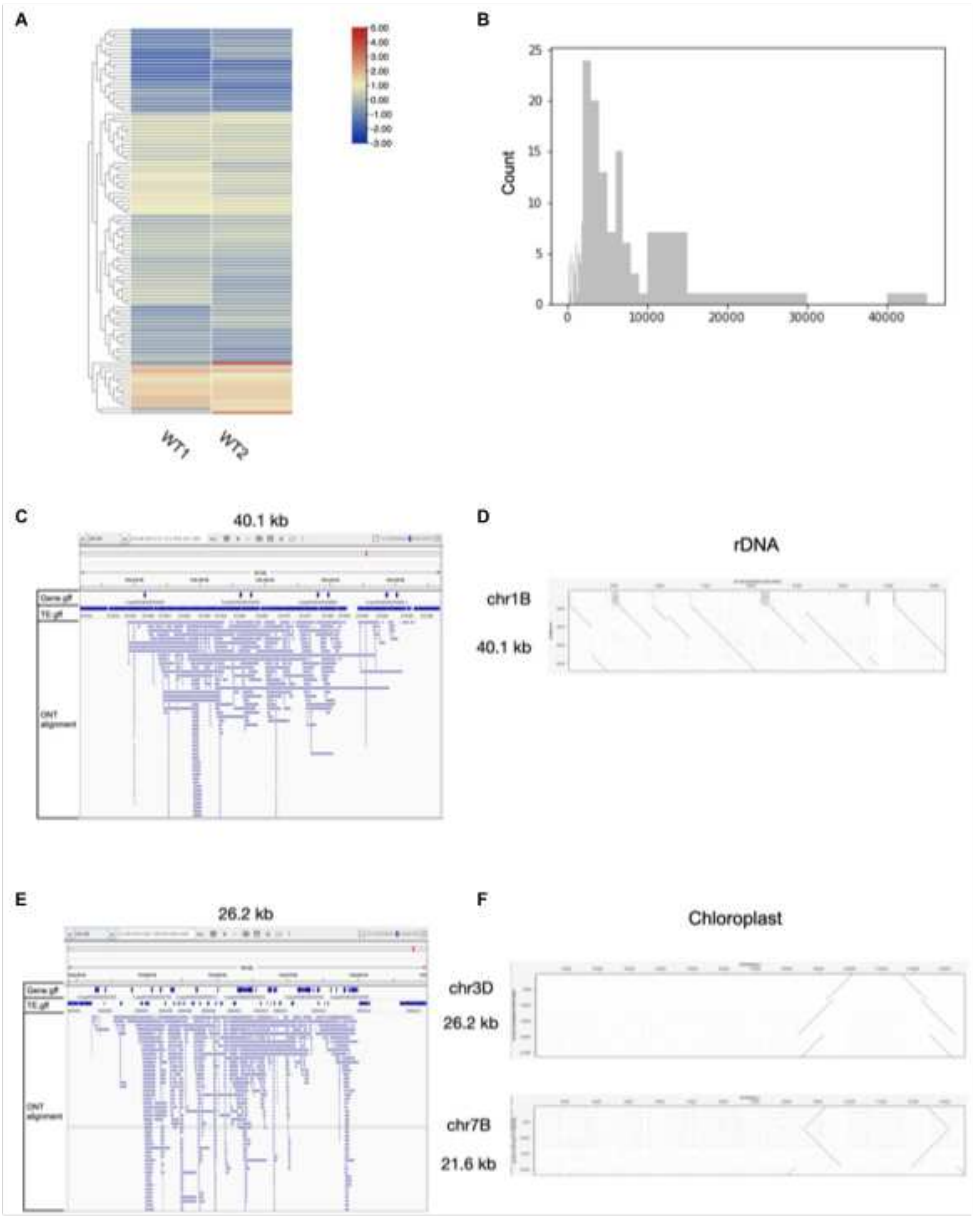


FIGURE 4 | Corresponding origins of eccDNAs loci detected by ecc_finder in wheat using mobilome-seq with Nanopore long reads. **(A)** Heat map of the 161 eccDNA sites detected using ecc_finder in the wheat wild-type (WT1, WT2) with two replicates, respectively. **(B)** Size distribution of the 161 detected eccDNAs. **(C)** Read distribution of the largest eccDNA-producing locus (40.1 kb) mapping to the reference genome. **(D)** Dot plot of the largest ecc-DNA-producing locus (26.2 kb) mapped to the reference genome. **(E)** Read distribution of the second largest eccDNA-producing locus (26.2 kb) mapped to the reference genome. **(F)** Dot plot of two eccDNA-producing loci (26.2 kb and 21.6 kb) and chloroplast genome, respectively.

DISCUSSION

EccDNA-producing loci can be repeated in the genome. However, current tools do not take into account the repeated nature of these loci, and the detected loci can thus be redundant, notably for TEs, rDNA, and satellites. In this case, identifying the exact locus producing eccDNA can be challenging. For a given family of long terminal repeats (LTR) retrotransposons producing eccDNA

for instance, all copies belonging to the same family and sharing the same LTR sequences will produce alignments of split and discordant reads at their boundaries. Only the copies producing *bona fide* eccDNA will thus display an even distribution of split and discordant reads throughout their internal region. ecc_finder implemented this step in its detection of eccDNA-producing loci in order to improve the detection of eccDNAs. Additionally, ecc_finder enables the use of eccDNA long-read sequencing data that is likely to become the standard in the coming years.

CONCLUSION

Although eccDNA was known for decades in yeasts, plants, and animals, growing evidence in recent years suggests that this peculiar form of DNA plays a role in rapid adaptation, for instance in cancer cells (Kumar et al., 2017; Verhaak et al., 2019; Kim et al., 2020; Wang et al., 2021) or herbicide resistant plants (Koo et al., 2018), by promoting overexpression and alternate epigenetic state of a selected set of genes. Characterizing the full repertoire of eccDNA is becoming crucial, and several protocols are available to enrich a DNA sample for eccDNA and sequence it with short or long reads. We believe that ecc_finder that was developed here will facilitate the downstream bioinformatic analysis of these data sets, notably for ONT long reads, and accelerate the discoveries linked to eccDNA biology in many species, including the ones with the largest genomes and high transposable element content.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: European Nucleotide Archive (ENA) repository under project number PRJEB46420 (https://www.ebi.ac.uk/ena/browser/view/PRJEB46420).

AUTHOR CONTRIBUTIONS

PZ produced mobilome-seq data, wrote the bioinformatic scripts, analyzed data, and wrote the manuscript. HP produced and analyzed mobilome-seq data and wrote the manuscript. CL produced mobilome-seq data. EB analyzed data and wrote the manuscript. MM designed the experiment, analyzed data, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

MM is supported by a grant from the French National Agency for Research (ANR-13-JSV6-0002 "*ExtraChrom*"). This study is set within the framework of the "Laboratoire d'Excellence (LABEX)" TULIP (ANR-10-LABX-41) and of the "Ecole Universitaire de Recherche (EUR)" TULIP-GS (ANR-18-EURE-0019). PZ, MM, and EB are members of the European Training Network "*EpiDiverse*" that received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965. HP is supported by China Scholarship Council (CSC) Grant (201806990012). EB and HP received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 725701, BUNGEE). MM and EB are supported by a grant from the Agence Nationale de la Recherche and Fonds National Suisse (ANR-21-PRCI-CE02, FNS 310030E_205554 "CropCircle").

ACKNOWLEDGMENTS

We would like to thank three reviewers for constructive comments and our EpiDiverse colleagues for discussions on mobilome-seq analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.743742/full#supplementary-material

REFERENCES

Abou-Ellail, M., Cooke, R., and Sáez-Vásquez, J. (2011). Variations in a team: major and minor variants of *Arabidopsis thaliana* rDNA genes. *Nucleus* 2, 294–299. doi: 10.4161/nucl.2.4.16561

Cloix, C., Tutois, S., Mathieu, O., Cuvillier, C., Espagnol, M. C., Picard, G., et al. (2000). Analysis of 5S rDNA arrays in *Arabidopsis thaliana*: physical mapping and chromosome-specific polymorphisms. *Genome Res.* 10, 679–690. doi: 10.1101/gr.10.5.679

Cohen, Z., Bacharach, E., and Lavi, S. (2006). Mouse major satellite DNA is prone to eccDNA formation via DNA ligase IV-dependent pathway. *Oncogene* 25, 4515–4524. doi: 10.1038/sj.onc.1209485

Cohen, S., and Méchali, M. (2002). Formation of extrachromosomal circles from telomeric DNA in *Xenopus laevis*. *EMBO Rep.* 3, 1168–1174. doi: 10.1093/embo-reports/kvf240

Cohen, S., Yacobi, K., and Segal, D. (2003). Extrachromosomal circular DNA of tandemly repeated genomic sequences in drosophila. *Genome Res.* 13, 1133–1145. doi: 10.1101/gr.907603

Deshpande, V., Luebeck, J., Nguyen, N.-P. D., Bakhtiari, M., Turner, K. M., Schwab, R., et al. (2019). Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* 10:392. doi: 10.1038/s41467-018-08200-y

Gao, Y., Liu, B., Wang, Y., and Xing, Y. (2019). TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. *Bioinformatics* 35, i200–i207. doi: 10.1093/bioinformatics/btz376

Hirochika, H., and Otsuki, H. (1995). Extrachromosomal circular forms of the tobacco retrotransposon *Ttol*. *Gene* 165, 229–232. doi: 10.1016/0378-1119(95)00581-P

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., et al. (2009). Fast mapping of short sequences with mismatches, insertions

and deletions using index structures. *PLoS Comput. Biol.* 5:e1000502. doi: 10.1371/journal.pcbi.1000502

Hotta, Y., and Bassel, A. (1965). Molecular size and circularity of DNA in cells of mammals and higher plants. *PNAS USA* 53, 356–362. doi: 10.1073/pnas.53.2.356

Kim, H., Nguyen, N.-P., Turner, K., Wu, S., Gujar, A. D., Luebeck, J., et al. (2020). Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.* 52, 891–897. doi: 10.1038/s41588-020-0678-2

Koche, R. P., Rodriguez-Fos, E., Helmsauer, K., Burkert, M., MacArthur, I. C., Maag, J., et al. (2020). Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nat. Genet.* 52, 29–34. doi: 10.1038/s41588-019-0547-z

Koo, D.-H., Molin, W. T., Saski, C. A., Jiang, J., Putta, K., Jugulam, M., et al. (2018). Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed *Amaranthus palmeri*. *Proc. Natl. Acad. Sci. U. S. A.* 115, 3332–3337. doi: 10.1073/pnas.1719354115

Kumar, P., Dillon, L. W., Shibata, Y., Jazaeri, A., Jones, D. R., and Dutta, A. (2017). Normal and cancerous tissues release extrachromosomal circular DNA (eccDNA) into the circulation. *Mol. Cancer Res. MCR* 15, 1197–1205. doi: 10.1158/1541-7786.MCR-17-0095

Kumar, P., Kiran, S., Saha, S., Su, Z., Paulsen, T., Chatrath, A., et al. (2020). ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines. *Sci. Adv.* 6:aba2489. doi: 10.1126/sciadv.aba2489

Lanciano, S., Carpentier, M.-C., Llauro, C., Jobet, E., Robakowska-Hyzorek, D., Lasserre, E., et al. (2017). Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLoS Genet.* 13:e1006630. doi: 10.1371/journal.pgen.1006630

Lanciano, S., Zhang, P., Llauro, C., and Mirouze, M. (2021). Identification of extrachromosomal circular forms of active transposable elements using Mobilome-Seq. *Methods Mol. Biol.* 2250, 87–93. doi: 10.1007/978-1-0716-1134-0_7

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinforma. Oxf. Engl.* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Mazzucco, G., Huda, A., Galli, M., Piccini, D., Giannattasio, M., Pessina, F., et al. (2020). Telomere damage induces internal loops that generate telomeric circles. *Nat. Commun.* 11:5297. doi: 10.1038/s41467-020-19139-4

Mehta, D., Cornet, L., Hirsch-Hoffmann, M., Zaidi, S. S. e. A., and Vanderschuren, H. (2020). Full-length sequencing of circular DNA viruses and extrachromosomal circular DNA using CIDER-Seq. *Nat. Protoc.* 15, 1673–1689. doi: 10.1038/s41596-020-0301-0

Møller, H. D., Mohiyuddin, M., Prada-Luengo, I., Sailani, M. R., Halling, J. F., Plomgaard, P., et al. (2018). Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nat. Commun.* 9:1069. doi: 10.1038/s41467-018-03369-8

Møller, H. D., Parsons, L., Jørgensen, T. S., Botstein, D., and Regenberg, B. (2015). Extrachromosomal circular DNA is common in yeast. *Proc. Natl Acad. Sci. USA* 112, E3114–E3122. doi: 10.1073/pnas.1508825112

Navrátilová, A., Koblížková, A., and Macas, J. (2008). Survey of extrachromosomal circular DNA derived from plant satellite repeats. *BMC Plant Biol.* 8:90. doi: 10.1186/1471-2229-8-90

Prada-Luengo, I., Krogh, A., Maretty, L., and Regenberg, B. (2019). Sensitive detection of circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads. *Bioinformatics* 20:663. doi: 10.1186/s12859-019-3160-3

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinform.* 70:e102. doi: 10.1002/cpbi.102

Saccone, C., Gissi, C., Lanave, C., Larizza, A., Pesole, G., and Reyes, A. (2000). Evolution of the mitochondrial genetic system: an overview. *Gene* 261, 153–159. doi: 10.1016/S0378-1119(00)00484-4

Sanchez, D. H., Gaubert, H., Hajk-Georg Drost, H. G., Radu Zabet, N., and Paszkowski, J. (2017). High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. *Nature Comm.* 8:1283. doi: 10.1038/s41467-017-01374-x

Sinclair, D. A., and Guarente, L. (1997). Extrachromosomal rDNA circles--a cause of aging in yeast. *Cell* 91, 1033–1042. doi: 10.1016/S0092-8674(00)80493-6

Turner, K. M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., et al. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* 543, 122–125. doi: 10.1038/nature21356

Verhaak, R. G. W., Bafna, V., and Mischel, P. S. (2019). Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat. Rev. Cancer* 19, 283–288. doi: 10.1038/s41568-019-0128-6

Wang, T., Zhang, H., Zhou, Y., and Shi, J. (2021). Extrachromosomal circular DNA: a new potential role in cancer progression. *J. Transl. Med.* 19:257. doi: 10.1186/s12967-021-02927-x

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-González, R. H., et al. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* 19:103. doi: 10.1186/s13059-018-1479-0

Wu, S., Turner, K. M., Nguyen, N., Raviram, R., Erb, M., Santini, J., et al. (2019). Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* 575, 699–703. doi: 10.1038/s41586-019-1763-5

Yan, Y., Guo, G., Huang, J., Gao, M., Zhu, Q., Zeng, S., et al. (2020). Current understanding of extrachromosomal circular DNA in cancer pathogenesis and therapeutic resistance. *J. Hematol. Oncol.J Hematol Oncol* 13:124. doi: 10.1186/s13045-020-00960-9

Zellinger, B., Akimcheva, S., Puizina, J., Schirato, M., and Riha, K. (2007). Ku suppresses formation of Telomeric circles and alternative telomere lengthening in *Arabidopsis*. *Mol. Cell* 27, 163–169. doi: 10.1016/j.molcel.2007.05.025

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Zhang, Peng, Llauro, Bucher and Mirouze. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

2.1.3 Update on ecc_finder

Ecc_finder has gained some attention after its release. Anonymous researchers have asked 10 questions on github about other types of input, output interpretation, computational memory requirements, etc. I have answered them all accordingly. I am also considering a second version of ecc_finder, ecc_finder2, for processing low-coverage whole-genome sequencing data and ATAC-seq data to explore its broader applications. Considering that ecc_finder2 needs reliable validation, it is best to train on datasets with known eccDNA loci, such as publicly available data from *Arabidopsis thaliana*.

During the writing of this thesis manuscript, I wondered whether eccDNAs could be either sequence-specific, locus-specific or tissue-specific. The recent study of eccDNA in mammal embryonic cells (using long read eccDNA-seq) gives a negative answer to sequence and locus specificity and mentions that eccDNAs may be generated by random breaks in genomic DNA, notably during apoptosis (Wang et al., 2021c). In this case the authors propose that eccDNAs originate from the entire genome, through random single fragment cyclizations and multifragment cyclizations of the genome producing between 200 bp and 3 kb circles (Wang et al., 2021c). Concerning the tissue specificity, in rice, eccDNAs generated by active TEs have been detected in endosperm tissue, but not in embryo or seed coat (Lanciano et al., 2017a). Similarly, tissue specific eccDNAs have been found in Arabidopsis flower, leaf, stem and root tissues (Wang et al., 2021b).

However, the latter study of eccDNA landscape in different tissues of *Arabidopsis thaliana* (Wang et al., 2021b) did not take into account that organelle DNA fragments transferred to the nucleus (such as NUPTs for nucleoplasmic DNA and NUMTs for nuclear mitochondrial DNA; Saccone et al., 2000; Yoshida et al., 2014) should be excluded from eccDNAs. Indeed these eccDNAs more likely originate from the organelles themselves, and not the NUPTs and NUMTs. For instance, there are 334 (45%) eccDNAs located in organelle genomes (mitochondria and chloroplast), 152 (20%) eccDNAs located in NUMTs and 79 (11%) eccDNAs located in NUPTs in the Wang et al. dataset. These eccDNAs derived from organelle sequences accounted for 76% of the total eccDNAs detected. In addition, 9 eccDNAs corresponded to ribosomal DNA repeats (rDNA) while 30 eccDNAs corresponded to centromeres (Figure II.1A). All

these eccDNAs from organelle genomes, NUPTs and NUMTs, rDNA and centromeric repeats should be discarded in downstream comparisons. Therefore, the 604/743 (81%) eccDNAs detected in this study do not correspond to novel eccDNAs.

In order to really evaluate the differences between eccDNAs detected from leaf, flower, stem and root tissues of Arabidopsis in terms of gene content, I reanalyzed this dataset, removing eccDNA originating from repeats and organelles listed above. This showed a dramatic reduction in the total number of root-specific eccDNAs (Figure II.1B). I could detect specific eccDNAs corresponding to flowers/leaves/stems/roots, as follows 94/104/96/38. Of note, there was no common eccDNA between roots and other different tissues (leaves, flowers and stems). To further validate this data, I analyzed eccDNAs from three replicates in the different tissues (Figure II.1C), and this showed a lack of consistency between samples. Additionally, the validation of *bona fide* eccDNAs is still questionable in this study, therefore I would recommend to use ecc_finder to re-do the step of detection.

In conclusion for this part, because the field of eccDNA detection is becoming a hot topic, I would suggest to apply stringent computational methods for their rigorous detection.

A

Category	eccDNA number	percentage
Mitochondria and Chloroplast	334	45%
NUMTs (nuclear mitochondrial DNA)	152	20%
NUPTs (nuclear plastid DNA)	79	11%
Ribosomal DNA	9	1%
Centromeric repeats	30	4%
TE	57	8%
Novel eccDNAs	82	11%
Total number of eccDNAs	743	

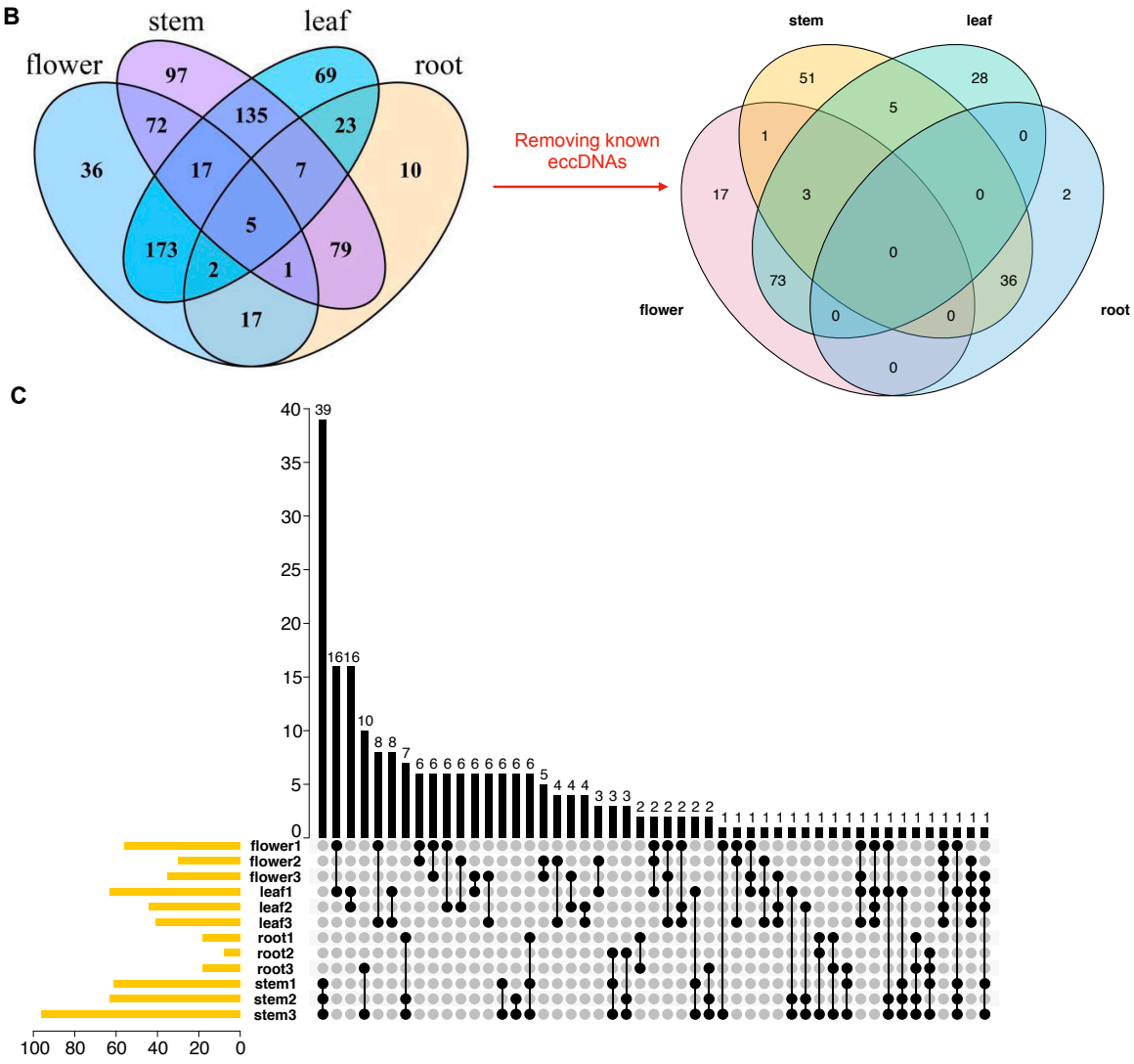


Figure II.1. Reanalysis of the characteristics of eccDNAs from leaf, flower, stem and root tissues of *Arabidopsis thaliana* in Wang et al. (2021). (**A**) Reclassification of 743 eccDNAs in Wang et al. (2021). All eccDNA loci in Supplemental Table 1 were intersected with annotations including NUPTs and NUMTs, rDNA, and centromeric repeats using bedtools and then re-annotated. (**B**) Comparison of differences between eccDNA detected from leaf, flower, stem, and root tissues of Arabidopsis after removal of positive controls. (**C**) Intersection of eccDNAs between eccDNA detected from leaf, flower, stem, and root tissues of Arabidopsis in three replicates.

2.2 SASAR: a new tool for meta-assembling plant genomes with long reads

2.2.1 My contribution to SASAR

I have developed 2 versions of SASAR, one for release in 2019 and one for the current version.

When I joined the lab in 2018, genome assembly with long-read sequencing had just started. At that time, about 14 plant genomes ranging in size from 120 Mb to 2.53 Gb had been sequenced with ONT, 5 of which were highly contiguous with N50 over 5 Mb. However, while ONT rapidly developed new chemistry and basecall workflows to improve read accuracy, some issues with downstream bioinformatics analysis were left to the attempted genomic projects for various laboratories.

Therefore, we sought to establish a standard benchmark for new assembly tools relative to genome size prior to scaffolding using ONT sequencing. We benchmarked five state-of-the-art assembly tools and developed a meta-assembly tool using a super assembly tool from assembly reconciliation, namely SASAR. Genome contiguity, genetic integrity and other quality measures indicated that the *A. thaliana* Columbia assembly was comparable to or better than the gold standard reference genome TAIR10.1, filling 73% of the centromeric gaps.

However, the SASAR manuscript was rejected in 2019 and is now surpassed by other genome assemblies of *A. thaliana*. To follow up on future breakthroughs in assembly I updated SASAR's algorithm to implement the assembly graph, especially to annotate SV in the assembly. Although its release has been delayed, the updated version of SASAR will provide a pan-SV together with a genome assembly, which will be novel in the field. In the next section I summarize and give an update on the main results from the 2019 manuscript.

2.2.2 SASAR manuscript (bioRxiv, 2022)

2.2.2 SASAR: a new tool for meta-assembling plant genomes with long reads, via assembly graph

Abstract

Long-read sequencing technologies such as PacBio and Oxford Nanopore Technology (ONT), which currently dominate genome research, have undoubtedly become platforms for genome assembly. Genome assembly tools optimized for long-read data came into being. However, different assemblies produced by different assemblers or the same assembler with different parameters have different performances, with the trade-off between improve contiguity, repetition resolution, and computational consumption. The final assembly is usually selected based on further evaluation, so not all produced assemblies, including hidden genetic variations, are used. Here, we developed SASAR (https://github.com/njaupan/SASAR) as a meta-assembly tool to reconcile the result of different long read assemblies. This strategy allows the reconciliation of multiple assemblies to increase the contiguity and coordination of structural variants via assembly graph. Using long-read assemblies produced from plant and animal species, SASAR achieved a contiguous genome assembly in an efficient time without a reference guide. It is worth noting that SASAR provides the latest updated Arabidopsis reference genome, which fills 73% of its N-stretches and corrects mis-assembly, most of which are derived from transposable elements, including the widely studied *ONSEN/ATCOPIA78* retrotransposon activated by heat stress.

Keywords: long read, genome assembly, assembly reconciliation, assembly graph.

Results

"Super ASsembly" from Assembly Reconciliation (SASAR)

SASAR is an efficient and accurate meta-assembly tool for merging assemblies. It is implemented as an open source Python3 command line utility, and its source code and documentation can be obtained from https://github.com/njaupan/SASAR on GitHub. The main goal of SASAR is to use the specificity of different assemblies from one species to optimize *de novo* assembly (Figure II.2). It starts by identifying "pan-contigs".

Whole genome alignment of every two assemblies is conducted using Minimap2 to capture accurate base alignment (Figure II.2: step 1). Then SASAR uses CIGAR string values to parse the PAF file to obtain statistics on alignment length, indel size and gap compression identity (which is defined by Minimap2 developer http://lh3.github.io/2018/11/25/on-the-definition-of-sequence-identity). The non-mapped and contigs that only have one hit are retained, and the contigs mapped multiple times in the entire area are replaced by the corresponding longest contigs in other assemblies. This will create the minimum “pan-contigs” to all assemblies. In the second phase, SASAR will extend "pan-contigs". SASAR will go through two ends of the “pan-contigs” and information about contig extensions are recorded. The certain regions are further added to the head or tail of the aligned contigs (Figure II.2: step 2). In the third phase, SASAR will correct mis-assemblies. Most assembly errors are due to repetitive sequences, especially transposable elements (TEs). Structural variants (SVs) detected from the “pan-contigs” via building assembly graph will be stored and annotated into different mechanisms. SVs present at least in any two assemblies will be used to correct “pan-contigs”. Regions with TE domain will be expanded to find the boundaries of TE, and the insertion or deletion of internal paralogs will be polished based on the consensus of TE within the “pan-contigs". Several rounds of polishing will be performed and the SV at the point of discordant alignments will be recorded. Note that users have the option to break the contigs at points of potential mis-assembly (Figure II.2: step 3). As the last step, SASAR will rescue some small contigs. According to the user's choice, contigs that appear once in a particular assembly can be rescued (Figure II.2: step 4).

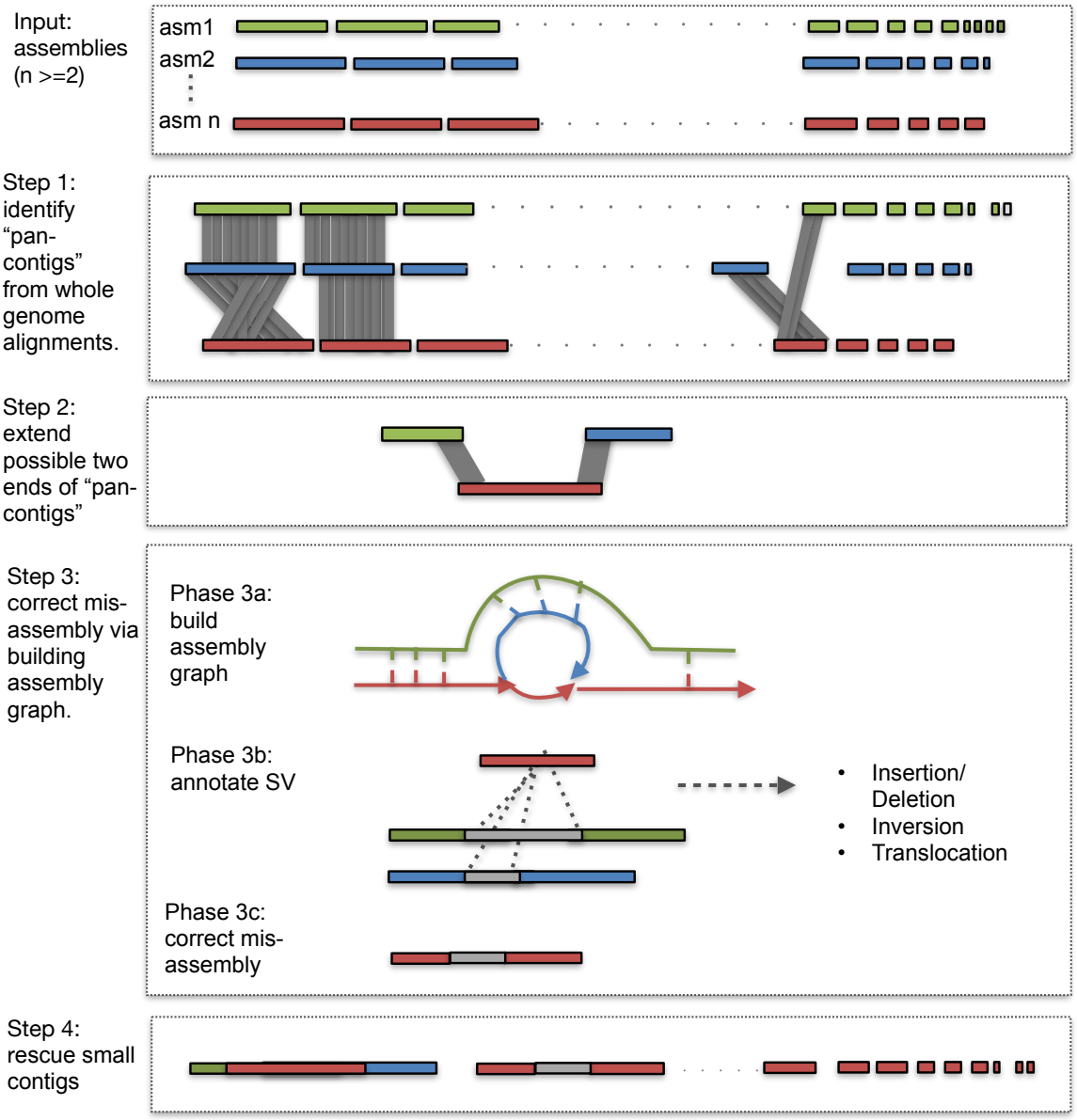


Figure II.2. Schematic pipeline of SASAR (Super ASsembly from Assembly Reconciliation). Different colors represent contigs assembled by different long read assemblers. Step 1: Identifying « pan-contigs », using whole genome alignment with minimap2. Step 2: Extending "pan-contigs". Step 3: Correct mis-assembly. Step 4: Rescuing some small contigs. According to the user's choice, contigs that appear once in a particular assembly can be rescued.

SASAR performance on merging assemblies from different assemblers

We evaluated SASAR performance on three datasets and compared it with assemblies generated by five state-of-the-art assemblers, namely Canu, Flye, Shasta, Canu-SMARTdenovo and wtdbg2 (Table 1). Each assembly was aligned to the corresponding reference genomes using minimap2, and basic metrics were extracted using QUAST-LG. In all assemblies of Arabidopsis datasets, SASAR assembly greatly balanced the longest NG50 of the SMARTdenovo assembly, the highest amount of BUSCO gene of the Flye assembly, and the highest resolution of centromeric repeats of the Canu assembly. It maintained an optimal continuity through 36 contigs and centromeric regions, with a total length approximately 10 Mb longer than TAIR10.1. In all assemblies of rice datasets, SASAR assembly kept the longest NG50 of the SMARTdenovo assembly and the highest score of BUSCO genes of the Canu assembly. In all assemblies of tomato datasets, SASAR assembly improved the reference assembly by 100Mbp (Table 1).

Notably, most of the extra 10 Mb genomic regions in the SASAR assembly of *A. thaliana* are centromeric regions. Compared to the estimated centromere size obtained from BAC sequencing, almost all centromeric sequences of chromosomes 2, 3, 4, 5 of *A. thaliana* have been assembled (Figure II.3). The SVs detected between our *A. thaliana* SASAR-assembly and the gold-standard reference genome (TAIR10) enabled us to close 80% of all N-stretches within TAIR10 (not shown). Most of SVs that filled N-stretch were aligned to transposable elements and microsatellites, consistent with the previous finding that the *A. thaliana* centromeric regions are rich in repetitive DNAs. In the non-centromeric regions of the *A. thaliana* genome, one 1 kb insertion and one solo *Copia* LTR insertion were detected in chromosome 5, located at Chr5:2610117 and Chr5:7876830, respectively. The 1 kb insertion was further amplified by PCR (Figure II.4) and was identified in the Col-XJTU assembly (Wang et al., 2021a), showing that this insertion is not restricted to our plant material. Most importantly, SASAR assembly corrected large mis-assemblies mediated by TEs, which were not shown in Col-XJTU assembly (Wang et al., 2021a) and Col-CEN assembly (Naish et al., 2021). For instance, the mis-assembly of *ONSEN/ATCOPIA78* in *A. thaliana* Col-0 (Figure II.5) was confirmed in different sequencing technologies and datasets generated by different studies and also validated by PCR (Gilly et al., 2014), indicating the sensitivity of SASAR assembly in the TE region.

Table 1. The quality and performance of long-read assemblies. The best value for each metric is highlighted in bold.

Genome	Metric	Reference	Canu	Canu+ SMARTdenovo	Flye	Shasta	wtdbg2	SASAR
A. thaliana Col-0 (132X)		TAIR10						
	Contig	7	530	**41**	116	642	292	**36**
	LG50	-	6	**4**	7	14	6	**4**
	NG50 (Mb)	-	7.50	**14.29**	6.19	3.03	9.11	**14.29**
	Largest contig (Mb)	-	13.53	16.06	11.40	12.61	13.33	**16.07**
	Assembly size (Mb)	119.67	131.31	118.97	118.37	118.69	119.79	**131.39**
	Alignment breakpoints	-	1140	**130**	123	131	213	**130**
	BUSCO (%)	-	98.1	99.2	98.8	99.2	99.2	**99.2**
	Centromeric repeats		**35549**	1487	1661	1611	2010	25554
O. sativa Nipponbare (34X)		IRGSP1.0						
	Contig	12+43	2318	760	908	3637	3152	**749**
	LG50	-	224	112	120	492	234	**112**
	NG50 (Mb)	-	0.46	1.01	0.91	0.22	0.39	**1.02**
	Largest contig (Mb)	-	2.50	3.92	**4.34**	1.58	3.57	**4.34**
	Assembly size (Mb)	373.80	372.60	372.17	365.65	342.64	**376.34**	372.60
	Alignment breakpoints	-	513	**423**	300	259	485	440
	BUSCO (%)	-	97.7	97.3	97.3	93.0	96.1	**97.7**
	Centromeric repeats		**4232**	1455	1001	1104	1260	3232
S. pennelli LA5240 (110X)		Ref.						
	Contig	**899**	2010	**899**	3180	9712	4986	**899**
	L50	**106**	169	**106**	165	1253	210	**106**
	NG50 (Mb)	**2.52**	1.55	**2.52**	1.68	0.16	1.23	**2.52**
	Largest contig (Mb)	**12.72**	10.01	**12.72**	17.77	1.76	15.01	**12.72**
	Assembly size (Mb)	915.60	961.83	915.60	1020.62	748.49	941.24	**1020.62**
	BUSCO (%)	**99.2**	99.2	**99.2**	99.6	99.6	98.9	**99.2**

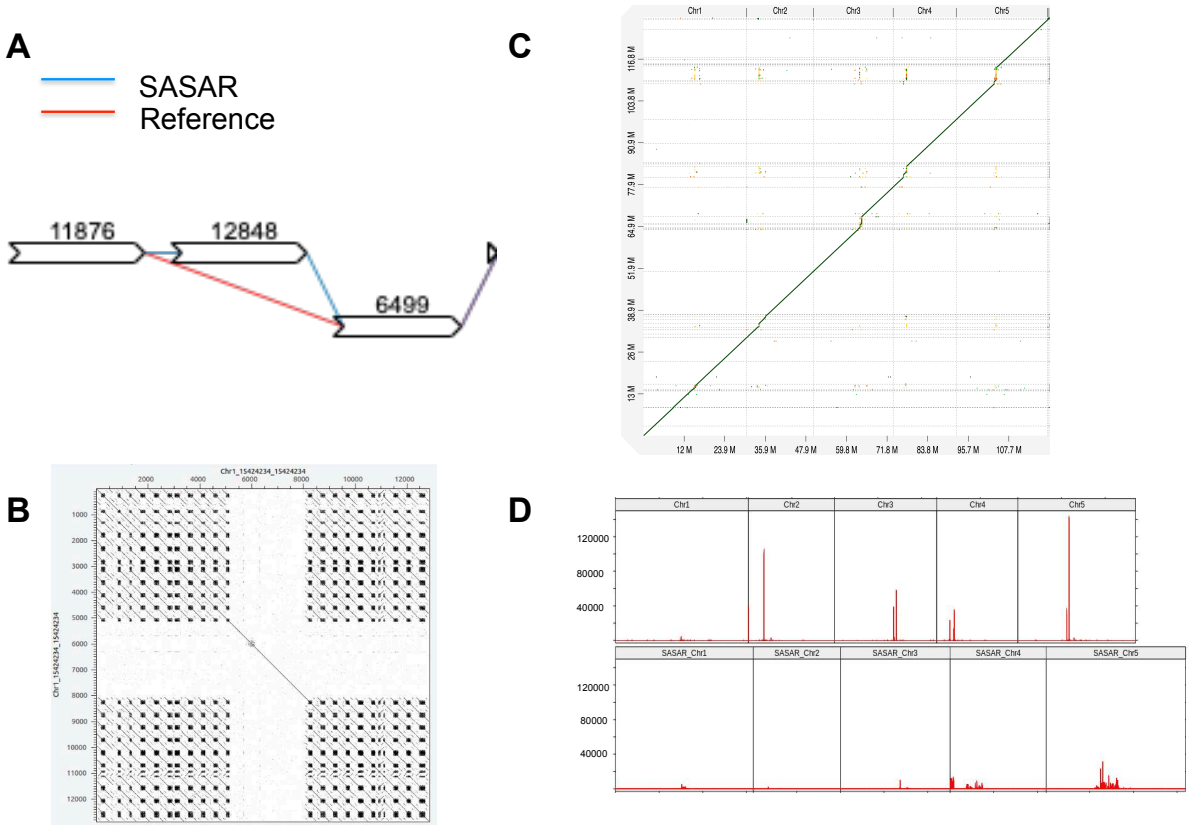


Figure II.3. SASAR performance in assembling the centromeres in *A. thaliana Col-0*. (A) A partial assembly graph consisting of chains of bubbles with the reference as the backbone. The bubble represents a 12848bp SV detected in the SASAR-assembly. (B) Dot plot of the 12848bp sequence against itself shows that the SV is a copy number variant of the 178bp centromere repeat. (C) Dot plot of SASAR-assembly versus reference genome TAIR10. SASAR assembled more centromeric regions of *A. thaliana* represented by “vertical lines”; (D) Raw reads coverage. Peaks of read coverage across the TAIR10 (upper panel) and the SASAR assembly (lower panel). Lower peaks indicate that more repeats are assembled. Note that SASAR assembly is 10 Mbp longer.

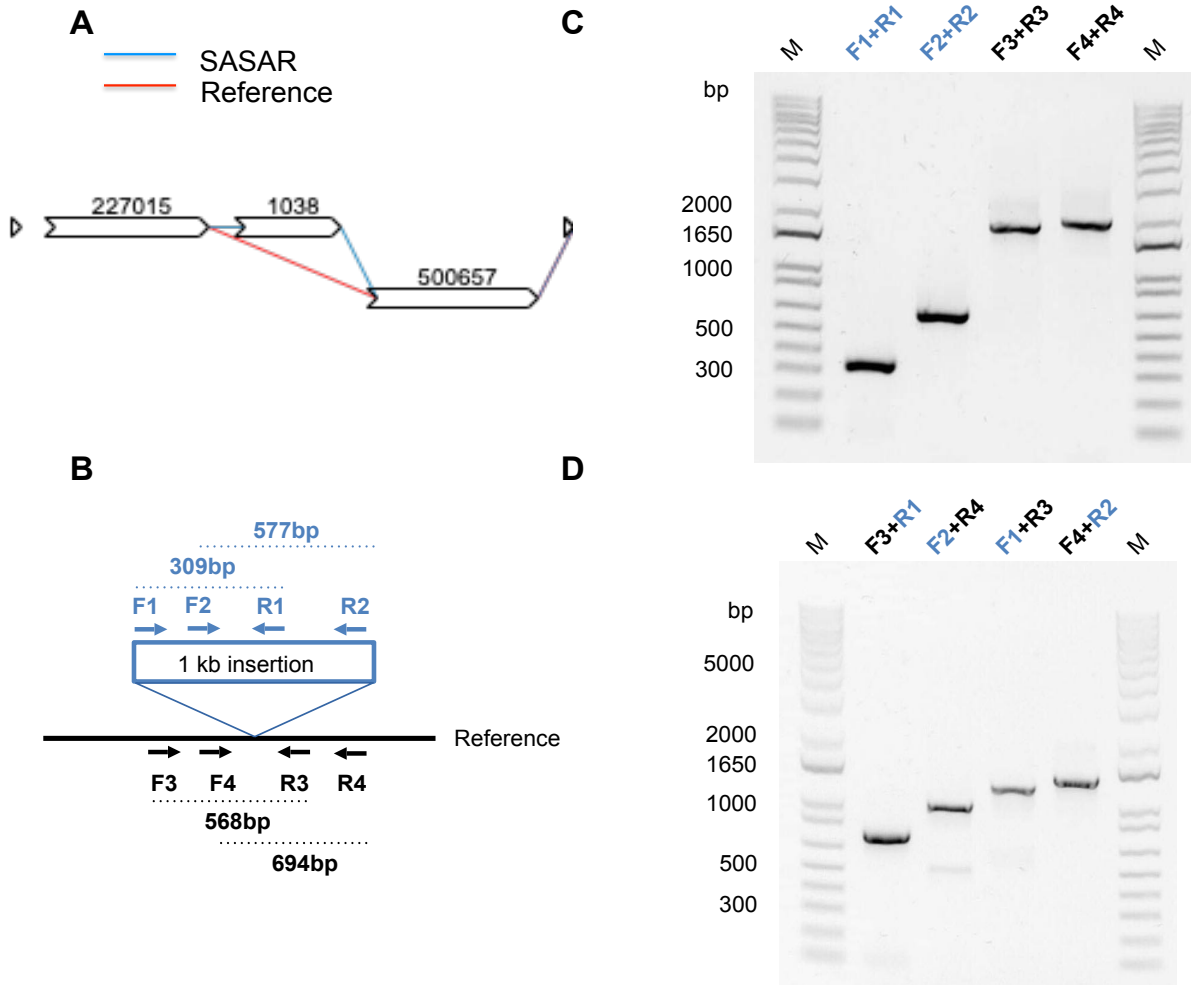


Figure II.4. SASAR performance in a 1kb insertion in the non-centromeric region of *A. thaliana* Col-0. (A) Partial assembly graph consisting of chains of bubbles with the reference as the backbone. The bubble represents a 1038bp insertion detected in the SASAR-assembly. (**B)** Scheme showing the position of primers used for PCR validation**. (C, D)** Gel electrophoresis of PCR products using Col-0 genomic DNA and primers combination as indicated. M: molecular marker. Taking into account the insertion, the estimated sizes are: F3+R3: 1583bp, F4+R4: 1709bp, F3+R1: 736bp, F2+R4: 978bp, F1+R3: 1115bp, F4+R2: 1309bp.

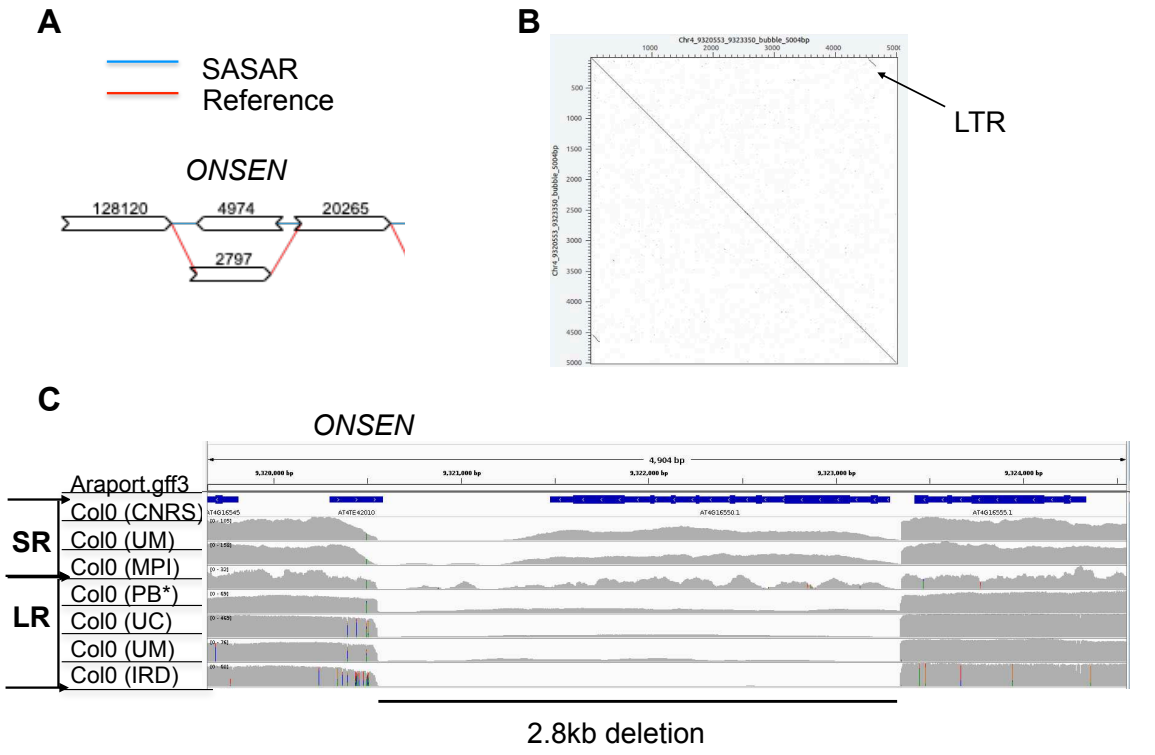


Figure II.5. SASAR improved the assembly of *ONSEN/ATCOPIA78* in *A. thaliana Col-0*. (A) Partial assembly graph consisting of chains of bubbles with the reference as the backbone. The bubble represents a SV detected in the SASAR-assembly. (B) Dot plot of the 4974bp sequence against itself shows that the SV is a LTR retrotransposon, corresponding to *ONSEN/ATCOPIA78.* (C) Misassembly of *ONSEN/ATCOPIA78* at the locus *AT4TE42010* was validated using Illumina short read (SR) and long read (LR) (ONT and PB *) from seven different institutes mapped on the reference genome TAIR10. The fragment corresponding to the 2.8kb is mis-assembled, instead it corresponds to a full-length *ONSEN* insertion, in reverse orientation.

Discussion

We have shown that there is a strong dependency of assembly contiguity on read length and sequencing coverage with reference-free approach. Although Hi-C has been widely adopted and undoubtedly improves the assembly into pseudomolecule level (Jupe et al., 2019), there are still some challenges that may hinder the ability of Hi-C to form chromosome-level pseudomolecules alone (Dudchenko et al., 2018). In principle, Hi-C data is relatively noisy, and this process relies on the aligning Hi-C reads to the draft assembly. Some small contigs with highly repetitive sequences failed to be accurately scaffolded because of conflicting Hi-C link data (Ghurye et al., 2019). This issue could be resolved by SASAR before or after using Hi-C.

Methods

Plant material

The *Arabidopsis thaliana* wildtype Col-0 plants were used in this study. Plants were grown in a growth chamber in soil under a 16h/8h (light/dark) cycle after a 2 days stratification step at 4°C. One-month-old plants (aerial parts) were harvested.

DNA preparation

Genomic DNA from plant material was extracted and ground to a fine powder in liquid nitrogen. The powder was resuspended in 10 ml of 65°C preheated CTAB2X extraction buffer (2% CTAB, 100 mM Tris-HCl pH 8, 20 mM EDTA pH 8, 1.4 M NaCl, 5% N-lauroylsarcosine di-sodium salt, 0.2% 2-mercaptoethanol) and incubated for 60 min at 65°C. Then, an equal volume of chloroform was added and the emulsion was maintained during 10 min before centrifugation at 4,500 rpm for 10 min at room temperature. The nucleic acids were precipitated with isopropanol (0.7 v/v) at -80 °C for 15 min and centrifuged at 4°C at 4,500 rpm for 45 min. Nucleic acids were further washed with 75% ethanol and centrifuged at 4°C at 4,500 rpm for 10 min. Finally, the pellet was air dried and DNA was resuspended in 300 µl TE (Tris-Hcl 10 Mm pH 8, EDTA 1mM pH 8) and was treated with 100 ng of RNase A (Qiagen, Hilden, Germany) for 30 min at 37°C. In order to purify the DNA, two steps were added: first, we used a Genomic DNA clean and concentrator column (Zymo, D4010, USA) on 10 µg of DNA

treated with RNase A; second, we precipitated the DNA with a 1/10 volume of sodium acetate 3M pH 5,2 and 2,5 volumes of ethanol 100 %. The pellet was dissolved in 100 µl TE and incubated overnight at 4°C. Both a Nanodrop and a Qubit quantification were performed to control that the ratios 260/280 and 260/230 were as recommended for the MinIon library preparation and to be sure that the results of these two different quantifications were identical or close.

MinIon library preparation

The MinIon sequencing library was generated with the sequencing kit SQK-LSK108 (ONT, Oxford, UK) according to the manufacturer's instructions. All DNA samples were end-repaired and dA-tailed using the NEBNext Ultra II end-repair/dA-tailing module (Biolabs, New England, USA, cat. no. E7546S) as per the manufacturer's instructions except for the thermocycler program performed for 30 min at 20 °C, followed by 30 min at 65°C and 5 min at 4°C. The DNA was further purified with one volume of Agencourt AMPure XP beads (Beckman Coulter, High Wycombe, UK). After two washes with 70% ethanol, beads were air-dried and the DNA was eluted with 31 µl UltraPure™ DNase/RNase-Free Distilled Water (Thermo-Fisher scientific, USA). A Qubit quantification was performed with 1 µl. A ligation was then performed by adding 50 µl Blunt/TA ligase master MIX (Biolabs, New England, US, cat. no. M0367S) and 20 µl of Adapter Mix 1D (AMX1D, ONT, Oxford, UK) to the 30 µl A-tailed library and incubated at RT for 30 min. The DNA was further purified with 0.4 X volume of Agencourt AMPure XP beads (Beckman Coulter, High Wycombe, UK). After two washes with 140 µl of adapter bead binding (ABB, ONT, Oxford, UK), beads were pelleted and air-dried a few seconds and the DNA was eluted with 15 µl of elution buffer (ELB, ONT, Oxford, UK). A Qubit quantification was performed with 1 µl.

Raw signal data processing

R9 ONT flow cells were used in this study. The number of available pores of each flow cell used was first recorded to evaluate the flow cell's quality with the MinKNOW™ software (version v0.51.1.62). The libraries were loaded on the flow cell following the manufacturer's instructions. Between 700 ng to 1650 ng of prepared library in 14 µl were added to 35 µl of RBF and 25,5 µl of LL buffers.This library mix was loaded on the flow cell via the SpotON sample. Raw signal FASTA5 files were base-called using

Guppy (v4.2.2) using the SQK-LSK108 library type. Comparison of read length distribution and quality between the runs displaying different fragmentation size for three biological samples was visualized in R using package ggplot2 (Wickham, 2009). Adapters of the ONT Ligation Sequencing Kit 1D (SQK-LSK108) were removed using Porechop (v0.2.1, https://github.com/rrwick/Porechop) with setting —discard_middle, from which reads with internal adapters were discarded. Reads sequenced from the same library of the same species were merged.

Long-read genome assemblies

Raw reads were corrected by Canu (v1.9), which includes a *de novo* correction module, using MHAP to align long reads against themselves and pbdagcon to correct the long reads by a consensus step. We applied 5 long read assemblers: Canu (v1.9), Minimap2/Miniasm (v0.3-r179-dirty) (Li, 2016), Flye (v2.7) (Kolmogorov et al., 2019), SMARTdenovo (Liu et al., 2021), Shasta and WTDBG2 (Ruan and Li, 2019) in our computational cluster with 2 nodes, 300G RAM. Canu was used to build consensus with module "–trim-assemble" and remaining parameters keep the same as described above in the correct method on each correct dataset. Flye was deployed by default parameters with an estimated genome size of corresponding genome size for three species. SMARTdenovo was performed on the Canu corrected reads and parameters were set with -c 1, -k17 as suggested by developers for large genome. WTDBG2 was run with default settings to produce assemblies.

Quality evaluation of draft assemblies

Whole genome alignments between the resulting assemblies and the corresponding reference genomes were rerun by minimap2 with two degrees of alignment divergence, module "asm5" for up to 1%, module "asm20" for up to 10%. Identity percentage was calculated by stats_from_paf.py. Genome–genome alignment dot plots were visualized using D-GENIES (Cabanettes and Klopp, 2018). Quality metrics for each assembly were produced using "stats_from_asm.py" and BUSCO (v3) (Waterhouse et al., 2018) using corresponding reference genomes. BUSCO was run with default parameters. The 'embryophyta_odb9' was used as a reference gene set. To assess the assembly resolution of repeated regions, centromeric repeats of *A. thaliana* (178 bp) were mapped to each corresponding draft assembly using blastn with a 100bp match

threshold. To count alignment breakpoints, we mapped all assemblies to the corresponding reference genomes with minimap2 under the option '--pafno-hit -cxasm20 -r2k -z1000,500'. Structural variant was called by paftools.js to collect various metrics (Whole command line: bash sv_from_asm.sh). N-stretches in the corresponding reference genomes were extracted using the script "findN.py" to check for misassembled regions. Breakpoints of shared structural variants detected were compared with N-stretch coordinates using bedtools intersect (Quinlan and Hall, 2010).

Data availability

Raw MinION sequencing data for *A. thaliana Col-0* were deposited in the European Nucleotide Archive under the project name PRJEB34954, with samples ERS3901322. For validation of mis-assembly in TAIR10, Illumina sequencing data of *A. thaliana Col-0* (MPI: Weigel's lab) were extracted from Sequence Read Archive SRR013327. Reads from the *S. pennellii* acc. LA5240/LYC1722 were obtained from https://plabipd.de/portal/solanum-pennellii.

Reference

Cabanettes, F., and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 6, e4958. doi:10.7717/peerj.4958.

Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C., Musial, N. T., Mostofa, R., et al. (2018). The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*, 254797. doi:10.1101/254797.

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., et al. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology* 15, e1007273. doi:10.1371/journal.pcbi.1007273.

Gilly, A., Etcheverry, M., Madoui, M.-A., Guy, J., Quadrana, L., Alberti, A., et al. (2014). TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *BMC Bioinformatics* 15, 377. doi:10.1186/s12859-014-0377-z.

Jupe, F., Rivkin, A. C., Michael, T. P., Zander, M., Motley, S. T., Sandoval, J. P., et al. (2019). The complex architecture and epigenomic impact of plant T-DNA insertions. *PLOS Genetics* 15, e1007819. doi:10.1371/journal.pgen.1007819.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37, 540–546. doi:10.1038/s41587-019-0072-8.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). 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.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32, 2103–2110. doi:10.1093/bioinformatics/btw152.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi:10.1093/bioinformatics/bty191.

Liu, H., Wu, S., Li, A., Ruan, J., Wu, S., Li, A., et al. (2021). SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte* 2021, 1–9. doi:10.46471/gigabyte.15.

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Schmücker, A., et al. (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* 374, eabi7489. doi:10.1126/science.abi7489.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033.

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 17, 155–158. doi:10.1038/s41592-019-0669-3.

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., et al. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* 38, 1044–1053. doi:10.1038/s41587-020-0503-6.

Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., et al. (2021). High-quality Arabidopsis thaliana genome assembly with nanopore and HiFi long reads. *Genomics, Proteomics & Bioinformatics*. doi:10.1016/j.gpb.2021.08.003.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* 35, 543–548. doi:10.1093/molbev/msx319.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag Available at: https://www.springer.com/gp/book/9780387981413 [Accessed July 29, 2019].

2.3 Tracking TE mobility and genome instability with long reads

2.3.1 My contribution to the discovery of genome instability in Arabidopsis epigenetic mutants

When I first got in touch with the long-read data of eccDNA-seq, I manually did a lots of dotplots to understand the fine structure of active TEs. I observed different structures of a certain TE, incomplete circles and chimeric circles. To investigate the origin of chimeric eccDNA including partial genes and partial TEs, we performed whole genome sequencing (WGS) using long reads.

However, there was no dedicated tools to detect TE insertion polymorphisms (TIPs) from long read sequencing that time (around 2020), I first developed a script based on the TRACKPOSON algorithm, namely CIGAR_SV (https://github.com/njaupan/CIGAR_SV). The CIGAR_SV uses CIGAR from read alignment to output insertion and deletion sequences. With the observation of partial TE in the WGS data, I further focused on the chimerism in the genome, especially "3-hit" reads that is to say reads mapping to three different locations in the genome (hence the name « Chimera » for the manuscript).

While focusing on the TIPs for each active TE, I found that there was a 2Mb large sequence inverted in the genome of the hypomethylated *ddm1* mutant. By performing many dotplots, I realized it was a true inversion. Then I collected all the public data associated with *ddm1* to understand the origin of inversion. I downloaded 123 EpiRILs WGS data and calculated the occurrence of the inversion. We produced two new WGS datasets of *ddm1* using ONT sequencing. In this material I further detected large duplications, opening a door to discuss the epigenetic control of genome stability.

In writing this manuscript, Hajk brought a lot of brainstorming to discuss the function of eccDNA, and Marie provided a lot of thoughtful discussion to help me understand the role of epigenetics. I hope to publish this work soon after we submit the thesis.

2.3.2 « Chimera » manuscript (bioRxiv, 2022)

eccDNA and structural variants analysis reveals massive genome instability in Arabidopsis epigenetic mutants

Panpan Zhang[1,2], Christel Llauro[2,3], Alain Ghesquière[1], R. Keith Slotkin[4], Frédéric Pontvianne[2,3], Marie Mirouze[1,2]

1 Institut de Recherche pour le Développement (IRD), UMR232 DIADE, 911 Avenue Agropolis, 34394 Montpellier, France
2 University of Perpignan, Laboratory of Plant Genome and Development, 58 Avenue Paul Alduy, 66860 Perpignan, France
3 Centre National de la Recherche Scientifique (CNRS), Laboratory of Plant Genome and Development, 58 Avenue Paul Alduy, 66860 Perpignan, France
4 Donald Danforth Plant Science Center, St. Louis, MO, 63132, USA

Abstract

The epigenome controls transposable element (TE) mobility and mutants affected in the epigenetic machinery display active TEs associated with extrachromosomal circular DNA (eccDNA). However, the interplay between eccDNA and genome stability is poorly understood. Here we show that *Arabidopsis* plants combining mutations in a chromatin remodeling, post-transcriptional silencing and RNA-directed DNA methylation have a high eccDNA load associated with integration of truncated TEs and genome instability. We analyzed the eccDNA and genome sequence of *ddm1 rdr6 pol4* (*Decrease DNA methylation 1, RNA dependent RNA polymerase 6, RNA polymerase 4*) triple mutant plants and uncovered TE mobility of full length and partial TEs. Additionally, TE movement was associated with gene movement of a disease resistance cluster named *RPP5*. We further discovered a large 2 Mbp inversion and show that this inversion is also present in *ddm1* single mutant plants, probably since its isolation 30 years ago. Finally, long read sequencing allowed the detection of two independent ~55 kb duplications in *ddm1* siblings. Our results highlight the role of the epigenome in protecting the genome not only against TE mobility but also against chaotic genome rearrangements and eccDNA-driven gene chimerism and reinforce the concept of a two-speed genome evolution in Arabidopsis, guided by the epigenome.

Key words: eccDNA, structural variant, *ddm1*, ONT, *Arabidopsis*

Introduction

Extrachromosomal circular DNA (eccDNA) has been described for decades in eukaryote cells including yeast, Drosophila, nematodes, plants and humans (Hotta and Bassel, 1965; Hirochika and Otsuki, 1995; Sinclair and Guarente, 1997; Cohen and Méchali, 2002; Kumar et al., 2017). Only recently this fraction of the cell genetic material has gained attention thanks to the development of specific eccDNA-seq sequencing approaches allowing its characterization (Møller et al., 2015; Lanciano et al., 2017). EccDNA is found in all eukaryotes cells where it originates from spurious homologous recombination between tandem copies (for example ribosomal DNA or telomeric DNA) or between micro homologies (Shibata et al., 2012) or from linearization of extrachromosomal linear DNA of active transposable elements (Møller et al., 2016; Lanciano et al., 2017) through HR or NHEJ. EccDNA is associated with senescence in yeast and with apoptosis in mammal cells (Wang et al., 2021; Arrey et al., 2022). In cancer cells eccDNA is abundant and can originate from chromothripsis, a phenomenon describing a massive and catastrophic event of genome rearrangement. In these cells eccDNA is associated with gene amplification and contributes to tumor evolution. The adaptive role of eccDNA has also been demonstrated in plants where eccDNA encoded genes can contribute to herbicide resistance in weed species (Koo et al., 2018). Despite the growing literature on eccDNA, its impact on the genome is not well described. In cattle an early work suggested that eccDNA could be linked to structural variations (Durkin et al., 2012). Other indirect evidence links the presence of highly active TEs and structural variations (Hufford et al., 2021). However, the lack of ongoing TE mobility has prevented a comprehensive analysis of the impact of eccDNA on genomic structural variants.

We thought to address this question in Arabidopsis by analyzing the eccDNA repertoire of mutants with a high eccDNA load. In plants TEs are controlled at different steps in their life cycle by a combination of epigenetic regulations involving notably DNA methylation (for gene silencing) and post-transcriptional gene silencing (PTGS) (Sigman and Slotkin, 2016; Nicolau et al., 2021; Lloyd and Lister, 2022). DNA methylation is maintained directly by methyltransferases and by the RNA-directed DNA methylation pathway (RdDM). It is also indirectly maintained by the chromatin

remodeler DDM1 (Decrease DNA Methylation 1) involved in the deposition of the heterochromatic histone variant H2A.W (Osakabe et al., 2021) at full length TEs. In order to increase the load of TE-driven eccDNA in *Arabidopsis thaliana*, we selected mutant plants with mutated *DDM1*, *RDR6* (involved in PTGS) and *NRPD1* or *POL4* (involved in RdDM). Mutant plants affected in one or a combination of these pathways have a high level of TE transcription (Panda et al., 2016; He et al., 2021a). We used eccDNA-seq and long read sequencing of their genome to analyse the impact of eccDNA on genomic stability. We discovered that the triple mutant plants display a high level of TE-derived eccDNAs originating from different TE families of LTR retrotransposons (such as *EVD/ATCOPIA93*, and *ATCOPIA21)* and DNA transposons (*VANDAL21*). Thanks to long read eccDNA-seq we could uncover the structure of these eccDNAs and show that while a fraction of the eccDNA is full-length as expected, the majority consist of truncated circles. Analyzing the genomic content of these mutant plants, we uncovered examples of truncated insertions of these TEs, suggesting that truncated copies are capable of integration. Most notably, we also identified chimeric gene-TE eccDNAs and show evidence of genomic integrations of such gene-TE elements. Finally, we serendipitously discovered large structural variations that promoted us to analyze the SVs in the single *ddm1* mutant. One large 2 Mbp inversion was detected in *ddm1* and could originate from its original EMS screen. In contrast, two 55 and 56 kb duplications were identified in single *ddm1* plants and suggest a high level of genome instability in this genetic background. Our work highlights the hidden consequence of lack of DDM1 on genomic stability and suggest that epigenetic control of genomic stability goes beyond TEs.

Results

Diverse TE families are present in the eccDNA repertoire of epigenetic mutants, and trigger new genomic insertions

We used the mobilome-seq or eccDNA-seq approach (Lanciano et al., 2017) to sequence eccDNAs from WT, *ddm1*, *ddm1 pol4*, *ddm1 rdr6* and *ddm1 pol4 rdr6* mutant plants using Illumina short reads. Briefly linear genomic DNA was digested and the circular DNA was amplified using random primed rolling circle and sequenced. The enrichment of the mapped reads as well as the circle-specific head-to-tail reads corresponding to eccDNAs were analyzed using ecc_finder (Zhang et al., 2021). The

two most abundant TE-eccDNAs corresponded to the long terminal repeat (LTR) retrotransposon *EVD* (Mirouze et al., 2009) and the DNA transposon of the Mutator family *VANDAL21* (Tsukahara et al., 2009). These eccDNAs were detected in *ddm1*, *ddm1 pol 4*, *ddm1 rdr6* and *ddm1 pol4 rdr6* libraries (Figure II.6A). Noticeably, *EVD*-eccDNA and *VANDAL21-eccDNA* reads accounted for 130,000 per million reads and 80,000 per million reads, respectively, in the *ddm1 pol4 rdr6* triple mutant library. On top of these highly represented eccDNAs, we detected eccDNAs from the LTR retrotransposons *ATCOPIA51*, *ATCOPIA52* and DNA transposon *VANDAL3* families (Figure II.6A), where the mobility of LTR retrotransposons is consistent with active reverse transcription (Panda and Slotkin, 2020) and VLP formation (Lee et al., 2020) in these mutant backgrounds. In order to capture the full picture of the eccDNA circular structures, we performed eccDNA-seq using ONT long reads (Lanciano et al., 2021; Zhang et al., 2021) on the *ddm1 pol4 rdr6* triple mutant plants. This analysis further validated that *EVD* (Figure II.6B) and *VANDAL21* (Figure II.6C) lead to the formation *bona fide* eccDNAs but also truncated ones (Figure II.6D-F).

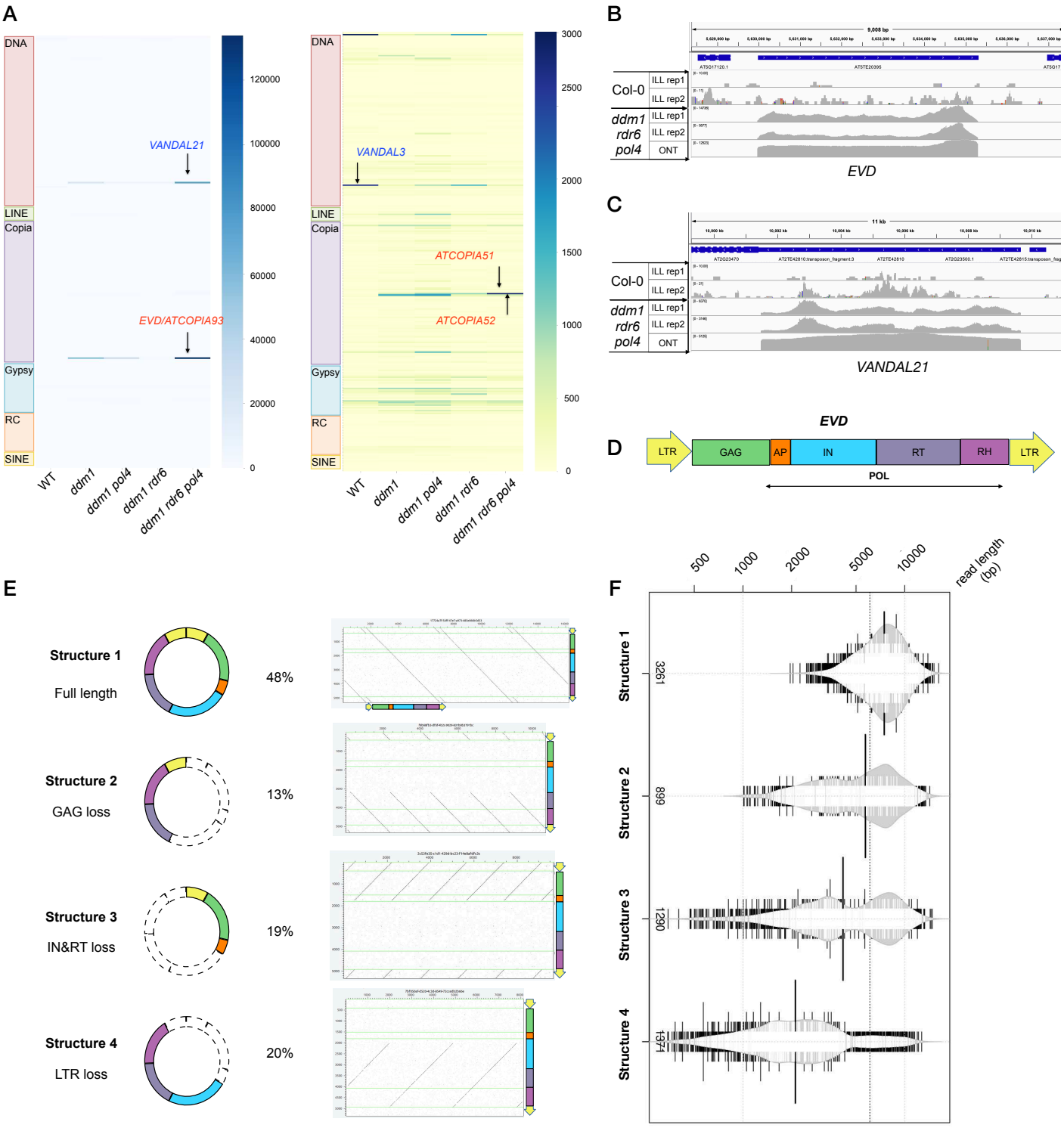


Figure II.6. The eccDNA repertoire in *A. thaliana* epigenetic mutants contains active full length and truncated TEs. (**A**) Heat map showing the profile of the extrachromosomal circular TEs among all 318 *Arabidopsis* TE families in Col-0, *ddm1*, *ddm1 pol4*, *ddm1 rdr6* and *ddm1 rdr6 pol4* triple mutant plants using Illumina eccDNA-seq (left panel). The number of circular TE is normalized by the number of mapped reads per million. Low abundant eccDNAs are visualized on a heat map without the top two abundant TEs (*EVD* and *VANDAL21*) (right panel) (**B**) **&** (**C**) Coverage of TE circles in the *ddm1 rdr6 pol4* triple mutant using short read (replicates rep1, rep2) and long read sequencing compared to the *Arabidopsis* wild-type Col-0 (replicates rep1, rep2) at the *EVD* (**B**) locus and the *VANDAL21* (**C**) loci (Illumina: ILL, Nanopore: ONT). ONT reads result in clearer boundaries and uniform coverage for active TEs. (**D**) Schematic view of *EVD* strcutre: 2 LTRs, a GAG (green) and a polyprotein (POL) cleaved into 4 active functional domains: AP (orange), IN (cyan), RT (purple) and RH (dark orchid). (**E**) Different *EVD*-eccDNA structures detected using ONT eccDNA-seq. Structure 1: A single read contains more than 2 full-length copies of *EVD*; Structure 2: A single read contains more than 5 incomplete *EVD* copies with lost GAG, AP and IN domains. Another single read contains more than 12 incomplete *EVD* copies with lost GAG, AP, IN and RT domains; Structure 3: A single read contains more than 5 incomplete *EVD* copies with lost IN and RT domains. Another single read contains more than 8 incomplete *EVD* copies with lost AP, IN and RT domains; Structure 4: Incomplete *EVD* with the loss of LTR (Figure 5E). (**F**) Distribution of read length and percentage of different *EVD* structures.

In order to detect new insertions of these TE families and compare with the eccDNA repertoire, we performed ONT genome resequencing of WT, *ddm1*, *ddm1 pol4*, *ddm1 rdr6* and *ddm1 pol4 rdr6* mutant plants. We detected insertions corresponding to the two most abundant TE families as eccDNA: *EVD* and *VANDAL21*, with up to 73 new insertions for *EVD* in the triple *ddm1 pol4 rdr6* mutant (Figure II.7A). Additionally, we detected insertions from two DNA transposons *ATENSPM3* and *ATMU5. Copia* retrotransposons (*EVD* and *ATCOPIA21*) inserted preferentially within exons, *VANDAL21* mainly targeted the 5' UTRs of active genes, and *ATENSPM3* insertion sites were more widely distributed next to genes (Figure II.7B). Their integration patterns are consistent with the discovery of the preferentially insertion sites in the *ddm1* derived epiRIL population (Quadrana et al., 2019). We did not detect any new insertion for the TE families *ATCOPIA51*, *ATCOPIA52* and *VANDAL3,* suggesting a dose dependent effect or additional mechanisms of control preventing their integration.

The eccDNA repertoire of epigenetic mutants contains truncated TEs, associated with truncated genomic insertions

A complete *Ty1/Copia* retrotransposon contains 2 LTRs, a GAG domain and a polyprotein (POL), which is cleaved into four active functional domains: AP, an aspartic protease, IN, an integrase, RT, a reverse-transcriptase and RH, a RNase H (Figure II.6D). On top of eccDNAs corresponding to expected full length *EVD* (Figure II.6E, structure 1), we identified eccDNAs with partial structures: loss of GAG domain (structure 2), loss of IN and RT domains (structure 3), loss of LTR (structure 4) (Figure II.6E). These truncated *EVD*-eccDNAs account for 52% of all *EVD*-eccDNAs (Figure II.6F), indicating that full length eccDNAs are probably only the tip of the iceberg of retrotransposon eccDNAs.

Given the high frequency of these partial eccDNAs we analyzed their potential impact on genomic SVs. In this aim we selected *EVD* and *COPIA21* containing reads in our ONT genome resequencing datasets for WT, *ddm1*, *ddm1 pol4*, and *ddm1 pol4 rdr6* mutants. We excluded *ddm1 rdr6* as there was no new detected retrotransposon insertion in this mutant background, for unknown reason. Insertions not present in the reference genome and corresponding to truncated structures were detected at 6 distinct loci (Figure II.7C-F). Some of these insertions of truncated retrotransposons contain clear target site duplications (TSDs). This observation suggests two possible scenarios.

Truncated copies of *EVD* and *ATCOPIA21* might form during reverse transcription and lead to linear and circular extrachromosomal DNA capable of new integrations. Alternately, recombination and deletion could have occurred soon after integration.

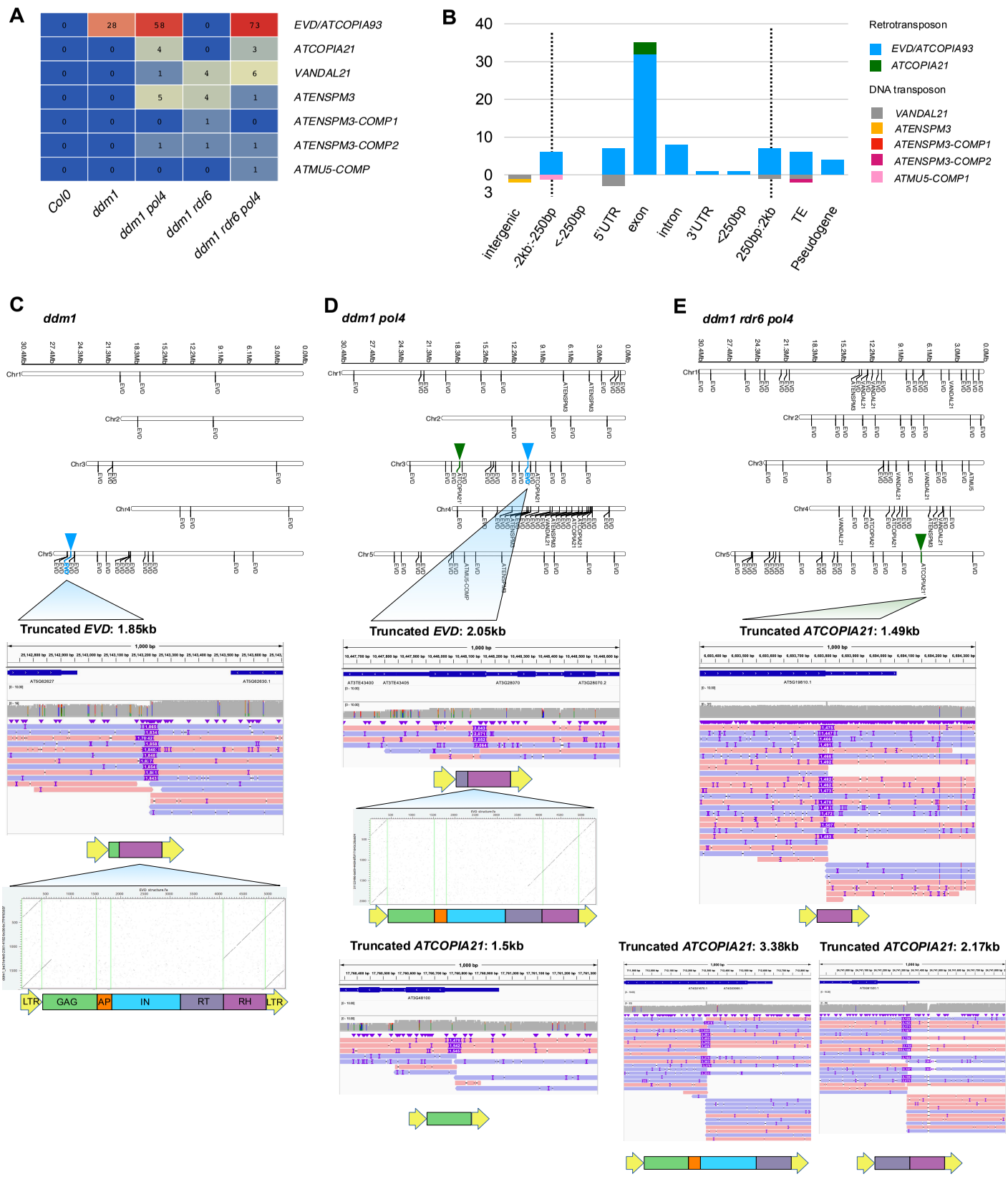


Figure II.7. TE insertion polymorphisms in *A. thaliana* epigenetic mutants revealed by ONT genome sequencing. (**A**) Insertion numbers of 7 mobilized TE families in the genome of the *ddm1, ddm1 rdr6, ddm1 pol4* and *ddm1 rdr6 pol4* mutants. (**B**) Target site preferences of different TE families in the *ddm1 rdr6 pol4* genome. (**C**), (**D**) and (**E**). Localisation of new insertions of *Copia* retrotransposons. Truncated insertions are highlighted (*EVD*, blue triangle, *ATCOPIA21* green triangle). The above figure shows the location of TE insertion sites in the corresponding genome: *ddm1* (**D**), *ddm1 pol4* (**E**) and *ddm1 rdr6 pol4* (**F**). For each new insertion of a truncated TE, an IGV view with a window size of 1kb is shown.

TE mobility can lead to gene mobility

In the ONT eccDNA-seq data from the triple mutant plants, we observed a striking example of chimeric eccDNA, defined as a circle originating from different genomic loci. In this *EVD*-eccDNA, supported by 7 copies in a single long read (Figure II.8A), each copy consists of a partial *EVD* fragment and a portion of the *AT5G66440* gene. To eliminate potential errors caused by ONT sequencing, *EVD* containing reads were extracted from both ONT and Illumina eccDNA-seq datasets for the triple mutant plants and mapped to the TAIR10 reference genome (Figure II.8B). The mapping profile for both eccDNA-seq datasets confirmed the presence of chimeric reads at the *AT5G66440* locus. The eccDNA repertoire of *ddm1 pol4 rdr6* plants thus contains chimeric eccDNAs containing a retrotransposon and a gene. To characterize the genomic locus at this gene we analyzed *ddm1 rdr6 pol4* ONT genomic data and detected an *EVD* insertion into the same gene (*AT5G66440*) at the same position, suggesting that the chimeric eccDNA corresponds to a new copy of *EVD* (Figure II.8C). Further, the insertion of *EVD* generated a 5 bp target site duplication, a signature of integrase-mediated insertion (ACGAA) (Figure II.8D).

To investigate whether this chimerism was due to the *EVD* TE family or could be detected for other TE families, we extracted *ATCOPIA21* long reads from two replicates of *ddm1 rdr6 pol4* ONT genomic datasets. We noticed that these *ATCOPIA21* containing long reads displayed complex re-arrangements when mapped to the TAIR10 reference genome (Figure II.9A). They were divided into three segments mapping to distinct genomic regions, and here-after referred to as "3-hit" reads. One "3-hit" read started from the *AT4G16950* gene encoding RPP5 (for RECOGNITION of PERONOSPORA PARASITICA 5), then spanned *ATCOPIA21*, and ended at the *AT4G16970* gene encoding a CRK19 (cysteine-rich RECEPTOR-like protein kinase 19) located 3 Mb away from the *RPP5* cluster (Figure II.9B). The insertion of *ATCOPIA21* at the CRK19 gene created an 8bp target site duplication (TATAGTAG) showing a proper integration event (Figure II.9C). Considering this data, we propose a model to explain how the *RPP5* gene moved close to the *CRK19* gene in the *ddm1 rdr6 pol4* triple mutant (Figure II.9D). The *RPP5* locus is located within a resistance (R) gene cluster that plays an important role in the innate immune response to pathogens in the *Arabidopsis thaliana* (Yi and Richards, 2007). We further detected TE polymorphisms in 64 natural accessions re-sequenced with long reads (Van de Weyer et al., 2019)

suggesting that the *RPP5* locus is a hotspot for TE insertions or that TE-polymorphisms are selected for in natural populations at this locus (not shown).

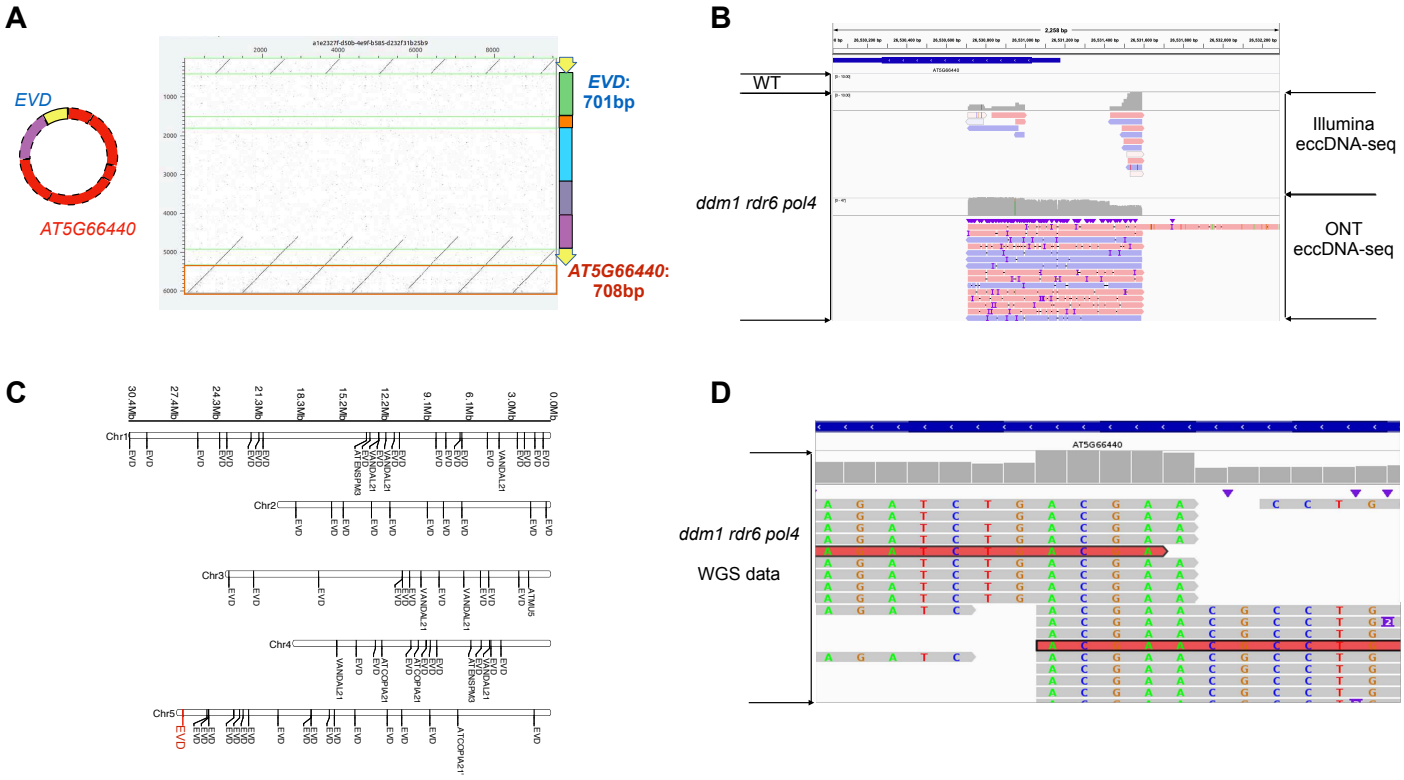


Figure II.8. Chimeric eccDNA containing a truncated *EVD*-gene fusion corresponding to a chimeric genomic integration in *A. thaliana ddm1 rdr6 pol4* mutant. (**A**) A chimeric eccDNA consisting of truncated *EVD* (701bp) fused to a truncated *AT5G66440* gene fragment (708bp). (**B**) eccDNA-seq reads (Illumina and ONT) containing *EVD* were selected and mapped to the TAIR10 reference genome and shown at the *AT5G66440* locus. The chimeric *EVD* reads were detected by both short and long read eccDNA-seq in the *ddm1 rdr6 pol4* mutant. (**C**) The chimeric *EVD-AT5G66440* corresponds to a new copy of *EVD* (in red) from whole genome re-sequencing. (**D**) The *EVD* insertion at *AT5G66440* created a 5bp target site duplication.

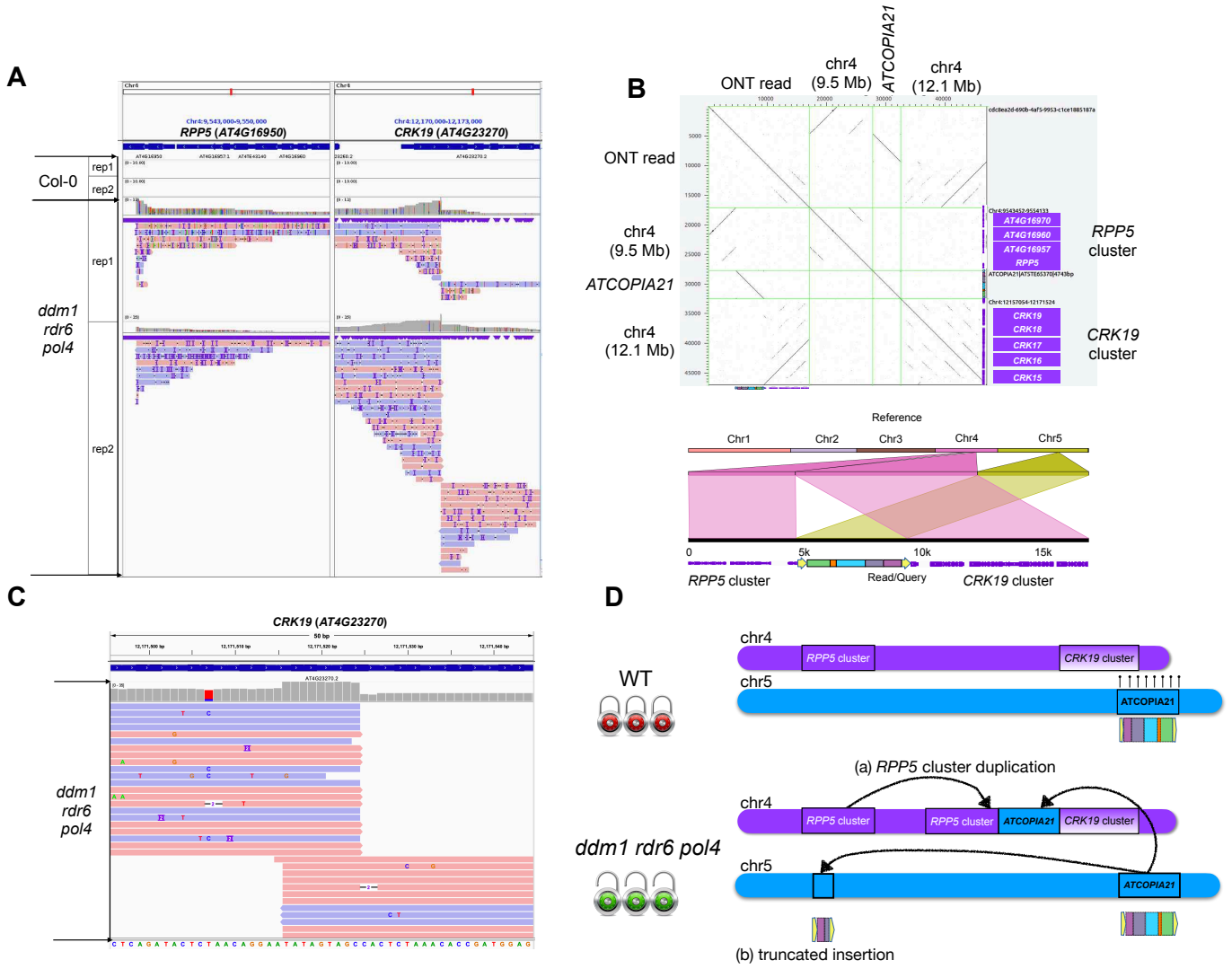


Figure II.9. *ATCOPIA21* mobility in the *A. thaliana ddm1 rdr6 pol4* triple mutant is associated with *RPP5* locus duplication. (**A**) ONT genomic reads from *ddm1 rdr6 pol4* and containing *ATCOPIA21* were selected and mapped to the TAIR10 reference genome. These aligned reads are displayed on IGV, at both the *RPP5* (*AT4G16950,* c. 9.5 Mb on chr. 4) and the *AT4G23270* loci (c. 12.1 Mb on chr. 4). (**B**) Dot plot of an ONT read from *ddm1 rdr6 pol4* genome mapping to 3 different genomic loci as indicated. (**C**) ONT reads showing a target site duplication generated by *ATCOPIA21* insertion at the *CRK19* locus in *ddm1 rdr6 pol4*. (**D**) Towards a model for the consequences of *ATCOPIA21* mobility in *ddm1 rdr6 pol4*, leading to *RPP5* locus duplication (a) and truncated new insertions (b). Chromosomes are not drawn to scale.

Structural variants in *ddm1* background go beyond TEs

To explore further the SVs in the genomes of *ddm1* single mutant and *ddm1 rdr6 pol4* triple mutants we assembled the corresponding genomes using ONT sequencing reads. For *ddm1* we used data obtained from two single plants with a coverage of 40.7x and 18.6x, respectively. For the triple mutant, as the plants were smaller, we used two datasets corresponding to five or six pooled plants (83.3x and 57.5x respectively). In the *ddm1* genome assembly, a fragment in a contig (tig00000083) was reversed end-to-end on chromosome 2 compared to the TAIR10 reference genome, suggesting a 2 Mbp inversion (Supp. Fig.1). This large inversion was also detected in the same region on chromosome 2 in the *ddm1 rdr6 pol4* genome assembly. This inversion (named Chr2-2M) breaks at the *Gypsy* retrotransposon *ATGP1* (*AT2TE07550*) in 5' and at a hypothetical gene (*AT2G07806*) in 3'. This gene corresponds to mitochondrial DNA integrated into the nuclear genome (Saccone et al., 2000) and is located only 117 kb away from the centromere, in the pericentromeric region. To verify that the Chr2-2M inversion was not due to genome mis-assembly, ONT reads spanning the two breakpoints were extracted and aligned to the TAIR10 reference genome. Dot plots of the reads to the reference validated the accuracy of the Chr2-2M inversion. Furthermore, the mappings of both ONT and Illumina reads were compared at the two breakpoints and showed consistent results (Supp. Fig.1). Surprisingly, the Chr2-2M inversion resulted in a 5 bp duplication (AATCT) and an 8 bp deletion (AGATGGTT), which facilitated the detection of the inversion. We noticed that the start of the inverted fragment (GTGA) is the reverse complement (with one mismatch) to the start of the second fragment (CTCA) suggesting that a micro homology might exist between the two breakpoints. Finally, the inversion was validated using PCR amplification (not shown). Of note, the Chr2-2M inversion breakpoints corresponds to cold points in the Hi-C map obtained by Feng et al. (2014).

In order to trace the origin of the Chr2-2M inversion, we thought to detect inversion from all available *ddm1* whole-genome sequencing data, including *ddm1-1*, *ddm1-2*, and *ddm1-2* derived epiRILs (Tsukahara et al., 2012; Fu et al., 2013; Cortijo et al., 2014). We could detect the Chr2-2M inversion in the genomes of *ddm1-2* generated by Vongs et al. in 1993 and in *ddm1-2* derived epiRiL generated by Johannes et al. in 2009 as well as *ddm1-1* generated by Vongs et al. in 1993 (Supp. Fig.1). In epiRILs 67 of the 123 lines that we re-analyzed for the inversion contained the homozygous Chr2-2M

inversion. The probability that epiRILs selected from eight generations of self-crosses generated after two backcrosses with wild-type Col-0 contained the Chr2-2M inversion is 55%, which is consistent with our observations (54,4%). Given the fact that the Chr2-2M inversion could be detected in both *ddm1* alleles, it most likely occurred during or before the EMS mutagenesis of wild-type Col-0 that was used to generate the *ddm1* mutants back in 1993. As the *DDM1* gene is located on chromosome 5, it is nevertheless intriguing that the inversion was not segregated away during the multiple backcrosses of this mutant over nearly 30 years. Nevertheless, the Chr2-2M inversion was not found in *ddm1 pol4*, nor in the recent whole-genome sequencing of the *ddm1-1* mutant from Zhu's group (He et al., 2021b), demonstrating that it can actually be segregated away.

Interestingly the left Chr2-2M breakpoint at *ATGP1* is completely hypomethylated in *ddm1* plants (not shown). Given that it was not possible to obtain whole-genome sequencing data of *ddm1* generated 30 years ago, the mechanism of the Chr2-2M inversion and its function in relation to DDM1 remains unclear. This observation nevertheless reveals that long studied mutants might still hide surprising genomic alterations.

Occurrence of large duplication events in the genome of Arabidopsis *ddm1-2* mutants

In addition to the Chr2-2M inversion, we detected a 55kb duplication (Chr1:5,548,395-5,603,615) and a 56kb duplication (Chr2:231,518-287,731) in two *ddm1* siblings, respectively. We use the detection of the 55kb duplication (starting at the 5'UTR of *AT1G16220* and ending within a *AT1G16390* exon) as an example. The coverage of *ddm1* ONT reads aligned to the TAIR10 reference genome increased 2-fold at the region of Chr1:5,548,395-5,603,615 whereas the coverage of Col0 reads showed a flat distribution (Figure II.10A). We analyzed long reads crossing the junction of two tandem copies (Figure II.10B) and confirmed the 55kb duplication by dot plots (Figure II.10C). We did not detect SNPs in the two tandem copies suggesting a recent event. To investigate the relationship between this new tandem repeats and DDM1-related DNA methylation, we examined the cytosine methylation pattern of the genes located at the breakpoints of the 55kb region. For this purpose, we used genes overlapping the junctions that could be unambiguously mapped. This showed that the second copy of

the two tandem replicates is hypomethylated at the left breakpoint, while the first one is not (Figure II.10D). The second detected 56kb duplication starts in the exon 2 of the *AT2G01510* gene and ends between two DNA transposons (*AT2TE01165* and *AT2TE01180*) (not shown).

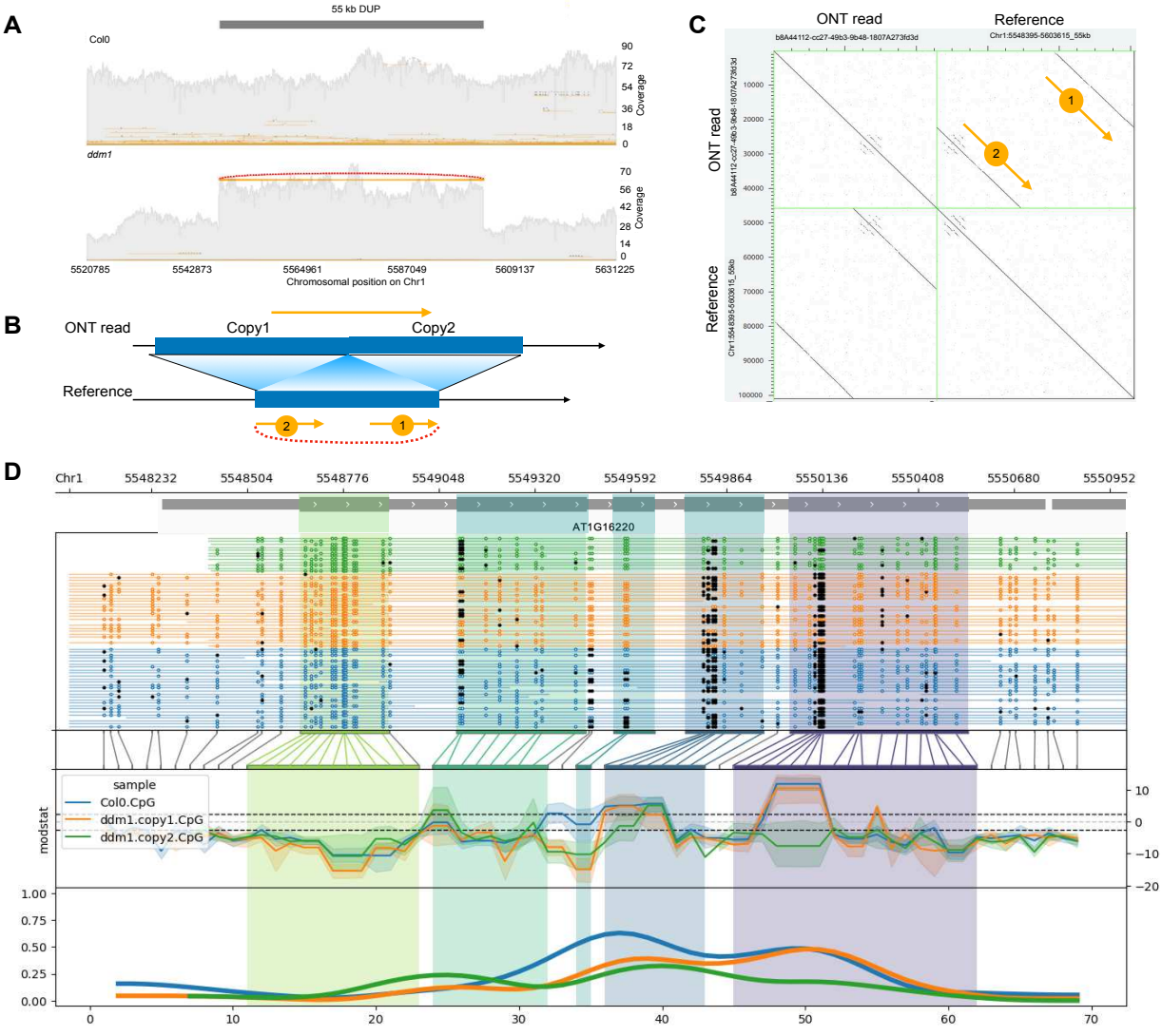


Figure II.10. Detection of a 55kb duplication on chromosome 1 in the *A. thaliana ddm1*-2 mutant genome. (**A**) Read depth of the *A. thaliana* wild type Col-0 and *ddm1-2* ONT reads aligned to the TAIR10 reference genome on chromosome 1. (**B**) Scheme indicating how the ONT reads (orange arrows) spanning the junction of the two tandem copies will be split and aligned on the reference. (**C**) Dot plot showing a raw 46kb long ONT read versus the reference at the duplicated region on chromosome 1. (**D**) Cytosine methylation at the duplicated gene at the start of the 55kb duplication. From top to bottom, the plot shows the genome coordinates, gene transcripts, ONT read mapping with modified bases as closed (methlated) or open (unmethylated) circles, raw log-likelihood ratios, and smoothed methylated fraction plot. Exons of the *AT1G16220* gene are highlighted.

Discussion

Truncated TEs in eccDNAs and as new genomic integrations

In this study, we investigated the overall genomic stability of *A. thaliana* epigenetic mutants in terms of eccDNA content and SVs (Figure II.11). In the eccDNA repertoire of *ddm1 rdr6 pol4* mutant plants, we detected a majority of eccDNAs composed of truncated TEs. Some truncated TE integrations were further detected in the *ddm1 rdr6 pol4* genome. eccDNA coming from TE lifecycle is generally considered as a dead-end, not being able to re-integrate the genome (Garfinkel et al., 2007). However, in yeast and cattle, integration of eccDNA has been suggested (Durkin et al., 2012; Thierry et al., 2015). More recently, studies in cancer cells have shown that eccDNAs that contain oncogenes (Verhaak et al., 2019; Wu et al., 2022) are able to integrate back into the genome (Koche et al., 2020). In our study, we present evidence that suggest the integration of truncated TEs in planta. Whether truncated eccDNAs are formed at the transcript level, in the VLP, or post-integration in the genome requires further investigations. Truncated TE insertions were also detected in the genome of *ddm1* single mutant and *ddm1 pol4* double mutant, suggesting that the *ddm1* mutant background itself is sufficient to promote these truncated TE integrations.

TE-induced SVs in *ddm1 rdr6 pol4*

SVs are enriched for repeated DNA including TEs (Audano et al., 2019; Carvalho and Lupski, 2016; Krasileva 2019), as exemplified for NLR genes in pepper retrogenes (Kim et al., 2017) or for the *sun* locus in tomato (Xiao et al., 2008). More recently, large scale studies of SV in tomato identified repeats in around 80% of SVs (Alonge et al., 2020). In a rice pan-genome analysis, half of the SVs (for the c. 80,000 for which they could assign a possible mechanism) were associated with TE insertions (Qin et al., 2021). To date, most TE-mediated SVs have been identified in natural variants, and the mechanism underlying these TE-mediated SVs is not yet clear. Here we show that retrotransposon mobility can lead to gene cluster duplication in *ddm1 rdr6 pol4* mutants. We have not observed TE-mediated SVs at genes in the single *ddm1* mutant suggesting that a heavy load of TE mobility is necessary to induce SVs.

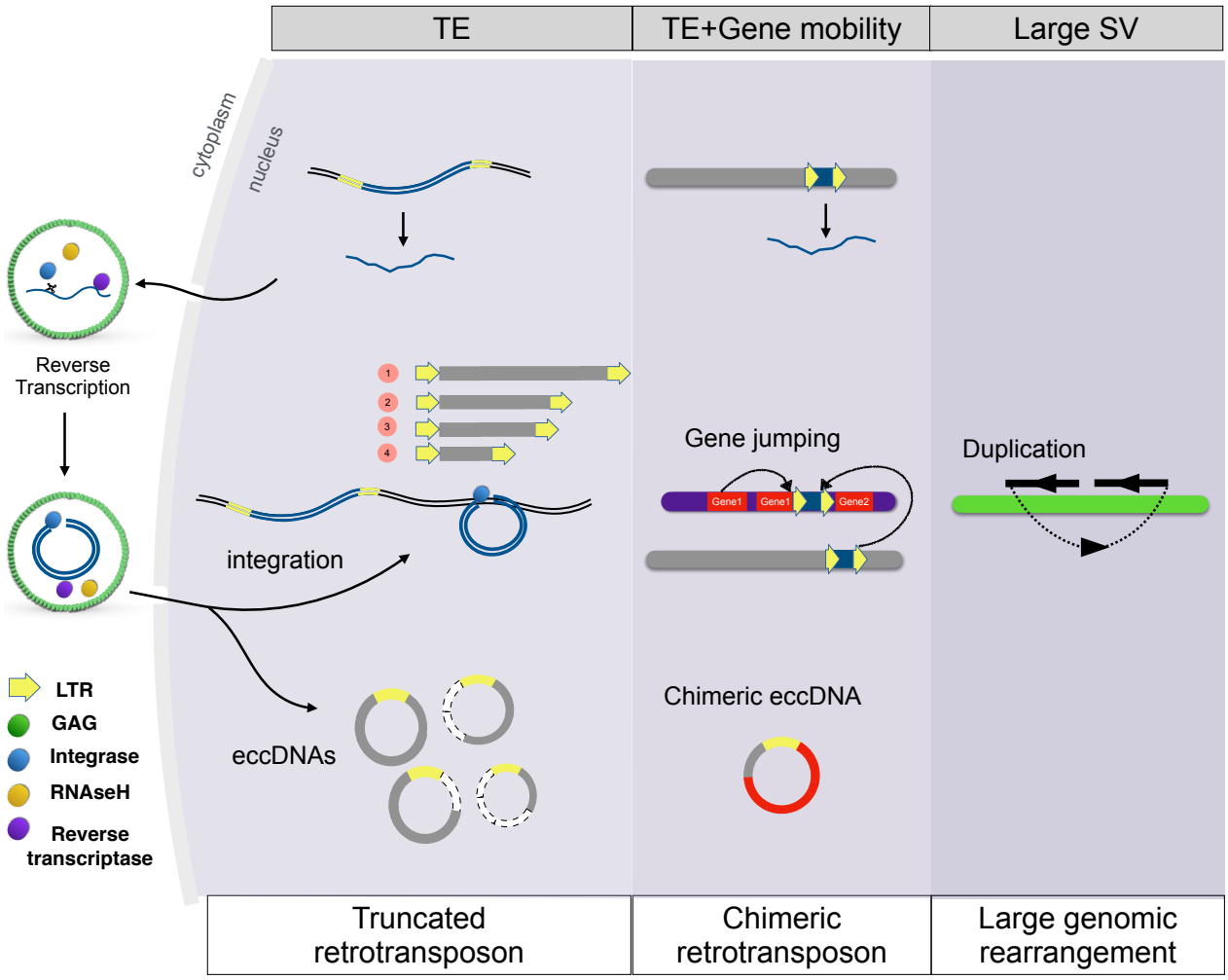


Figure II.11. Overall genomic instability detected in this study in *A. thaliana* epigenetic mutants. The role of the epigenome in protecting the genome not only against transposable element mobility but also against gene chimerism as well as chaotic genome rearrangements is highlighted.

ddm1 mutants carry a 2Mb inversion

Large genomic rearrangements are frequently found in T-DNA mutants and mutagenized plants (Jupe et al., 2019; Pucker et al., 2021). The 2Mb inversion that we have uncovered in *ddm1* mutant plants probably occurred nearly 30 years ago during the original EMS screen, as two alleles from this screen (*ddm1-1* and *ddm1-2*) display the inversion. This inversion segregated in a Mendelian fashion in the *ddm1*-epiRIL population, and was recently segregated away in *ddm1* (He et al., 2021b). The reason why this inversion was never segregated away in our *ddm1* plants remains obscure. It could have been present in the original line with which the *ddm1* alleles were backcrossed (Vongs et al., 1993). Nevertheless, this observation illustrates the power of long reads in detecting SVs in long-studied mutants and advocate for a closer examination of the genomes or commonly used genetic mutants.

Three examples of independent large duplications in *ddm1*

We have identified two independent large duplications (55 and 56 kb) in two *ddm1-2* individuals that occurred in the lab and was thus not due to EMS mutagenesis. A large duplication has already been observed in the *ddm1* mutant background at the *bal* locus (Yi and Richards, 2008, 2009), suggesting that duplications are frequent in this background. The *bal* duplication, also 55 kb in length, is associated with a dwarf plant phenotype due to overexpression of a duplicated gene at the *RPP5 locus*, a disease resistance locus containing tandem repeats of R genes clustered with TEs (Yi and Richards, 2009). Large tandem duplications have also been reported in lines derived from a *fas2* mutant with a decrease number of rDNA copies (Picart-Picolo et al. 2020). Loss of *DDM1* or *FAS2* correlates with higher rates of meiotic (Melamed-Bessudo and Levy, 2012) or homologous recombination (Endo et al., 2006; Muchova et al., 2015), respectively. Our study suggests that DDM1 is involved in genome stability. The detection of duplications was probably facilitated by the use of ONT sequencing together with the use of single plants for sequencing, increasing our SV detection sensitivity.

DDM1, a DHL chromatin remodeler involved in DNA repair?

DDM1 has recently been defined as belonging to DHL chromatin remodelers, an acronym for a family comprising DDM1 (in plants), HELLS (Human helicase lymphoid specific, in humans) and LSH (lymphoid specific helicase, in mice) (Berger et al., 2022). Both DDM1 and HELLS are involved in the deposition of heterochromatic histone variants, namely H2A.W and macroH2A1.2, respectively (Berger et al., 2022; Osakabe et al., 2021). Recent studies suggest an increasing role of histone variants in DNA repair (Davarinejad et al., 2022). HELLS plays a role in homologous recombination repair of heterochromatic breaks (Ni et al., 2020; Xu et al., 2021; Caron et al., 2021). HELLS and LSH are expressed in lymphoid where they participate in V(D)J recombination but are also expressed in testis. During meiosis, HELLS interacts with PRDM9 to open chromatin and direct recombination DSBs (Spruce et al., 2020; Imai et al., 2020). LSH promotes DNA repair (Burrage et al., 2012) in mice. Indirectly, DHL remodelers have an impact on DNA methylation, and their loss is associated with TE reactivation. We propose that the two phenotypes observed in this study highlight these dual roles of DDM1. In one hand, DDM1 is indirectly involved in DNA methylation: its loss releases silencing at TEs, leading to a high eccDNA load and new TE insertions. When combined with *rdr6* and *pol4* this mutation leads to truncated and complex SVs. On the other hand, DDM1 is involved in H2A.W deposition and ensures HR repair of DSBs: its loss leads to tandem duplications. More examples of epigenetically induced SVs will be instrumental to address the precise role of DDM1 in genome stability.

Methods

Detection of eccDNAs from Illumina and ONT eccDNA-seq

The eccDNA producing loci from each Arabidopsis epigenetic mutant were detected using ecc_finder (Zhang et al., 2021) with default parameters of short-read-mapping and long-read-mapping mode (for Illumina and data ONT data, respectively). eccDNAs originating from organelle DNA fragments mapping to their nucleic copy (such as NUPTs for nucleoplasmic DNA and NUMTs for nuclear mitochondrial DNA (Saccone et al., 2000; Yoshida et al., 2014)), from ribosomal DNA repeats (rDNA) and from centromeric repeats were removed. The remaining eccDNAs were grouped into TE

families by mapping them with BWA to the reference genome and normalized with FPKM (fragments per kilo base per million mapped reads, paired-end).

EVD functional annotation

The two LTR sequences were identified by self-to-self alignment using BLAST. The sh-GAG structure of *EVD* was identified previously, indicating splicing of AP (Oberlin et al., 2017), and long read RNA-seq (Panda and Slotkin, 2020) further confirmed the sequences of GAG and AP. RH sequence was identified using VLP data from the same *ddm1* mutant (Lee et al., 2020).

Detection of truncated and chimeric eccDNAs from ONT eccDNA-seq

EccDNA-seq data were filtered for reads containing *EVD*. These *EVD* reads were mapped on the annotated structural domains of the *EVD* sequence using minimap2. Different profiles, such as the loss of different structural domains, were visualized using dotplot and then systematically grouped using bedtools groupby. In the next step, *EVD* reads were remapped to the reference genome using minimap2 to check for unmapped sequences. Chimeric eccDNAs supported by at least 5 reads were retained. The same approach was used for *ATCOPIA21* eccDNAs.

Structural variant detection from *de novo* assembly of ONT reads

ONT genomic reads were assembled using Canu v2.0 (Koren et al., 2017) with with the following modification of default parameters: -nanopore genomeSize=130m corMhapSensitivity=high corMinCoverage=0 corOutCoverage=100. The producing genome assemblies were polished with Pilon (Walker et al., 2014) and then matched against the TAIR10 reference genome using Mummer4 (Marçais et al., 2018). The raw match results were further filtered using delta-filter and then subjected SV detection using Syri (Goel et al., 2019).

Detection of TE insertion polymorphisms from ONT genomic reads

Reads spanning the entire insertion and deletion sequence do not cause alignment breakpoints, but are flagged in the CIGAR. We developed a script to filter the CIGAR

output. This pipeline can be found on GitHub (https://github.com/njaupan/CIGAR_SV). Briefly, in order to extract all reads containing insertions and deletions, we generated a PAF file from minimap2 (-cs -cx map-ont) (Li, 2018, 2) . The CIGAR metrics were then indexed until the position of the SV on the reference genome. Breakpoints with more than 5 supported reads at the same position were selected. The start and end positions of the breakpoints were extracted from the PAF file and grouped to generate common breakpoints displayed in BED format. The breakpoint locations were finally filtered for the presence of 5-20bp TSDs supporting *bona fide* TE insertion polymorphisms.

Identification of inversion and duplication

SVs from minimap2 alignments were detected using Sniffles v2.0 (Sedlazeck et al., 2018), which generated a VCF file for each Arabidopsis mutant separately. The VCF file were furthered filtered for duplications, inversions and translocations larger than 1kb. In order to detect inversions, reads at the two junctions were extracted and visualized with dotplot. The final Chr2-2M inversion was visually validated using Jbrowse (Buels et al., 2016) and re-constructed manually. For duplications, read depth was calculated with samtools depth (Li et al., 2009) and visualized with samplot (Belyeu et al., 2021). Read spanning the junctions of two tandem duplications were extracted to identify the different copies.

Detection of DNA methylation from ONT genomic reads

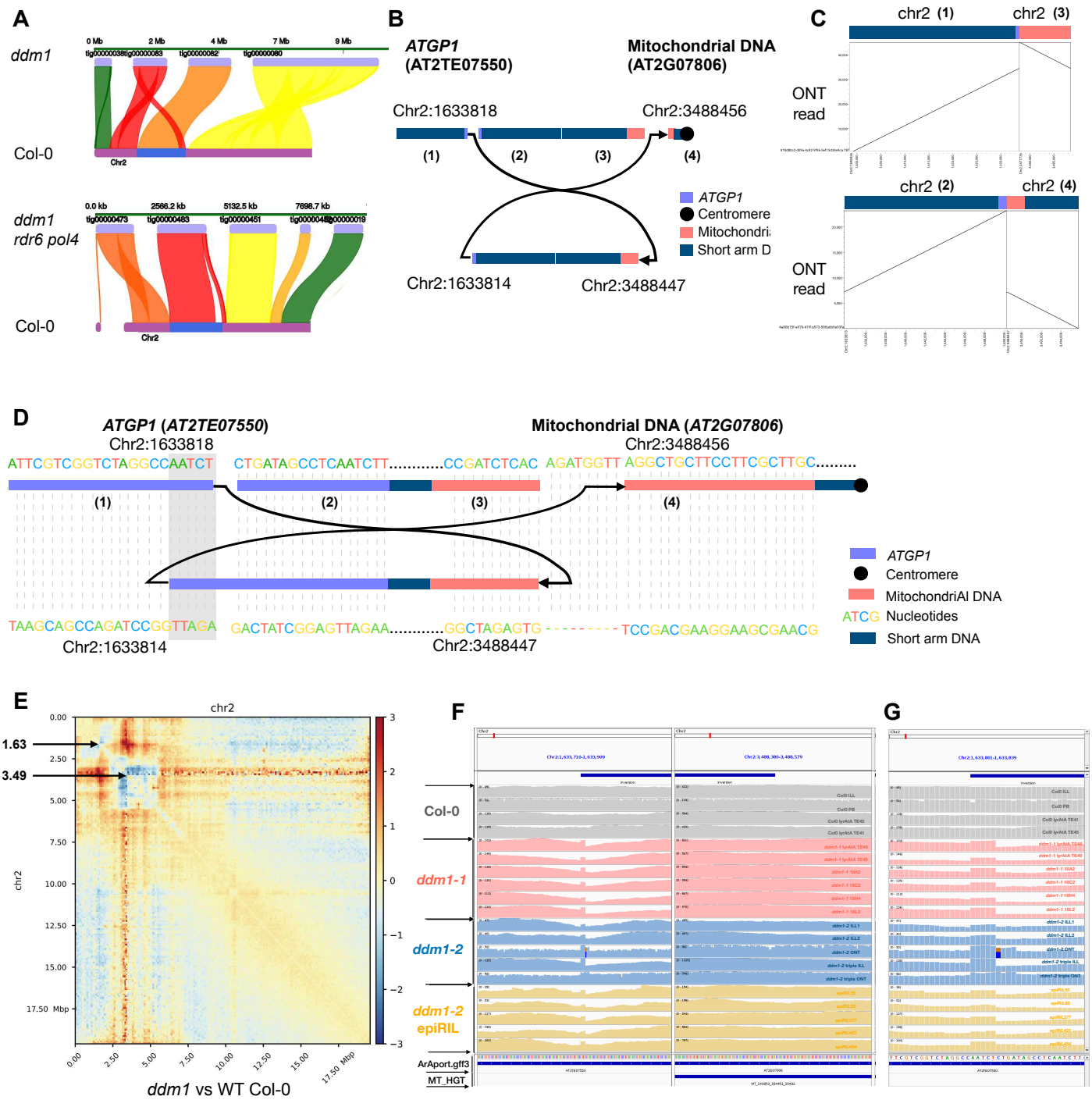
Cytosine methylation patterns were detected from ONT reads using Nanopolish (Simpson et al., 2017). The methylation patterns at the two ends of the Chr2-2M inversion and the two duplications were parsed and plotted using methylartist (https://github.com/adamewing/methylartist).

Data and code availability

All high-throughput sequencing data generated in this study have been the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena) under the PRJEBXXX project. Source codes are available at https://github.com/njaupan/eccDNA-Genome.

Acknowledgments

We thank our colleagues from the IRD and LGDP laboratories, our EpiDiverse colleagues for stimulating discussions. MM is supported by a grant from the French National Agency for Research (ANR-13-JSV6-0002 “*ExtraChrom*”). This study is set within the framework of the "Laboratoire d’Excellence (LABEX)" TULIP (ANR-10-LABX-41) and of the "Ecole Universitaire de Recherche (EUR)" TULIP-GS (ANR-18-EURE-0019). PZ and MM are members of the European Training Network “*EpiDiverse*” that received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965.



Supplementary Figure 1. A large inversion in the *A. thaliana ddm1*-2 mutant genome. (**A**) Comparison of *ddm1-2* genome assembly and reference genome on chromosome 2, and *ddm1 rdr6 pol4* genome assembly and reference genome on chromosome 2. (**B**) Diagram of the Chr2-2M inversion which breaks at the retrotransposon *ATGP1* and a hypothetical gene (AT2G07806) corresponding to mitochondrial DNA integrated into the nuclear genome (Saccone et al., 2000). (**C**) Alignment of raw reads spanning the two breakpoints of the Chr2-2M inversion in dot plots. (**D**) Schematic representation of Chr2-2M inversion at the nucleotide level. (**E**) Inversion breakpoints correspond to "cold" (blue) spots revealed by Hi-C data (Feng et al., 2014). (**F**) **&** (**G**) Read mapping of whole-genome sequencing of *ddm1-1* and *ddm1-2* generated by Vongs et al. in 1993, and *ddm1-2* EpiRiL mutants by Vongs et al. in 2014. Different sequencing technologies (Illumina: ILL, PacBio: PB and Nanopore: ONT) are indicated. The left side alignment of *ATGP1 (AT2TE07550)* resulted in a 5bp duplication (**G**).

References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* 182, 145-161.e23. doi:10.1016/j.cell.2020.05.021.

Arrey, G., Keating, S. T., and Regenberg, B. (2022). A unifying model for extrachromosomal circular DNA load in eukaryotic cells. *Seminars in Cell & Developmental Biology*. doi:10.1016/j.semcdb.2022.03.002.

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* 176, 663-675.e19. doi:10.1016/j.cell.2018.12.019.

Belyeu, J. R., Chowdhury, M., Brown, J., Pedersen, B. S., Cormier, M. J., Quinlan, A. R., et al. (2021). Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol* 22, 161. doi:10.1186/s13059-021-02380-5.

Berger, F., Muegge, K., and Richards, E. J. (2022). Seminars in cell and development biology on histone variants remodelers of H2A variants associated with heterochromatin. *Seminars in Cell & Developmental Biology*. doi:10.1016/j.semcdb.2022.02.026.

Bourguet, P., Picard, C. L., Yelagandula, R., Pélissier, T., Lorković, Z. J., Feng, S., et al. (2021). The histone variant H2A.W and linker histone H1 co-regulate heterochromatin accessibility and DNA methylation. *Nat Commun* 12, 2683. doi:10.1038/s41467-021-22993-5.

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology* 17, 66. doi:10.1186/s13059-016-0924-1.

Burrage, J., Termanis, A., Geissner, A., Myant, K., Gordon, K., and Stancheva, I. (2012). The SNF2 family ATPase LSH promotes phosphorylation of H2AX and efficient repair of DNA double-strand breaks in mammalian cells. *Journal of Cell Science* 125, 5524–5534. doi:10.1242/jcs.111252.

Caron, P., Pobega, E., and Polo, S. E. (2021). DNA Double-Strand Break Repair: All Roads Lead to HeterochROMAtin Marks. *Frontiers in Genetics* 12. Available at: https://www.frontiersin.org/article/10.3389/fgene.2021.730696 [Accessed April 13, 2022].

Carvalho, C. M. B., and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 17, 224–238. doi:10.1038/nrg.2015.25.

Cohen, S., and Méchali, M. (2002). Formation of extrachromosomal circles from telomeric DNA in Xenopus laevis. *EMBO Rep* 3, 1168–1174. doi:10.1093/embo-reports/kvf240.

Cortijo, S., Wardenaar, R., Colomé-Tatché, M., Gilly, A., Etcheverry, M., Labadie, K., et al. (2014). Mapping the epigenetic basis of complex Traits. *Science* 343, 1145–1148. doi:10.1126/science.1248127.

Davarinejad, H., Huang, Y.-C., Mermaz, B., LeBlanc, C., Poulet, A., Thomson, G., et al. (2022). The histone H3.1 variant regulates TONSOKU-mediated DNA repair during replication. *Science* 375, 1281–1286. doi:10.1126/science.abm5320.

Durkin, K., Coppieters, W., Drögemüller, C., Ahariz, N., Cambisano, N., Druet, T., et al. (2012). Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* 482, 81–84. doi:10.1038/nature10757.

Endo, M., Ishikawa, Y., Osakabe, K., Nakayama, S., Kaya, H., Araki, T., et al. (2006). Increased frequency of homologous recombination and T-DNA integration in Arabidopsis CAF-1 mutants. *EMBO J* 25, 5579–5590. doi:10.1038/sj.emboj.7601434.

Feng, S., Cokus, S. J., Schubert, V., Zhai, J., Pellegrini, M., and Jacobsen, S. E. (2014). Genome-wide Hi-C analyses in wild type and mutants reveal high-resolution chromatin interactions in Arabidopsis. *Mol Cell* 55, 694–707. doi:10.1016/j.molcel.2014.07.008.

Fu, Y., Kawabe, A., Etcheverry, M., Ito, T., Toyoda, A., Fujiyama, A., et al. (2013). Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. *EMBO J* 32, 2407–2417. doi:10.1038/emboj.2013.169.

Garfinkel, D. J., Stefanisko, K. M., Nyswaner, K. M., Moore, S. P., Oh, J., and Hughes, S. H. (2006). Retrotransposon Suicide: Formation of Ty1 Circles and Autointegration via a Central DNA Flap. *Journal of Virology* 80, 11920–11934. doi:10.1128/JVI.01483-06.

Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* 20, 277. doi:10.1186/s13059-019-1911-0.

He, J., Babarinde, I. A., Sun, L., Xu, S., Chen, R., Shi, J., et al. (2021a). Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. *Nat Commun* 12, 1456. doi:10.1038/s41467-021-21808-x.

He, L., Zhao, C., Zhang, Q., Zinta, G., Wang, D., Lozano-Durán, R., et al. (2021b). Pathway conversion enables a double-lock mechanism to maintain DNA methylation and genome stability. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2107320118. doi:10.1073/pnas.2107320118.

Hirochika, H., and Otsuki, H. (1995). Extrachromosomal circular forms of the tobacco retrotransposon Ttol. *Gene* 165, 229–232. doi:10.1016/0378-1119(95)00581-P.

Hotta, Y., and Bassel, A. (1965). Molecular size and circularity of DNA in cells of mammals and high plants. *Proc Natl Acad Sci U S A* 53, 356–362.

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 373, 655–662. doi:10.1126/science.abg5289.

Imai, Y., Biot, M., Clément, J. A., Teragaki, M., Urbach, S., Robert, T., et al. (2020). PRDM9 activity depends on HELLS and promotes local 5-hydroxymethylcytosine enrichment. *eLife* 9, e57117. doi:10.7554/eLife.57117.

Jupe, F., Rivkin, A. C., Michael, T. P., Zander, M., Motley, S. T., Sandoval, J. P., et al. (2019). The complex architecture and epigenomic impact of plant T-DNA insertions. *PLoS Genet* 15. doi:10.1371/journal.pgen.1007819.

Kim, S., Park, J., Yeom, S.-I., Kim, Y.-M., Seo, E., Kim, K.-T., et al. (2017). 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.

Koche, R. P., Rodriguez-Fos, E., Helmsauer, K., Burkert, M., MacArthur, I. C., Maag, J., et al. (2020). Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nat Genet* 52, 29–34. doi:10.1038/s41588-019-0547-z.

Koo, D.-H., Molin, W. T., Saski, C. A., Jiang, J., Putta, K., Jugulam, M., et al. (2018). Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed Amaranthus palmeri. *Proc Natl Acad Sci U S A* 115, 3332–3337. doi:10.1073/pnas.1719354115.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). 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.

Krasileva, K. V. (2019). The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. *Current Opinion in Plant Biology* 48, 18–25. doi:10.1016/j.pbi.2019.01.004.

Kumar, P., Dillon, L. W., Shibata, Y., Jazaeri, A., Jones, D. R., and Dutta, A. (2017). Normal and cancerous tissues release extrachromosomal circular DNA (eccDNA) into the circulation. *Mol Cancer Res* 15, 1197–1205. doi:10.1158/1541-7786.MCR-17-0095.

Lanciano, S., Carpentier, M.-C., Llauro, C., Jobet, E., Robakowska-Hyzorek, D., Lasserre, E., et al. (2017). Sequencing the extrachromosomal circular mobilome

reveals retrotransposon activity in plants. *PLoS Genet* 13, e1006630. doi:10.1371/journal.pgen.1006630.

Lanciano, S., Zhang, P., Llauro, C., and Mirouze, M. (2021). Identification of Extrachromosomal Circular Forms of Active Transposable Elements Using Mobilome-Seq. *Methods Mol Biol* 2250, 87–93. doi:10.1007/978-1-0716-1134-0_7.

Lee, S. C., Ernst, E., Berube, B., Borges, F., Parent, J.-S., Ledon, P., et al. (2020). Arabidopsis retrotransposon virus-like particles and their regulation by epigenetically activated small RNA. *Genome Res.* doi:10.1101/gr.259044.119.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Lloyd, J. P. B., and Lister, R. (2022). Epigenome plasticity in plants. *Nat Rev Genet* 23, 55–68. doi:10.1038/s41576-021-00407-y.

Lopez, F. B., Fort, A., Tadini, L., Probst, A. V., McHale, M., Friel, J., et al. (2021). Gene dosage compensation of rRNA transcript levels in Arabidopsis thaliana lines with reduced ribosomal gene copy number. *Plant Cell* 33, 1135–1150. doi:10.1093/plcell/koab020.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* 14, e1005944. doi:10.1371/journal.pcbi.1005944.

Melamed-Bessudo, C., and Levy, A. A. (2012). Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis. *Proc Natl Acad Sci U S A* 109, E981-988. doi:10.1073/pnas.1120742109.

Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., et al. (2009). Selective epigenetic control of retrotransposition in Arabidopsis. *Nature* 461, 427–430. doi:10.1038/nature08328.

Møller, H. D., Larsen, C. E., Parsons, L., Hansen, A. J., Regenberg, B., and Mourier, T. (2016). Formation of Extrachromosomal Circular DNA from Long Terminal Repeats of Retrotransposons in Saccharomyces cerevisiae. *G3: Genes, Genomes, Genetics* 6, 453–462. doi:10.1534/g3.115.025858.

Møller, H. D., Parsons, L., Jørgensen, T. S., Botstein, D., and Regenberg, B. (2015). Extrachromosomal circular DNA is common in yeast. *Proc Natl Acad Sci USA* 112, E3114–E3122. doi:10.1073/pnas.1508825112.

Muchová, V., Amiard, S., Mozgová, I., Dvořáčková, M., Gallego, M. E., White, C., et al. (2015). Homology-dependent repair is involved in 45S rDNA loss in plant CAF-1 mutants. *Plant J* 81, 198–209. doi:10.1111/tpj.12718.

Ni, K., Ren, J., Xu, X., He, Y., Finney, R., Braun, S. M. G., et al. (2020). LSH mediates gene repression through macroH2A deposition. *Nat Commun* 11, 5647. doi:10.1038/s41467-020-19159-0.

Nicolau, M., Picault, N., and Moissiard, G. (2021). The Evolutionary Volte-Face of Transposable Elements: From Harmful Jumping Genes to Major Drivers of Genetic Innovation. *Cells* 10, 2952. doi:10.3390/cells10112952.

Oberlin, S., Sarazin, A., Chevalier, C., Voinnet, O., and Marí-Ordóñez, A. (2017). A genome-wide transcriptome and translatome analysis of *Arabidopsis* transposons identifies a unique and conserved genome expression strategy for *Ty1/Copia* retroelements. *Genome Res.* 27, 1549–1562. doi:10.1101/gr.220723.117.

Osakabe, A., Jamge, B., Axelsson, E., Montgomery, S. A., Akimcheva, S., Kuehn, A. L., et al. (2021). The chromatin remodeler DDM1 prevents transposon mobility through deposition of histone variant H2A.W. *Nat Cell Biol* 23, 391–400. doi:10.1038/s41556-021-00658-1.

Panda, K., Ji, L., Neumann, D. A., Daron, J., Schmitz, R. J., and Slotkin, R. K. (2016). Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome Biol* 17, 170. doi:10.1186/s13059-016-1032-y.

Panda, K., and Slotkin, R. K. (2020). Long-Read cDNA Sequencing Enables a “Gene-Like” Transcript Annotation of Transposable Elements. *Plant Cell* 32, 2687–2698. doi:10.1105/tpc.20.00115.

Picart-Picolo, A., Grob, S., Picault, N., Franek, M., Llauro, C., Halter, T., et al. (2020). Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response. *Genome Res.* 30, 1583–1592. doi:10.1101/gr.261586.120.

Pucker, B., Kleinbölting, N., and Weisshaar, B. (2021). Large scale genomic rearrangements in selected Arabidopsis thaliana T-DNA lines are caused by T-DNA insertion mutagenesis. *BMC Genomics* 22, 599. doi:10.1186/s12864-021-07877-8.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184, 3542-3558.e16. doi:10.1016/j.cell.2021.04.046.

Quadrana, L., Etcheverry, M., Gilly, A., Caillieux, E., Madoui, M.-A., Guy, J., et al. (2019). Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nat Commun* 10, 3421. doi:10.1038/s41467-019-11385-5.

Saccone, C., Gissi, C., Lanave, C., Larizza, A., Pesole, G., and Reyes, A. (2000). Evolution of the mitochondrial genetic system: an overview. *Gene* 261, 153–159. doi:10.1016/s0378-1119(00)00484-4.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., et al. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 15, 461–468. doi:10.1038/s41592-018-0001-7.

Shibata, Y., Kumar, P., Layer, R., Willcox, S., Gagan, J. R., Griffith, J. D., et al. (2012). Extrachromosomal MicroDNAs and Chromosomal Microdeletions in Normal Tissues. *Science* 336, 82–86. doi:10.1126/science.1213307.

Sigman, M. J., and Slotkin, R. K. (2016). The First Rule of Plant Transposable Element Silencing: Location, Location, Location. *Plant Cell* 28, 304–313. doi:10.1105/tpc.15.00869.

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 14, 407–410. doi:10.1038/nmeth.4184.

Sinclair, D. A., and Guarente, L. (1997). Extrachromosomal rDNA circles--a cause of aging in yeast. *Cell* 91, 1033–1042. doi:10.1016/s0092-8674(00)80493-6.

Spruce, C., Dlamini, S., Ananda, G., Bronkema, N., Tian, H., Paigen, K., et al. (2020). HELLS and PRDM9 form a pioneer complex to open chromatin at meiotic recombination hot spots. *Genes Dev.* 34, 398–412. doi:10.1101/gad.333542.119.

Thierry, A., Khanna, V., Créno, S., Lafontaine, I., Ma, L., Bouchier, C., et al. (2015). Macrotene chromosomes provide insights to a new mechanism of high-order gene amplification in eukaryotes. *Nat Commun* 6, 6154. doi:10.1038/ncomms7154.

Tsukahara, S., Kawabe, A., Kobayashi, A., Ito, T., Aizu, T., Shin-i, T., et al. (2012). Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. *Genes Dev.* 26, 705–713. doi:10.1101/gad.183871.111.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. *Nature* 461, 423–426. doi:10.1038/nature08351.

Van de Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., et al. (2019). A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell* 178, 1260-1272.e14. doi:10.1016/j.cell.2019.07.038.

Verhaak, R. G. W., Bafna, V., and Mischel, P. S. (2019). Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat Rev Cancer* 19, 283–288. doi:10.1038/s41568-019-0128-6.

Vongs, A., Kakutani, T., Martienssen, R. A., and Richards, E. J. (1993). Arabidopsis thaliana DNA Methylation Mutants. *Science*. doi:10.1126/science.8316832.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9, e112963. doi:10.1371/journal.pone.0112963.

Wang, Y., Wang, M., Djekidel, M. N., Chen, H., Liu, D., Alt, F. W., et al. (2021). eccDNAs are apoptotic products with high innate immunostimulatory activity. *Nature* 599, 308–314. doi:10.1038/s41586-021-04009-w.

Wu, S., Turner, K. M., Nguyen, N., Raviram, R., Erb, M., Santini, J., et al. (2019). Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* 575, 699–703. doi:10.1038/s41586-019-1763-5.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., and van der Knaap, E. (2008). A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit. *Science* 319, 1527–1530. doi:10.1126/science.1153040.

Xu, X., Ni, K., He, Y., Ren, J., Sun, C., Liu, Y., et al. (2021). The epigenetic regulator LSH maintains fork protection and genomic stability via MacroH2A deposition and RAD51 filament formation. *Nat Commun* 12, 3520. doi:10.1038/s41467-021-23809-2.

Yi, H., and Richards, E. J. (2007). A Cluster of Disease Resistance Genes in Arabidopsis Is Coordinately Regulated by Transcriptional Activation and RNA Silencing. *Plant Cell* 19, 2929–2939. doi:10.1105/tpc.107.051821.

Yi, H., and Richards, E. J. (2008). Phenotypic instability of Arabidopsis alleles affecting a disease Resistance gene cluster. *BMC Plant Biol* 8, 36. doi:10.1186/1471-2229-8-36.

Yi, H., and Richards, E. J. (2009). Gene Duplication and Hypermutation of the Pathogen Resistance Gene SNC1 in the Arabidopsis bal Variant. *Genetics* 183, 1227–1234. doi:10.1534/genetics.109.105569.

Yoshida, T., Furihata, H. Y., and Kawabe, A. (2014). Patterns of Genomic Integration of Nuclear Chloroplast DNA Fragments in Plant Species. *DNA Res* 21, 127–140. doi:10.1093/dnares/dst045.

Zhang, P., Peng, H., Llauro, C., Bucher, E., and Mirouze, M. (2021). ecc_finder: a robust and accurate tool for detecting extrachromosomal circular DNA from sequencing data. *Frontiers in Plant Science* 12. Available at: doi:10.3389/fpls.2021.743742.

3. General discussion and perspectives

3.1 Future trends to understand the role of eccDNAs

3.1.1 Remaining questions on eccDNA inheritance and the emergence of new genes

The inheritance of eccDNA is still a mystery. Due to the absence of centromeres, eccDNA segregation may be heterogeneous, resulting in progeny cells containing different eccDNA copy numbers (Verhaak et al., 2019). Cells with a high copy number of eccDNAs may have a selective advantage for adaptation to certain environments, as shown for the glyphosate resistance in *Amaranthus palmeri* (Koo et al., 2018b).

In order to better understand eccDNA inheritance and its role in chimeric gene formation, it would be interesting to find a model system in which eccDNA might be abundant and passed to the next generation. In *C. elegans* spermatocytes, a brief temperature shift significantly increase the *Tc1/mariner* transposition activity (Kurhanewicz et al., 2020), increasing DNA double-strand breaks in spermatocytes by 25-fold. Despite this, heat shocked males produced offspring with surprisingly low non-viability, with only a 3-fold increase, demonstrating the biological stability of spermatogenesis (Bhalla, 2020; Kurhanewicz et al., 2020). Transposition of transposons during spermatogenesis also validates a hypothesis about the origin of new genes--out of testis (Nyberg and Carthew, 2017). Because testis is subject to extremely strong selective pressures, such as competition between spermatocytes and gender conflict, testicular tissue is evolutionarily extremely voracious and able to accommodate a variety of new genes. Many studies have shown that genes related to sperm formation are often subject to strong natural selection pressures (Kaessmann, 2010; Wu et al., 2011; Zhao et al., 2014; Oss and Carvunis, 2019). In mammals and flies, newly formed genes tend to be specifically expressed only in testis (Wu et al., 2011; Zhao et al., 2014). Moreover, there is substantial DNA demethylation during formation, providing excellent conditions for transposon activation. Although the presence of eccDNA was not yet described in this model, it would be a nice experimental system to investigate the impact of transposition on SVs and possibly on chimeric gene formation, transmitted the next generation.

Young gene emergence was also studied in Arabidopsis and rice genomes and this revealed that a number of recently evolved young genes are involved in defense and reproductive processes. Transcriptomic analysis further showed that plant male reproductive cells are associated with high expression of young genes (Cui et al., 2015).

Therefore, it is exciting to consider future experiments to address eccDNA inheritance patterns during sexual reproduction. In the case of hypomethylated *ddm1* mutants, for example, it would be interesting to detect fluorescently labeled *EVADE* and get a live-cell image during cell division and reproductive development. Of note, DDM1 plays an important role in pollen development. It is expressed in the sperm cells but lost in the vegetative nucleus (Slotkin et al., 2009; Calarco et al., 2012).

3.1.2 Towards eccDNA detection directly from genomic data

Plant population genomes assembled from long read sequences are emerging, for instance in *Arabidopsis thaliana* (Jiao and Schneeberger, 2020), tomato (Gao et al., 2019; Alonge et al., 2020) and rice (Qin et al., 2021). In particular, a large number of SVs associated with phenotypes in these species, including duplication, TE insertions, and translocations have been resolved. DNA double-strand breaks, chromosomal rearrangements, and other possible chromosomal events can lead to DNA fragment cyclization to form eccDNA (Liao et al., 2020). Nevertheless, it is still questionable whether hotspots of SVs will be the hotspots that generate eccDNAs.

To address this question, it would be important to characterize eccDNA producing loci directly from genomic datasets (not only from eccDNA-seq). However, most software currently available for eccDNA detection are based on the input from eccDNA-seq data, such as ecc_finder (Zhang et al., 2021b). Is it possible to use the whole genome sequencing (WGS) data in hand to predict the loci that produce eccDNA? The answer is clearly yes.

If to consider the short read mapping, read pairs can be grouped by chromosome and orientation. Circular specific reads (discordant read pairs with outward-facing labels and split read pairs with inward-facing labels) used for identification in eccDNA-seq data will remain the target of WGS data. Could we speed up the computational performance (compared to Python scripts), e.g. by using SAMBLASTER (Faust and Hall, 2014)? So far a few tools such as SAMBLASTER can detect discordant and split read in linear DNA. Similarly, loci displaying more than two sub-reads alignment in the long-read mapping will be candidate eccDNA producing loci. However, more tests to improve the

computational performance are needed to develop this algorithm. The sensitivity of this approach would also have to be evaluated.

3.2 Future trends to obtain high-quality genome assembly

3.2.1 Choice of long read sequencing technologies

PacBio and ONT are currently the mainstream long read sequencing technologies, with PacBio HiFi reads having an advantage in high accuracy and ONT reads having an advantage in read length. How to select sequencing technologies and obtain high quality genome assemblies is critical. Recently, the differences between PacBio and ONT in genome assembly of spruce, rice, and maize species have been compared, and we take maize as an example for a detailed description.

Using PacBio sequencing (62X coverage), ONT sequencing (50X coverage), Liu et al. (2020) compared the assembly results of PacBio alone, ONT alone, and PacBio +ONT+physical mapping, respectively, and the results showed that the contiguity of PacBio assembly was 20 times higher than that of ONT, but ONT assembly showed better results in assembling large repeats and high heterozygous regions. When combining PacBio with ONT+physical map assembly, the overall maize genome Contig N50 reached 162 Mb, and the Gap number was only 1.3 Mb, which is the best maize genome assembly contiguity so far (Liu et al., 2020a).

Furthermore, Mascher et al. (2021) compared all barley assemblies from (1) PacBio continuous long-reads (CLR), (2) PacBio circular consensus sequencing reads (CCS), (3) ONT, and (4) lllumina short-read data (TRITEX). The CLR data were assembled separately using MECAT and wtdbg2 softwares; CLR and TRITEX data were mixed using Wengan software; CCS data were assembled using Hi-Canu (Nurk et al., 2020) and Falcon (https://github.com/PacificBiosciences/FALCON), respectively; ONT data were assembled using Smartdenovo (https://github.com/ruanjue/smartdenovo). Five evaluation criteria were used after genome assembly: (1) evaluation of basic statistics such as contig or scaffold N50/N90; (2) match with Bionano optical profiles; (3) comparison rate with barley reference genome (Morex V2); (4) evaluation of conserved datasets (BUSCO); (5) barley transcriptome data matching rate. The results showed that long-read sequencing is significantly better than short read length sequencing; the choice of the assembly algorithm has a great influence on the assembly results; mixing

short read with long read length data does not work well; complex sequences of genome need longer read length to be resolved (Mascher et al., 2021).

To date, plant genome studies have shown that the analysis of the very large genomes and polyploid plant genomes requires a smart material selection. The combination of genome + transcriptome + resequencing is still a necessary routine. Finally, HiFi + ONT ultra long reads provides technical support for the completion of plant genomes. With even complex genome assemblies in our hands, the most exciting part of genomics resides now in the characterization of “core" and "dispensable" genome for a given species.

3.2.2 Emerging tools to characterize SVs in pan-genomes

Using a pan-genome approach rather than a single reference genome allows for a more comprehensive characterization of genetic variation and can improve genomic analysis used widely. However, there are very few pan-genomic related tools and they are still in the developmental stage. Giraffe is a new tool that efficiently maps genomic sequences to a "pangenome" representing a wide range of different human genomic sequences. From 5,202 different individuals, Giraffe identified 167,000 SVs using short read mapping and further identified haplotypes based on the sequence graph, demonstrating its broad applicability and functionality (Sirén et al., 2021).

Benchmark of all long-read SV callers reveals the strengths and weaknesses of each SV detection algorithm and provided the basis for integrating multiple algorithms in a new SV detection pipeline, namely combiSV (Dierckxsens et al.,2021). The Perl script combines VCF output from Sniffles, pbsv, and so on into a superior call set. CombiSV achieves higher recall, precision and accuracy than SURVIVOR, an existing algorithm for generating consensus VCFs (Dierckxsens et al., 2021).

A just-published tool, TEsorter, can accurately classify LTR retrotransposons in the maize and rice genomes and can be drawn upon for integration into a pan-genomic SVs tool (Zhang et al., 2022). Based on the classified conserved protein structural domains (database source REXdb or GyDB), TEsorter first searches for structural domains using Hidden Markov Models and then filters hits for classification. The classification results

at the clade level are highly consistent with the phylogenetic tree. By comparing five commonly used transposon classification software (RepeatModeler, DeepTE, TERR, LTR_retriever and LTRclassifier), TEsorter has a clear advantage in terms of accuracy and computational speed (Zhang et al., 2022).

In the future, more effective SV detection algorithms will consist in a combination of multiple methods in order to produce better results. These above-mentioned methods may still be surpassed by emerging tools to facilitate breakpoint resolution. There is still a lot of room for development in this area.

3.2.3 Having a pan-genome alternative

Although not everyone can have a pan-genome due to sequencing price, is it possible to obtain a whole class of genes or gene families, namely pan-genes? The key question to address is how extensive is the diversity for a gene family of interest, in the target population and how much genome assembly is required to capture most of the variation.

In this respect, the construction of pan-genes for the class of disease resistance genes, the intracellular nucleotide-binding site leucine repeat receptors (NLRs) in *A. thaliana* provides a learning blueprint (Van de Weyer et al., 2019). Based on combining resistance gene enrichment sequencing with PacBio SMRT sequencing technology from 64 geographically distributed *A. thaliana* accessions, a nearly complete pan-NLRome was identified. The 40 randomized lines have outlined more than 98% of the pan-NLRome. It demonstrates the combined sequencing strategy is a cost-effective alternative to whole-genome sequencing. It can identify the complex diversity of NLRs, and also provides a viable method to identify pan-NLRome to accelerate the elucidation of NLR specificity in disease resistance (Van de Weyer et al., 2019; Barragan and Weigel, 2021).

3.3 Future trends to uncover the relationship between TEs, eccDNAs and SVs

3.3.1 On the role of VLP in fast TE and TE-gene chimerism evolution

Given the capture capacity of TEs, well described for Pack-MULEs in rice and Pack-TIRs in 100 species of animals, TEs have been described to promote adaptive evolution by forming new genes (Talbert and Chandler, 1988; Jiang et al. 2004, 2011; Tan et al. 2021). Furthermore, fusions between DNA transposons and protein-coding genes in all tetrapod genomes demonstrate that TEs provide a recurrent supply for shaping novel protein structures (Cosby et al., 2021). However, little is still known about fusions between TE, notably retrotransposons, and gene in terms of biogenesis, stability and transgenerational impact.

Considering that during retrotransposon life-cycle, transcripts are encapsidated in the VLP, how do TE and gene transcripts fuse together? Do they peel off from the genome and form chimeric ecDNA? Or are they wrapped together in the VLP and then transferred to the nucleus as chimeric eccDNA? To answer these questions, recent VLP-related studies studies on domesticated gag proteins open new perspectives.

Arc, for instance, is a gag domesticated protein present in human neuronal cells and derived from *Gypsy* retrotransposons. *Arc* proteins form a VLP in which their own mRNA is encapsulated and transferred from one neuronal cell to another, participating in memory consolidation (Pastuzyn et al., 2018). Similarly, other animals have independently evolved their own *Arc*. The *Arc* gene in Drosophila also transports RNA between neurons in a VLP (Ashley et al., 2018). Similarly, the *Gypsy* retrotransposons-derived protein PEG10 is also capable of transferring or binding RNA and has also been reported to be involved in the formation of the mammalian placenta (Ono et al., 2006; Korb and Finkbeiner, 2011). Since genes derived from gag homolog can form VLP and then serve as RNA delivery, can they deliver also other gene transcripts of interest? The paper by Segel et al. (2021) is noteworthy as it is the first example of a specific biotechnological RNA delivery within a cell. The authors show that target mRNAs can be reprogrammed with the untranslated region flanking Peg10 allowing their encapsidation by PEG10 for RNA delivery. Excitingly, 500 bp of the 3' UTR of the mouse PEG10 were sufficient to efficiently transfer exogenous mRNA into target reporter cells (Segel et al., 2021).

The power of genes derived from gag is clearly established, and thus it will not be too surprising that endogenous genes can gain extra ability from integrase, protease, and so on. For example, the *Gin2* gene (*Gypsy Integrase 2*) was domesticated from a retrotransposon integrase in fish, at least 500 million years ago. The Gin2 protein retains the HHCC zinc finger motif suggesting its ability to bind DNA or RNA (Marín, 2010; Chalopin et al., 2012). Although the mechanisms for TE domestication are not known, the VLP offers a possibility for quick evolution of TEs. Indeed, partial retrotransposons were found in the VLP, but also in the eccDNA fraction and in the genome of *A. thaliana ddm1* mutants, suggesting a possible route for rapid TE evolution. Re-analyzing the available VLP data of Arabidopsis *ddm1* mutants in the genic regions (Lee et al., 2020) could give a hint on the genes and/or TE-gene chimeras entering the VLPs.

3.3.2 Interactions between eccDNA and the genome in the context of the 3D genome

The three-dimensional (3D) structure of chromatin allows for interactions between DNA elements. In tumour cells, the chromatin of eccDNA is highly opened: eccDNA contains histone modifications of enhancers and promoters (H3K4me1/3, H3K27ac), but lacks repressive histone modifications (Wu et al., 2019). Chromatin loop formation mediates the interaction between enhancers and promoters, which drives gene expression. In tumor cells eccDNA enhancers can come in close contact to genes leading to their over-expression. New ultra-long distance chromatin interactions can thus occur within eccDNAs (Wu et al., 2019).

Chen et al. (2021) developed a new technique for studying ecDNA chromatin openness at the single molecule level. Genomic DNA is processed using m6A MTase methyltransferase to obtain m6A DNA methylation modifications in open regions of chromatin. Exonuclease is also introduced to remove linear genomic DNA and the complete eccDNA is sequenced by ONT sequencing. The finding that eccDNA chromatin accessibility is mostly highly open compared to that of linear DNA reinforces the general view that eccDNA amplification leads to higher transcription of oncogenes (Chen et al., 2021).

Therefore, it would be interesting to check chromatin accessibility at loci generating chimeric eccDNA in *ddm1* mutants (Zhong et al., 2021). In addition, the expression of

chimeric genes located on these eccDNA and integrated in the genome should be measured and compared to the expression of the endogenous gene

Finally, one could use Hi-C data for the direct detection of SVs. Some methods were already developed in this direction, such as Hic_breakfinder that can potentially identify all types of SVs, while others, such as HiCnv (Chakraborty and Ay, 2018) aims to detect only copy number variants (CNVs) and translocations, respectively (Spielmann et al., 2018).

In the thesis, I investigated TE mobility, SVs in *de novo* assembly and read mapping, as well as development of tools to explore eccDNA landscape, SVs in genomes and mechanisms of eccDNA-genome interaction. Inspired by TE-gene chimera, I will further explore new gene formation in large genomes, or communications between different cells mediated by VLP cargo, as a postdoc in Cédric Feschotte's lab.

4. Bibliography

A Howard, E., and S Dennis, E. (1984). Transposable elements in maize - the Activator-Dissociation (Ac-Ds) System. *Aust. Jnl. Of Bio. Sci.* 37, 307. doi:10.1071/BI9840307.

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nat Methods* 8, 61–65. doi:10.1038/nmeth.1527.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182, 145-161.e23. doi:10.1016/j.cell.2020.05.021.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., et al. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166, 481–491. doi:10.1016/j.cell.2016.05.063.

Arkhipova, I. R., and Yushenova, I. A. (2019). Giant transposons in eukaryotes: is bigger better? *Genome Biol Evol* 11, 906–918. doi:10.1093/gbe/evz041.

Ashley, J., Cordy, B., Lucia, D., Fradkin, L. G., Budnik, V., and Thomson, T. (2018). Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* 172, 262-274.e11. doi:10.1016/j.cell.2017.12.022.

Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., et al. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* 33, 296–300. doi:10.1038/nbt.3103.

Barragan, A. C., and Weigel, D. (2021). Plant NLR diversity: the known unknowns of pan-NLRomes. *The Plant Cell* 33, 814–831. doi:10.1093/plcell/koaa002.

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nat. Plants* 6, 914–920. doi:10.1038/s41477-020-0733-0.

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., et al. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants* 4, 879–887. doi:10.1038/s41477-018-0289-4.

Belyeu, J. R., Chowdhury, M., Brown, J., Pedersen, B. S., Cormier, M. J., Quinlan, A. R., et al. (2021). Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol* 22, 161. doi:10.1186/s13059-021-02380-5.

Bhalla, N. (2020). Meiosis: Is Spermatogenesis Stress an Opportunity for Evolutionary Innovation? *Current Biology* 30, R1471–R1473. doi:10.1016/j.cub.2020.10.042.

Bolger, M., Schwacke, R., Gundlach, H., Schmutzer, T., Chen, J., Arend, D., et al. (2017). From plant genomes to phenotypes. *Journal of Biotechnology* 261, 46–52. doi:10.1016/j.jbiotec.2017.06.003.

Bourguet, P., Picard, C. L., Yelagandula, R., Pélissier, T., Lorković, Z. J., Feng, S., et al. (2021). The histone variant H2A.W and linker histone H1 co-regulate heterochromatin accessibility and DNA methylation. *Nat Commun* 12, 2683. doi:10.1038/s41467-021-22993-5.

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology* 17, 66. doi:10.1186/s13059-016-0924-1.

Cabanettes, F., and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 6, e4958. doi:10.7717/peerj.4958.

Calarco, J. P., Borges, F., Donoghue, M. T. A., Van Ex, F., Jullien, P. E., Lopes, T., et al. (2012). Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell* 151, 194–205. doi:10.1016/j.cell.2012.09.001.

Cameron, D. L., Di Stefano, L., and Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* 10, 3240. doi:10.1038/s41467-019-11146-4.

Capy, P., Vitalis, R., Langin, T., Higuet, D., and Bazin, C. (1996). Relationships between transposable elements based upon the integrase-transposase domains: Is there a common ancestor? *J Mol Evol* 42, 359–368. doi:10.1007/BF02337546.

Carpentier, M.-C., Manfroi, E., Wei, F.-J., Wu, H.-P., Lasserre, E., Llauro, C., et al. (2019). Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nat Commun* 10, 24. doi:10.1038/s41467-018-07974-5.

Carvalho, C. M. B., and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 17, 224–238. doi:10.1038/nrg.2015.25.

Castel, S. E., and Martienssen, R. A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat Rev Genet* 14, 100–112. doi:10.1038/nrg3355.

Cerbin, S., and Jiang, N. (2018). Duplication of host genes by transposable elements. *Current Opinion in Genetics & Development* 49, 63–69. doi:10.1016/j.gde.2018.03.005.

Chakraborty, A., and Ay, F. (2018). Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics* 34, 338–345. doi:10.1093/bioinformatics/btx664.

Chalopin, D., Galiana, D., and Volff, J.-N. (2012). Genetic Innovation in Vertebrates: Gypsy Integrase Genes and Other Genes Derived from Transposable Elements. *Int J Evol Biol* 2012, 724519. doi:10.1155/2012/724519.

Chen, W., Weng, Z., Xie, Z., Xie, Y., Zhang, C., Chen, Z., et al. (2021). Sequencing of methylase-accessible regions in integral circular extrachromosomal DNA reveals differences in chromatin structure. *Epigenetics & Chromatin* 14, 40. doi:10.1186/s13072-021-00416-5.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 18, 170–175. doi:10.1038/s41592-020-01056-5.

Cohen, S., and Méchali, M. (2002). Formation of extrachromosomal circles from telomeric DNA in Xenopus laevis. *EMBO Rep* 3, 1168–1174. doi:10.1093/embo-reports/kvf240.

Cohen, Z., Bacharach, E., and Lavi, S. (2006). Mouse major satellite DNA is prone to eccDNA formation via DNA Ligase IV-dependent pathway. *Oncogene* 25, 4515–4524. doi:10.1038/sj.onc.1209485.

Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 29, 987–991. doi:10.1038/nbt.2023.

Cortijo, S., Wardenaar, R., Colome-Tatche, M., Gilly, A., Etcheverry, M., Labadie, K., et al. (2014). Mapping the Epigenetic Basis of Complex Traits. *Science* 343, 1145–1148. doi:10.1126/science.1248127.

Cosby, R. L., Judd, J., Zhang, R., Zhong, A., Garry, N., Pritham, E. J., et al. (2021). Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* 371, eabc6405. doi:10.1126/science.abc6405.

Cost, G. J., Feng, Q., Jacquier, A., and Boeke, J. D. (2002). Human L1 element target-primed reverse transcription in vitro. *EMBO J* 21, 5899–5910. doi:10.1093/emboj/cdf592.

Cui, X., Lv, Y., Chen, M., Nikoloski, Z., Twell, D., and Zhang, D. (2015). Young Genes out of the Male: An Insight from Evolutionary Age Analysis of the Pollen Transcriptome. *Molecular Plant* 8, 935–945. doi:10.1016/j.molp.2014.12.008.

Deragon, J.-M., and Zhang, X. (2006). Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Systematic Biology* 55, 949–956. doi:10.1080/10635150601047843.

Dierckxsens, N., Li, T., Vermeesch, J. R., and Xie, Z. (2021). A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biol* 22, 342. doi:10.1186/s13059-021-02551-4.

Dong, F., Miller, J. T., Jackson, S. A., Wang, G.-L., Ronald, P. C., and Jiang, J. (1998). Rice (Oryza sativa) centromeric regions consist of complex DNA. *PNAS* 95, 8135–8140. doi:10.1073/pnas.95.14.8135.

Drost, H.-G., and Sanchez, D. H. (2019). Becoming a Selfish Clan: Recombination Associated to Reverse-Transcription in LTR Retrotransposons. *Genome Biology and Evolution* 11, 3382–3392. doi:10.1093/gbe/evz255.

Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., et al. (2019). Origin and evolution of the octoploid strawberry genome. *Nat Genet* 51, 541–547. doi:10.1038/s41588-019-0356-4.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. 323, 7.

Faust, G. G., and Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 30, 2503–2505. doi:10.1093/bioinformatics/btu314.

Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* 3, 329–341. doi:10.1038/nrg793.

Feschotte, C., and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41, 331–368. doi:10.1146/annurev.genet.40.110405.090448.

Fultz, D., Choudury, S. G., and Slotkin, R. K. (2015). Silencing of active transposable elements in plants. *Current Opinion in Plant Biology* 27, 67–76. doi:10.1016/j.pbi.2015.05.027.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). 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.

Garg, S. (2021). Computational methods for chromosome-scale haplotype reconstruction. *Genome Biology* 22, 101. doi:10.1186/s13059-021-02328-9.

Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* 20, 277. doi:10.1186/s13059-019-1911-0.

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

Havecker, E. R., Gao, X., and Voytas, D. F. (2004). The diversity of LTR retrotransposons. *Genome Biology* 5, 225. doi:10.1186/gb-2004-5-6-225.

Hirochika, H., and Otsuki, H. (1995). Extrachromosomal circular forms of the tobacco retrotransposon Ttol. *Gene* 165, 229–232. doi:10.1016/0378-1119(95)00581-P.

Hotta, Y., and Bassel, A. (1965). Molecular size and circularity of DNA in cells of mammals and high plants. *Proc Natl Acad Sci U S A* 53, 356–362.

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 373, 655–662. doi:10.1126/science.abg5289.

Hurwitz, B. L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S. A., et al. (2010). Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza. *The Plant Journal* 63, 990–1003. doi:10.1111/j.1365-313X.2010.04293.x.

Inagaki, S. (2021). Silencing and anti-silencing mechanisms that shape the epigenome in plants. *Genes Genet. Syst.*, 21–00041. doi:10.1266/ggs.21-00041.

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V. S., Gundlach, H., Monat, C., et al. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* 588, 284–289. doi:10.1038/s41586-020-2947-8.

Jiang, N., Bao, Z., Zhang, X., Eddy, S. R., and Wessler, S. R. (2004a). Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431, 569–573. doi:10.1038/nature02953.

Jiang, N., Ferguson, A. A., Slotkin, R. K., and Lisch, D. (2011). Pack-Mutator–like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *PNAS* 108, 1537–1542. doi:10.1073/pnas.1010814108.

Jiang, N., Feschotte, C., Zhang, X., and Wessler, S. R. (2004b). Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Current Opinion in Plant Biology* 7, 115–119. doi:10.1016/j.pbi.2004.01.004.

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., et al. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* 21, 189. doi:10.1186/s13059-020-02107-y.

Jiao, W.-B., and Schneeberger, K. (2020). Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun* 11, 989. doi:10.1038/s41467-020-14779-y.

Johannes, F., Porcher, E., Teixeira, F. K., Saliba-Colombani, V., Simon, M., Agier, N., et al. (2009). Assessing the impact of transgenerational epigenetic variation on complex traits. *PLOS Genetics* 5, e1000530. doi:10.1371/journal.pgen.1000530.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 1313–1326. doi:10.1101/gr.101386.109.

Kapitonov, V. V., and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A* 98, 8714–8719. doi:10.1073/pnas.151269298.

Kapitonov, V. V., and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics* 23, 521–529. doi:10.1016/j.tig.2007.08.004.

Kasschau, K. D., Fahlgren, N., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., et al. (2007). Genome-Wide Profiling and Analysis of Arabidopsis siRNAs. *PLOS Biology* 5, e57. doi:10.1371/journal.pbio.0050057.

Kato, M., Miura, A., Bender, J., Jacobsen, S. E., and Kakutani, T. (2003). Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis. *Curr Biol* 13, 421–426. doi:10.1016/s0960-9822(03)00106-4.

Kazazian, H. H. (2004). Mobile Elements: Drivers of Genome Evolution. *Science* 303, 1626–1632. doi:10.1126/science.1089670.

Kilzer, J. M., Stracker, T., Beitzel, B., Meek, K., Weitzman, M., and Bushman, F. D. (2003). Roles of host cell factors in circularization of retroviral dna. *Virology* 314, 460–467. doi:10.1016/S0042-6822(03)00455-0.

Kim, E. Y., Wang, L., Lei, Z., Li, H., Fan, W., and Cho, J. (2021). Ribosome stalling and SGS3 phase separation prime the epigenetic silencing of transposons. *Nat. Plants* 7, 303–309. doi:10.1038/s41477-021-00867-4.

Kim, H., Nguyen, N.-P., Turner, K., Wu, S., Gujar, A. D., Luebeck, J., et al. (2020). Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet* 52, 891–897. doi:10.1038/s41588-020-0678-2.

Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A., and Voytas, D. F. (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res.* 8, 464–478. doi:10.1101/gr.8.5.464.

Kinoshita, Y., Saze, H., Kinoshita, T., Miura, A., Soppe, W. J. J., Koornneef, M., et al. (2006). Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats: FWA gene silencing in A. thaliana. *The Plant Journal* 49, 38–45. doi:10.1111/j.1365-313X.2006.02936.x.

Koche, R. P., Rodriguez-Fos, E., Helmsauer, K., Burkert, M., MacArthur, I. C., Maag, J., et al. (2020). Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nat Genet* 52, 29–34. doi:10.1038/s41588-019-0547-z.

Kofler, R., Gómez-Sánchez, D., and Schlötterer, C. (2016). PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Molecular Biology and Evolution* 33, 2759–2764. doi:10.1093/molbev/msw137.

Koo, D.-H., Molin, W. T., Saski, C. A., Jiang, J., Putta, K., Jugulam, M., et al. (2018). Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed Amaranthus palmeri. *PNAS* 115, 3332–3337.

Korb, E., and Finkbeiner, S. (2011). Arc in synaptic plasticity: from gene to behavior. *Trends Neurosci* 34, 591–598. doi:10.1016/j.tins.2011.08.007.

Kumar, P., Dillon, L. W., Shibata, Y., Jazaeri, A., Jones, D. R., and Dutta, A. (2017). Normal and cancerous tissues release extrachromosomal circular DNA (eccDNA) into the circulation. *Mol Cancer Res* 15, 1197–1205. doi:10.1158/1541-7786.MCR-17-0095.

Kurhanewicz, N. A., Dinwiddie, D., Bush, Z. D., and Libuda, D. E. (2020). Elevated Temperatures Cause Transposon-Associated DNA Damage in C. elegans Spermatocytes. *Curr Biol* 30, 5007-5017.e4. doi:10.1016/j.cub.2020.09.050.

Lanciano, S., Carpentier, M.-C., Llauro, C., Jobet, E., Robakowska-Hyzorek, D., Lasserre, E., et al. (2017). Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLOS Genetics* 13, e1006630. doi:10.1371/journal.pgen.1006630.

Lanciano, S., Zhang, P., Llauro, C., and Mirouze, M. (2021). Identification of Extrachromosomal Circular Forms of Active Transposable Elements Using Mobilome-Seq. *Methods Mol Biol* 2250, 87–93. doi:10.1007/978-1-0716-1134-0_7.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062.

Law, J. A., and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11, 204–220. doi:10.1038/nrg2719.

Lee, S. C., Ernst, E., Berube, B., Borges, F., Parent, J.-S., Ledon, P., et al. (2020). Arabidopsis retrotransposon virus-like particles and their regulation by epigenetically activated small RNA. *Genome Res.* doi:10.1101/gr.259044.119.

Lerat, E., and Capy, P. (1999). Retrotransposons and retroviruses: analysis of the envelope gene. *Molecular Biology and Evolution* 16, 1198–1207. doi:10.1093/oxfordjournals.molbev.a026210.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, L., Olvera, J. M., Yoder, K. E., Mitchell, R. S., Butler, S. L., Lieber, M., et al. (2001). Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. *EMBO J* 20, 3272–3281. doi:10.1093/emboj/20.12.3272.

Liao, Z., Jiang, W., Ye, L., Li, T., Yu, X., and Liu, L. (2020). Classification of extrachromosomal circular DNA with a focus on the role of extrachromosomal DNA (ecDNA) in tumor heterogeneity and progression. *Biochimica et Biophysica*

Acta (BBA) - Reviews on Cancer 1874, 188392. doi:10.1016/j.bbcan.2020.188392.

Linheiro, R. S., and Bergman, C. M. (2012). Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster. *PLOS ONE* 7, e30008. doi:10.1371/journal.pone.0030008.

Lippman, Z., Gendrel, A.-V., Black, M., Vaughn, M. W., Dedhia, N., Richard McCombie, W., et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430, 471–476. doi:10.1038/nature02651.

Liu, H., Wang, X., Wang, G., Cui, P., Wu, S., Ai, C., et al. (2021). The nearly complete genome of Ginkgo biloba illuminates gymnosperm evolution. *Nat. Plants* 7, 748–756. doi:10.1038/s41477-021-00933-x.

Liu, J., Seetharam, A. S., Chougule, K., Ou, S., Swentowsky, K. W., Gent, J. I., et al. (2020a). Gapless assembly of maize chromosomes using long-read technologies. *Genome Biology* 21, 121. doi:10.1186/s13059-020-02029-9.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., et al. (2020b). Pan-genome of wild and cultivated soybeans. *Cell* 182, 162-176.e13. doi:10.1016/j.cell.2020.05.023.

Liu, Z., Tavares, R., Forsythe, E. S., André, F., Lugan, R., Jonasson, G., et al. (2016). Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. *Nat Commun* 7, 13026. doi:10.1038/ncomms13026.

Lloyd, J. P. B., and Lister, R. (2022). Epigenome plasticity in plants. *Nat Rev Genet* 23, 55–68. doi:10.1038/s41576-021-00407-y.

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., et al. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet* 45, 884–890. doi:10.1038/ng.2678.

Lyons, D. B., and Zilberman, D. (2017). DDM1 and Lsh remodelers allow methylation of DNA wrapped in nucleosomes. *eLife* 6, e30674. doi:10.7554/eLife.30674.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biol* 20, 246. doi:10.1186/s13059-019-1828-7.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* 14, e1005944. doi:10.1371/journal.pcbi.1005944.

Marín, I. (2010). GIN transposons: genetic elements linking retrotransposons and genes. *Mol Biol Evol* 27, 1903–1911. doi:10.1093/molbev/msq072.

Marx, V. (2021). Long road to long-read assembly. *Nat Methods* 18, 125–129. doi:10.1038/s41592-021-01057-y.

Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C. S., et al. (2021). Long-read sequence assembly: a technical evaluation in barley. *The Plant Cell* 33, 1888–1906. doi:10.1093/plcell/koab077.

McClintock, B. (1948). “Mutable loci in maize,” in (Cold Spring Harbor, New York: Carnegie Institution of Washington), 155–169. Available at: https://archive.org/stream/yearbookcarne47194748carn#page/154/mode/2up/search/Mcclintock [Accessed April 11, 2022].

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *PNAS* 36, 344–355. doi:10.1073/pnas.36.6.344.

McClure, M. A. (1991). Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Molecular Biology and Evolution* 8, 835–856. doi:10.1093/oxfordjournals.molbev.a040686.

Mehta, D. (2020). Full-length sequencing of circular DNA viruses and extrachromosomal circular DNA using CIDER-Seq. *NATURE PROTOCOLS*, 19.

Mérot, C., Oomen, R. A., Tigano, A., and Wellenreuther, M. (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology & Evolution* 35, 561–572. doi:10.1016/j.tree.2020.03.002.

Michael, T. P., and VanBuren, R. (2020). Building near-complete plant genomes. *Current Opinion in Plant Biology* 54, 26–33. doi:10.1016/j.pbi.2019.12.009.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 34, i142–i150. doi:10.1093/bioinformatics/bty266.

Miller, C. A., Hampton, O., Coarfa, C., and Milosavljevic, A. (2011). ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. *PLoS One* 6, e16327. doi:10.1371/journal.pone.0016327.

Møller, H. D., Parsons, L., Jørgensen, T. S., Botstein, D., and Regenberg, B. (2015). Extrachromosomal circular DNA is common in yeast. *Proc Natl Acad Sci USA* 112, E3114–E3122. doi:10.1073/pnas.1508825112.

Muñoz-López, M., and García-Pérez, J. L. (2010). DNA transposons: nature and applications in genomics. *Curr Genomics* 11, 115–128. doi:10.2174/138920210790886871.

Nagarajan, N., and Pop, M. (2013). Sequence assembly demystified. *Nat Rev Genet* 14, 157–167. doi:10.1038/nrg3367.

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Schmücker, A., et al. (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* 374, eabi7489. doi:10.1126/science.abi7489.

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., et al. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30, 1291–1305. doi:10.1101/gr.263566.120.

Nuthikattu, S., McCue, A. D., Panda, K., Fultz, D., DeFraia, C., Thomas, E. N., et al. (2013). The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs. *Plant Physiology* 162, 116–131. doi:10.1104/pp.113.216481.

Nyberg, K. G., and Carthew, R. W. (2017). Out of the testis: biological impacts of new genes. *Genes Dev.* 31, 1825–1826. doi:10.1101/gad.307496.117.

Ogiwara, I., Miya, M., Ohshima, K., and Okada, N. (1999). Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. *Molecular Biology and Evolution* 16, 1238–1250. doi:10.1093/oxfordjournals.molbev.a026214.

Ono, R., Nakamura, K., Inoue, K., Naruse, M., Usami, T., Wakisaka-Saito, N., et al. (2006). Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat Genet* 38, 101–106. doi:10.1038/ng1699.

Osakabe, A., Jamge, B., Axelsson, E., Montgomery, S. A., Akimcheva, S., Kuehn, A. L., et al. (2021). The chromatin remodeler DDM1 prevents transposon mobility through deposition of histone variant H2A.W. *Nat Cell Biol* 23, 391–400. doi:10.1038/s41556-021-00658-1.

Oss, S. B. V., and Carvunis, A.-R. (2019). De novo gene birth. *PLOS Genetics* 15, e1008160. doi:10.1371/journal.pgen.1008160.

Panda, K., Ji, L., Neumann, D. A., Daron, J., Schmitz, R. J., and Slotkin, R. K. (2016). Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome Biol* 17, 170. doi:10.1186/s13059-016-1032-y.

Pastuzyn, E. D., Day, C. E., Kearns, R. B., Kyrke-Smith, M., Taibi, A. V., McCormick, J., et al. (2018). The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* 172, 275-288.e18. doi:10.1016/j.cell.2017.12.024.

Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G. W., et al. (2015). WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J Comput Biol* 22, 498–509. doi:10.1089/cmb.2014.0157.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Research* 16, 1262–1269. doi:10.1101/gr.5290206.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 36, 983–987. doi:10.1038/nbt.4235.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184, 3542-3558.e16. doi:10.1016/j.cell.2021.04.046.

Reinders, J., Wulff, B. B. H., Mirouze, M., Marí-Ordóñez, A., Dapp, M., Rozhon, W., et al. (2009). Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. *Genes Dev* 23, 939–950. doi:10.1101/gad.524609.

Rizzi, R., Beretta, S., Patterson, M., Pirola, Y., Previtali, M., Della Vedova, G., et al. (2019). Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. *Quant Biol* 7, 278–292. doi:10.1007/s40484-019-0181-x.

Robb, S. M. C., Lu, L., Valencia, E., Burnette, J. M., III, Okumoto, Y., Wessler, S. R., et al. (2013). The Use of RelocaTE and Unassembled Short Reads to Produce High-Resolution Snapshots of Transposable Element Generated Diversity in Rice. *G3 Genes|Genomes|Genetics* 3, 949–957. doi:10.1534/g3.112.005348.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative Genomics Viewer. *Nat Biotechnol* 29, 24–26. doi:10.1038/nbt.1754.

Roudier, F., Teixeira, F. K., and Colot, V. (2009). Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. *Trends Genet* 25, 511–517. doi:10.1016/j.tig.2009.09.013.

Sabot, F., and Schulman, A. H. (2006). Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity* 97, 381–388. doi:10.1038/sj.hdy.6800903.

Saccone, C., Gissi, C., Lanave, C., Larizza, A., Pesole, G., and Reyes, A. (2000). Evolution of the mitochondrial genetic system: an overview. *Gene* 261, 153–159. doi:10.1016/s0378-1119(00)00484-4.

Sanchez, D. H., Gaubert, H., Drost, H.-G., Zabet, N. R., and Paszkowski, J. (2017). High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. *Nat Commun* 8, 1283. doi:10.1038/s41467-017-01374-x.

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

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*. doi:10.1126/science.1178534.

Schulman, A. H. (2013). Retrotransposon replication in plants. *Curr Opin Virol* 3, 604–614. doi:10.1016/j.coviro.2013.08.009.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., et al. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 15, 461–468. doi:10.1038/s41592-018-0001-7.

Segel, M., Lash, B., Song, J., Ladha, A., Liu, C. C., Jin, X., et al. (2021). Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. *Science* 373, 882–889. doi:10.1126/science.abg6155.

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). "BUSCO: Assessing Genome Assembly and Annotation Completeness," in *Gene Prediction: Methods and Protocols* Methods in Molecular Biology., ed. M. Kollmar (New York, NY: Springer), 227–245. doi:10.1007/978-1-4939-9173-0_14.

Shibata, Y., Kumar, P., Layer, R., Willcox, S., Gagan, J. R., Griffith, J. D., et al. (2012). Extrachromosomal MicroDNAs and Chromosomal Microdeletions in Normal Tissues. *Science* 336, 82–86. doi:10.1126/science.1213307.

Sin, S. T. K., Jiang, P., Deng, J., Ji, L., Cheng, S. H., Dutta, A., et al. (2020). Identification and characterization of extrachromosomal circular DNA in maternal plasma. *PNAS* 117, 1658–1665.

Sinclair, D. A., and Guarente, L. (1997). Extrachromosomal rDNA circles--a cause of aging in yeast. *Cell* 91, 1033–1042. doi:10.1016/s0092-8674(00)80493-6.

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 374, abg8871. doi:10.1126/science.abg8871.

Slotkin, R. K., Vaughn, M., Borges, F., Tanurdžić, M., Becker, J. D., Feijó, J. A., et al. (2009). Epigenetic Reprogramming and Small RNA Silencing of Transposable Elements in Pollen. *Cell* 136, 461–472. doi:10.1016/j.cell.2008.12.038.

Smolka, M., Paulin, L. F., Grochowski, C. M., Mahmoud, M., Behera, S., Gandhi, M., et al. (2022). Comprehensive Structural Variant Detection: From Mosaic to Population-Level. 2022.04.04.487055. doi:10.1101/2022.04.04.487055.

Sohn, J., and Nam, J.-W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics* 19, 23–40. doi:10.1093/bib/bbw096.

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat. Plants* 6, 34–45. doi:10.1038/s41477-019-0577-7.

Song, J.-M., Xie, W.-Z., Wang, S., Guo, Y.-X., Koo, D.-H., Kudrna, D., et al. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. *Molecular Plant* 14, 1757–1767. doi:10.1016/j.molp.2021.06.018.

Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018). Structural variation in the 3D genome. *Nat Rev Genet* 19, 453–467. doi:10.1038/s41576-018-0007-0.

Stankiewicz, P., and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med* 61, 437–455. doi:10.1146/annurev-med-100708-204735.

Sun, H., Jiao, W.-B., Campoy, J. A., Krause, K., Goel, M., Folz-Donahue, K., et al. (2021a). Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. doi:10.1101/2021.05.15.444292.

Sun, X., Zhu, S., Li, N., Cheng, Y., Zhao, J., Qiao, X., et al. (2020). A chromosome-level genome assembly of garlic (Allium sativum) provides insights into genome evolution and allicin biosynthesis. *Mol Plant* 13, 1328–1339. doi:10.1016/j.molp.2020.07.019.

Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., and Guo, L. (2021b). Twenty years of plant genome sequencing: achievements and challenges. *Trends in Plant Science*, S1360138521002818. doi:10.1016/j.tplants.2021.10.006.

Sundaresan, V., and Freeling, M. (1987). An extrachromosomal form of the Mu transposons of maize. *Proc Natl Acad Sci U S A* 84, 4924–4928. doi:10.1073/pnas.84.14.4924.

Talbert, L. E., and Chandler, V. L. (1988). Characterization of a highly conserved sequence related to mutator transposable elements in maize. *Mol Biol Evol* 5, 519–529. doi:10.1093/oxfordjournals.molbev.a040510.

Tan, S., Cardoso-Moreira, M., Shi, W., Zhang, D., Huang, J., Mao, Y., et al. (2016). LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Res* 26, 1663–1675. doi:10.1101/gr.204925.116.

Tan, S., Ma, H., Wang, J., Wang, M., Wang, M., Yin, H., et al. (2021). DNA transposons mediate duplications via transposition-independent and -dependent mechanisms in metazoans. *Nat Commun* 12, 4280. doi:10.1038/s41467-021-24585-9.

Tao, Y., Zhao, X., Mace, E., Henry, R., and Jordan, D. (2019). Exploring and Exploiting Pan-genomics for Crop Improvement. *Molecular Plant* 12, 156–169. doi:10.1016/j.molp.2018.12.016.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome.” *PNAS* 102, 13950–13955. doi:10.1073/pnas.0506758102.

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

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. *Nature* 461, 423–426. doi:10.1038/nature08351.

Turner, B. A., Miorin, T. R., Stewart, N. B., Reid, R. W., Moore, C. C., and Rogers, R. L. (2021). Chromosomal rearrangements as a source of local adaptation in island Drosophila. *arXiv:2109.09801 [q-bio]*. Available at: http://arxiv.org/abs/2109.09801 [Accessed April 11, 2022].

Turner, K. M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., et al. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* 543, 122–125. doi:10.1038/nature21356.

Van de Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., et al. (2019). A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell* 178, 1260-1272.e14. doi:10.1016/j.cell.2019.07.038.

Verhaak, R. G. W., Bafna, V., and Mischel, P. S. (2019). Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat Rev Cancer* 19, 283–288. doi:10.1038/s41568-019-0128-6.

Vongs, A., Kakutani, T., Martienssen, R. A., and Richards, E. J. (1993). Arabidopsis thaliana DNA Methylation Mutants. *Science*. doi:10.1126/science.8316832.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., et al. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature* 588, 277–283. doi:10.1038/s41586-020-2961-x.

Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., et al. (2021a). High-quality Arabidopsis thaliana genome assembly with nanopore and HiFi long reads. *Genomics, Proteomics & Bioinformatics*. doi:10.1016/j.gpb.2021.08.003.

Wang, K., Tian, H., Wang, L., Wang, L., Tan, Y., Zhang, Z., et al. (2021b). Deciphering extrachromosomal circular DNA in Arabidopsis. *Computational and Structural Biotechnology Journal* 19, 1176–1183. doi:10.1016/j.csbj.2021.01.043.

Wang, Y., Wang, M., Djekidel, M. N., Chen, H., Liu, D., Alt, F. W., et al. (2021c). eccDNAs are apoptotic products with high innate immunostimulatory activity. *Nature* 599, 308–314. doi:10.1038/s41586-021-04009-w.

Wassenegger, M., Heimes, S., Riedel, L., and Sänger, H. L. (1994). RNA-directed de novo methylation of genomic sequences in plants. *Cell* 76, 567–576. doi:10.1016/0092-8674(94)90119-8.

Wells, J. N., and Feschotte, C. (2020). A field guide to eukaryotic transposable elements. *Annu. Rev. Genet.* 54, 539–561. doi:10.1146/annurev-genet-040620-022145.

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R. H., et al. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology* 19, 103. doi:10.1186/s13059-018-1479-0.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8, 973–982. doi:10.1038/nrg2165.

Wu, D.-D., Irwin, D. M., and Zhang, Y.-P. (2011). De Novo Origin of Human Protein-Coding Genes. *PLOS Genetics* 7, e1002379. doi:10.1371/journal.pgen.1002379.

Wu, S., Turner, K. M., Nguyen, N., Raviram, R., Erb, M., Santini, J., et al. (2019). Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* 575, 699–703. doi:10.1038/s41586-019-1763-5.

Xiong, X., Gou, J., Liao, Q., Li, Y., Zhou, Q., Bi, G., et al. (2021). The Taxus genome provides insights into paclitaxel biosynthesis. *Nat. Plants* 7, 1026–1036. doi:10.1038/s41477-021-00963-5.

Yan, Y., Guo, G., Huang, J., Gao, M., Zhu, Q., Zeng, S., et al. (2020). Current understanding of extrachromosomal circular DNA in cancer pathogenesis and therapeutic resistance. *Journal of Hematology & Oncology* 13, 124. doi:10.1186/s13045-020-00960-9.

Yoshida, T., Furihata, H. Y., and Kawabe, A. (2014). Patterns of Genomic Integration of Nuclear Chloroplast DNA Fragments in Plant Species. *DNA Res* 21, 127–140. doi:10.1093/dnares/dst045.

Yu, C., Han, F., Zhang, J., Birchler, J., and Peterson, T. (2012). A transgenic system for generation of transposon Ac/Ds-induced chromosome rearrangements in rice. *Theor Appl Genet* 125, 1449–1462. doi:10.1007/s00122-012-1925-4.

Yu, T., Huang, X., Dou, S., Tang, X., Luo, S., Theurkauf, W. E., et al. (2021). A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies. *Nucleic Acids Res* 49, e44. doi:10.1093/nar/gkab010.

Yu, Y., Yao, W., Wang, Y., and Huang, F. (2019). shinyChromosome: An R/Shiny Application for Interactive Creation of Non-circular Plots of Whole Genomes.

Genomics, Proteomics & Bioinformatics 17, 535–539. doi:10.1016/j.gpb.2019.07.003.

Zhang, C., Yang, Z., Tang, D., Zhu, Y., Wang, P., Li, D., et al. (2021a). Genome design of hybrid potato. *Cell* 184, 3873-3883.e12. doi:10.1016/j.cell.2021.06.006.

Zhang, P., Peng, H., Llauro, C., Bucher, E., and Mirouze, M. (2021b). ecc_finder: a robust and accurate tool for detecting extrachromosomal circular DNA from sequencing data. *Frontiers in Plant Science* 12. Available at: doi:10.3389/fpls.2021.743742.

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., et al. (2022). TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research*, uhac017. doi:10.1093/hr/uhac017.

Zhao, L., Saelao, P., Jones, C. D., and Begun, D. J. (2014). Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science* 343, 769–772. doi:10.1126/science.1248286.

Zhong, Z., Feng, S., Duttke, S. H., Potok, M. E., Zhang, Y., Gallego-Bartolomé, J., et al. (2021). DNA methylation-linked chromatin accessibility affects genomic architecture in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2023347118. doi:10.1073/pnas.2023347118.

Zhou, H., Ma, R., Gao, L., Zhang, J., Zhang, A., Zhang, X., et al. (2021). A 1.7-Mb chromosomal inversion downstream of a PpOFP1 gene is responsible for flat fruit shape in peach. *Plant Biotechnol J* 19, 192–205. doi:10.1111/pbi.13455.

Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J. P., et al. (2020). Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat Genet* 52, 1018–1023. doi:10.1038/s41588-020-0699-x.

Zhu, Y., Gujar, A. D., Wong, C.-H., Tjong, H., Ngan, C. Y., Gong, L., et al. (2021). Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell* 39, 694-707.e7. doi:10.1016/j.ccell.2021.03.006.

5. Appendix

A short summary of my 3 contributions

In the paper by Picart-Picolo et al. (Genome research 2020), I contributed to the de novo assembly and downstream structural variation detection of two Arabidopsis *fas* mutants.

In the paper by Lanciano et al. (Plant Transposable Elements 2021), I contributed to the pipeline for the detection of eccDNA.

In the paper by Nunn et al. (Plant biotechnology journal 2021), I contributed to the TE annotation and the analysis of eccDNA detection.

Research

Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response

Ariadna Picart-Picolo,[1,2] Stefan Grob,[3] Nathalie Picault,[1,2] Michal Franek,[4] Christel Llauro,[1,2] Thierry Halter,[5] Tom R. Maier,[6] Edouard Jobet,[1,2] Julie Descombin,[1,2] Panpan Zhang,[2,7] Vijayapalani Paramasivan,[6] Thomas J. Baum,[6] Lionel Navarro,[5] Martina Dvořáčková,[4] Marie Mirouze,[2,7] and Frédéric Pontvianne[1,2]

[1]*CNRS,* [2]*UPVD, LGDP UMR5096, Université de Perpignan, 66860 Perpignan, France;* [3]*Institute of Plant and Microbial Biology, University of Zurich, CH-8008 Zurich, Switzerland;* [4]*Mendel Centre for Plant Genomics and Proteomics, CEITEC, Masaryk University, 625 00 Brno, Czech Republic;* [5]*ENS, IBENS, CNRS/INSERM, PSL Research University, 75005 Paris, France;* [6]*Department of Plant Pathology and Microbiology, Iowa State University, Ames, Iowa 50011, USA;* [7]*IRD, UMR232 DIADE, 34394 Montpellier, France*

Rapid plant genome evolution is crucial to adapt to environmental changes. Chromosomal rearrangements and gene copy number variation (CNV) are two important tools for genome evolution and sources for the creation of new genes. However, their emergence takes many generations. In this study, we show that in *Arabidopsis thaliana*, a significant loss of ribosomal RNA (rRNA) genes with a past history of a mutation for the chromatin assembly factor I (CAFI) complex causes rapid changes in the genome structure. Using long-read sequencing and microscopic approaches, we have identified up to 15 independent large tandem duplications in direct orientation (TDDOs) ranging from 60 kb to 1.44 Mb. Our data suggest that these TDDOs appeared within a few generations, leading to the duplication of hundreds of genes. By subsequently focusing on a line only containing 20% of rRNA gene copies (20rDNA line), we investigated the impact of TDDOs on 3D genome organization, gene expression, and cytosine methylation. We found that duplicated genes often accumulate more transcripts. Among them, several are involved in plant–pathogen response, which could explain why the 20rDNA line is hyper-resistant to both bacterial and nematode infections. Finally, we show that the TDDOs create gene fusions and/or truncations and discuss their potential implications for the evolution of plant genomes.

[Supplemental material is available for this article.]

In most eukaryotes, hundreds of ribosomal RNA (rRNA) genes compose the nucleolus organizer region (NOR). In *Arabidopsis thaliana Columbia* ecotype (Col-0), 375 tandem 45S rRNA gene copies are located at the top of both Chromosomes 2 (NOR2) and 4 (NOR4) (Copenhaver and Pikaard 1996). Only a portion of these copies is actively transcribed in the nucleolus to produce ribosomes. Most rRNA genes indeed remain transcriptionally inactive and accumulate repressive chromatin modification marks (Pontvianne et al. 2010, 2012, 2013; Grummt and Längst 2013). As in many species, rRNA gene copy numbers are highly variable among *A. thaliana* populations (Dopman and Hartl 2007; Kobayashi 2011; Gibbons et al. 2015; Rabanal et al. 2017). In natural inbred lines found in Sweden, rRNA copy number heterogeneity can account for up to 10% of genome size variation (Long et al. 2013). Worldwide, *A. thaliana* ecotypes can be found with a rRNA gene copy number ranging from 500 to 2500 in haploid cells (Long et al. 2013). Therefore, 500 copies could be considered as the lowest rRNA gene copy number found *in natura* so far (Rabanal et al. 2017). In budding yeast and in *Drosophila*, previous studies suggest that a minimum amount of inactive rRNA genes is necessary for global genome stability (Ide et al. 2010). One to two hundred rRNA gene units are usually found in budding yeast, but genome engineering allowed the creation of viable yeast lines with only 40 rRNA gene units (Takeuchi et al. 2003). Similarly, shifts in rRNA gene copy number affect genome-wide chromatin marks and alter gene expression in flies (Paredes and Maggert 2009; Paredes et al. 2011). In plants, neither the impact of this variability nor the consequences of having few copies of rRNA genes are known.

FASCIATA (FAS) 1 and 2 are part of the chromatin assembly factor (CAF) complex required for proper deposition of histones H3 and H4 upon DNA replication (Ramirez-Parra and Gutierrez 2007). In *A. thaliana*, their knockouts accumulate several signs of genomic instability, including double-stranded breaks (DSBs), telomere shortening, and drastic changes in rRNA gene copy number (Mozgová et al. 2010; Varas et al. 2017). Consequently, the intranuclear positioning of the NORs as well as their epigenetic state are modified in these mutants (Mozgová et al. 2010; Pontvianne et al. 2013). Crossing *fas1-4* and *fas2-4* mutants and subsequent inbreeding by self-fertilization led to the creation of *A. thaliana* wild-type segregant *FAS* genes lines with only 20% of the amount of rRNA gene copies in comparison to wild-type Col-0 (Fig. 1A; Pavlištová et al. 2016). This 20rDNA line (hereafter

Corresponding author: fpontvia@univ-perp.fr
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.261586.120. Freely available online through the *Genome Research* Open Access option.

© 2020 Picart-Picolo et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

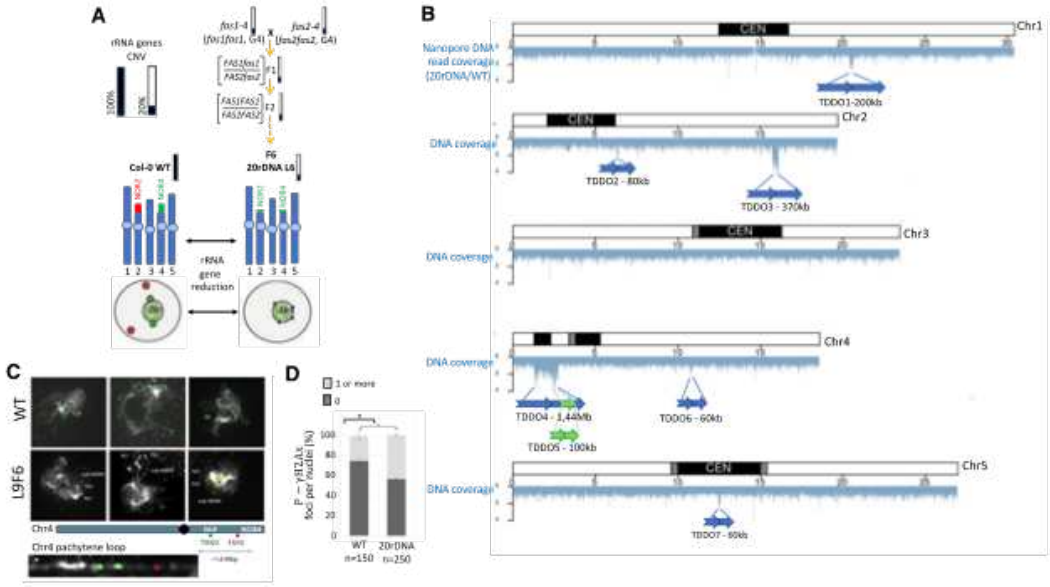


Figure 1. Genomic instability in the 20rDNA line L6F6. (*A*) Schematic representation of obtaining the 20rDNA L6F6 and its relative content in rRNA gene copies. NOR2 and NOR4 rRNA gene copies are both affected by the reduction, and according to the DNA-FISH experiment, it changes their nuclear distribution in the nucleus compared to wild-type Col-0 (WT), because all rRNA gene copies associate with the nucleolus (Pavlištová et al. 2016). (*B*) Distribution of the reads obtained by nanopore sequencing along all chromosomes in the 20rDNA L6F6 line compared to wild-type Col-0. The blue arrows represent the localization and the orientation of the TDDO identified in L6F6, except for TDDO5, which is represented by a green arrow. (*C*) DNA-FISH analyses of two loci present on the short arm of Chromosome 4 (kr4s), distant by 1.6 Mb: one present on the TDDO4 (BAC T5H22; green) and one located outside the TDDO4 (BAC F5I10; red). Three Pachytene chromosomes from WT Col-0 and in 20rDNA L6F6 are shown. A zoom of the unlooped kr4s from the *middle* panel of the 20rDNA line6 is presented at the *bottom* of the panel. (*D*) Histogram showing the percentage of nuclei displaying at least one P-γ-H2Ax foci in WT Col-0 versus L6F6.

named 20rDNA L6) has a wild-type phenotype and retained a low amount of rRNA genes for five generations (referred as F5) (Pavlištová et al. 2016). In this study, we took advantage of this plant material to test the impact of a low amount of rRNA genes on plant genome stability during several generations. We found unexpected consequences on the genome structure and stability, as well as on its 3D genome organization. We also show the short-term consequences of gene copy number variation (CNV) on their expression and potentially their role in plant phenotypic traits such as pathogen responses.

Results

The 20rDNA L6 line accumulates features of genomic instability

The 20rDNA L6 contains only 20% of rRNA genes compared to the wild-type Col-0 and was obtained as a wild-type segregant from a cross between *fas1-4* and *fas2-4* mutant lines as described (Fig. 1A; Pavlištová et al. 2016). To confirm and precisely map the potential consequence of genomic instability in 20rDNA L6F6 (F6 for the sixth generation after F1), we performed long-read resequencing using nanopore technology. We obtained 6.4 Gb of total sequences with a midsize of 6 kb. We then analyzed the sequencing coverage against the TAIR10 *A. thaliana* Col-0 reference genome to identify the highly covered regions (Fig. 1B). We have detected seven large duplications, corresponding to tandem duplications in direct orientation (TDDO), named TDDO1 to TDDO7. The largest region, TDDO4, represents 1.44 Mb, spanning the heterochromatic *knob* on the short arm of Chromosome 4 (*hk4s*), a large heterochromatic region outside the pericentromeres, and a euchromatic region distal to the *knob*. Other TDDOs range in size from 60 to 370 kb long and are present on Chromosomes 1, 2, 4, and 5 (Fig. 1B). The absence/presence of TDDO4, the largest duplication, was also confirmed by DNA-fluorescence in situ hybridization (FISH) (Fig. 1C). We used two probes generated from BAC clones: one recognizing a portion of TDDO4 (*hk4s*–T5H22) and one recognizing an unduplicated genomic region located between TDDO4 and the NOR4 (F5I10). Different cell types were analyzed from vegetative as well as reproductive tissues: in both, more signals corresponding to TDDO4 were detected in the 20rDNA L6F6 nuclei compared to wild-type Col-0 cells (Supplemental Fig. S1). Analyses of pachytene chromosomes clearly showed that the additional signal actually belonged to the same chromosome, which confirms the duplication hypothesis (Fig. 1C).

The occurrence of duplication events is a sign of genomic instability. Thus, the chromosomal rearrangements observed in 20rDNA L6F6 could be the consequence of double-stranded breaks (DSBs). To test this hypothesis, we compared the amount of spontaneous DSBs between 20rDNA L6F6 and wild-type Col-0 cells by performing immunostaining of serine 139-phosphorylated H2Ax histone variant (P-γ-H2Ax), which is a marker of DSB (Charbonnel et al. 2010). P-γ-H2Ax foci were detected at a higher rate in 20rDNA L6F6 nuclei compared to wild-type nuclei from leaf tissues (Fig. 1D; Supplemental Fig. S2A). Accumulation of DSB foci can potentially be associated with a DNA repair defect. This hypothesis is supported by an increased susceptibility of the

20rDNA L6F6 line to a treatment with the genotoxin bleomycin (Supplemental Fig. S2B).

Appearance of the duplication events in the 20rDNA line

To show a potential link between low rDNA copies and TDDO appearance, it is crucial to know when these duplication events occurred. Like 20rDNA L6F6 line, 20rDNA L9F6 is an independent inbred line deriving from the cross between *fas1-4* and *fas2-4* mutants that both also display low amounts of rDNA copies (Supplemental Fig. S3; Mozgová et al. 2010; Pavlištová et al. 2016). We then performed long-read resequencing using nanopore technology and identified TDDO in the 20rDNA L9F6 line, as well as in the offspring of the parental lines *fas1-4* and *fas2-4* used to generate the initial cross (Fig. 2A).

In *fas1-4*, none of the seven TDDOs identified in 20rDNA L6F6 were detected, but we found six new TDDOs ranging from 57 to 175 kb long (Fig. 2A; Supplemental Fig. S4A). In *fas2-4*, only TDDO4 is present, as well as two additional duplications named TDDO8 (286 kb) and TDDO9 (106 kb) (Fig. 2A; Supplemental Fig. S4B). We also identified a deletion of 9.75 kb named DEL1 on Chromosome 4 (Fig. 2A; Supplemental Fig. S5). Analyses of 20rDNA L9F6 revealed that TDDOs 1, 5, and 7 are shared between L9F6 and L6F6, which suggests their presence in the F1 (Fig. 2A; Supplemental Fig. S4C). We could not find TDDO4, although this duplication is in one of the parents. Further analyses by quantitative PCR and DNA-FISH revealed that TDDO4 has been segregated out between the L9F2 and L9F4 (Supplemental Fig. S6). In parallel, the absence of TDDOs 2, 3, and 6 in the *fas* parents or in the L9 strongly suggests the appearance of these TDDOs between L6F2 and L6F6 (Supplemental Fig. S4D). This hypothesis is supported at least for TDDO3 by qPCR analyses (Fig. 2B). However, owing to a lack of long-read sequences obtained for L9F6, we are not able to determine the existence of novel TDDO.

In summary, our analyses identified 15 TDDOs that appeared independently, either in the parental line or in the two independent inbred lines resulting from the *fas1-4* and *fas2-4* cross (Fig. 2A; Supplemental Figs. S4A,D, S7). This hypothesis is supported by the absence of TDDO4 in all generation 1 (G1) mutants *fas2-4* and *fas2-5* analyzed by PCR (Supplemental Fig. S4E,F), suggesting the also very recent appearance of TDDO4 in the parental *fas2-4* (G4) line.

Impact of low rDNA and TDDO on 3D genome organization

The nucleolus plays an important role in the spatial organization of the chromosomes (Bersaglieri and Santoro 2019; Pontvianne and Liu 2020; Santos et al. 2020). Nucleolus-associated chromatin domains (NADs), essentially composed of repressed chromatin domains, localize at the nucleolar periphery (Németh et al. 2010; van Koningsbruggen et al. 2010; Pontvianne et al. 2016b). Because rRNA gene nuclear distribution has a critical impact in NADs identity both in plant and animal cells (Quinodoz et al. 2018; Picart-Picolo et al. 2019, 2020), we analyzed NADs composition and 3D organization in 20rDNA L6F6. The fifth generation of the 20rDNA L6 (20rDNA L6F5) was transformed with a transgene ectopically expressing the FIBRILLARIN 2 nucleolar protein fused to the yellow fluorescent protein (FIB2:YFP). Using the FIB2:YFP nucleolar marker, we isolated nuclei and nucleoli from the transformants and identified NADs nuclear and nucleolar DNA sequences as previously described (Pontvianne

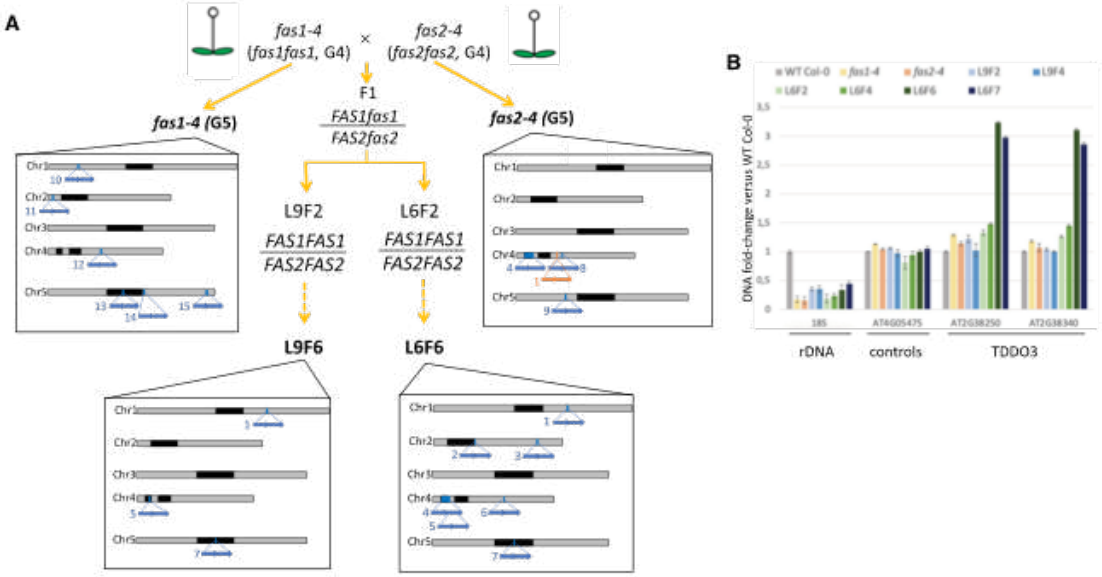


Figure 2. Identification of TDDO in the parental lines and in L9F6. (*A*) Identification of TDDO and deletion in the offspring of the parental lines *fas1-4* and *fas2-4* used to generate the L6 and L9 lines. Lines in bold were sequenced using nanopore technology. Relative distribution and names of each chromosome rearrangement identified are represented by blue (TDDO) and orange (deletion) arrows along chromosomes. Characteristics of each TDDO can be found in Supplemental Table S2. (*B*) CNV of genes present or not in TDDO3 and of rRNA genes were determined by quantitative PCR. Their relative enrichment was determined in WT Col-0, the parent lines *fas1-4* and *fas2-4*, and in L6 and L9 at several generations. CNVs of rRNA genes were determined using probes amplifying the 18S, the loci AT4G05475 and AT4G16580 that are not duplicated are controls, and the loci AT4TE12140 and AT4G05030 allow the identification of the largest duplication of TDDO4.

et al. 2016a; Carpentier et al. 2018). Preceding studies have clearly shown that in wild-type Col-0 leaf cells, NOR4-derived rRNA genes are expressed and associate with the nucleolus. Conversely, NOR2 is excluded from the nucleolus, and NOR2-derived rRNA genes are silent (Pontvianne et al. 2013; Chandrasekhara et al. 2016). As a result, NADs are essentially distributed in the entire short arm of Chromosome 4 (kr4s), which juxtaposes the active NOR4 and associates with the nucleolus (Pontvianne et al. 2016b). Compared to the wild-type, NADs in 20rDNA L6F6 are enriched from genomic regions located on both Chromosomes 2 and 4 short arms (Fig. 3A; Supplemental Fig. S8). Among the 434 genes that gained nucleolar association, 144 belong to Chromosome 2 (33%) (Supplemental Table S1). In contrast, only 19 genes on Chromosome 4 gain nucleolar association. These results are consistent with the rDNA transcriptional state, as all leftover NOR2 and NOR4-derived rRNA genes are actively transcribed and associate with the nucleolus (Pavlištová et al. 2016). We also detected an enrichment of centromeric sequences associating with nucleoli in L6F6 compared to wild type. This type of reorganization was previously shown to associate with changes in NOR subnuclear organization (Pontvianne et al. 2016b; Pontvianne and Grob 2020). As in wild type, subtelomeric regions remain associated with the nucleolus in the 20rDNA L6F6 line (Fig. 3A; Supplemental Fig. S8). In summary, NAD identification in 20rDNA L6F6 revealed that 5.6 Mb of chromatin domains mainly enriched in silent epigenetic marks changed their subnuclear distribution, which suggests a substantial reorganization of the nuclear genome.

To get a global view of the chromatin 3D organization, we analyzed all chromatin–chromatin interactions using genome-wide chromosome-conformation capture (Hi-C). We generated triplicate Hi-C samples from both wild-type and 20rDNA L6F6 14-d-old seedlings (Fig. 3B; Supplemental Fig. S9). To assess differences between two given sets of Hi-C samples statistically, we took advantage of our triplicate Hi-C data sets and performed student *t*-tests on each contact frequency (pixel of the Hi-C matrix) and determined whether contact frequencies significantly changed between the wild-type and 20rDNA L6F6 (Fig. 3C,E). Contact frequencies assayed by Hi-C can be used to detect chromosomal rearrangements (Himmelbach et al. 2018). In our case, a duplication would lead to a twofold increase in coverage of the affected region, thus doubling of interaction frequencies at this region. We indeed found several regions displaying a significant ($P<0.01$) increase of contact frequencies at several chromosomal locations, all corresponding to the previously described TDDO1 to TDDO7. Analyzing the genome-wide coverage using unpaired raw Hi-C sequencing reads confirmed the presence of significant increase in coverage of the affected regions (Supplemental Fig. S10). We subsequently normalized our Hi-C matrices for the assayed coverage. However, coverage-normalized Hi-C data showed that short-range contact frequencies within the duplicated regions are significantly depleted. Whether this depletion of contact frequencies is biologically significant or represents an artifact of the normalization procedure is extremely difficult to determine.

To further examine potential differences in 3D folding principles between wild-type Col-0 and 20rDNA L6F6, we performed a

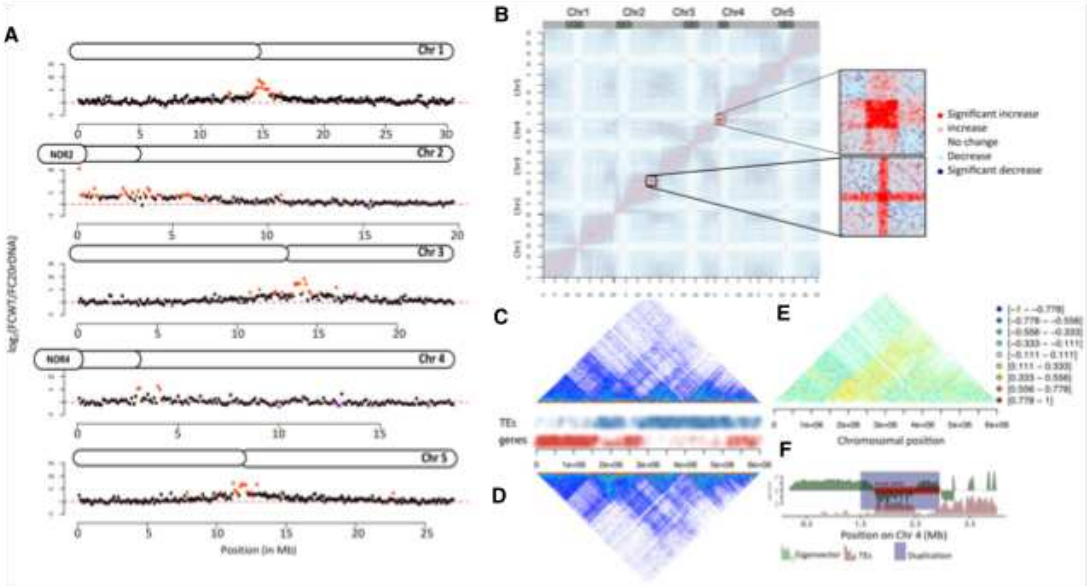


Figure 3. 3D genome organization in L6F6. (*A*) Chromosome plots displaying the relative enrichment of a given genomic segment with the nucleolus. The *y*-axis displays the fold change nucleolus enrichment between wild-type Col-0 and the 20rDNA L6F6. Each dot represents a 100-kb window. Nucleolus-enriched genomic regions above the threshold are red, and depleted regions are violet. (*B*) Coverage-normalized *t*-test difference matrix (50-kb bins). The color of each pixel of the matrix is defined by the result of a *t*-test using the triplicate contact frequencies from wild-type and 20rDNA coverage-normalized Hi-C samples. The two magnified areas correspond to the two regions displaying the highest level of contact frequency changes. (*C*,*D*) Non-normalized Hi-C snapshot showing the contact frequencies on the short arm of Chromosome 4 in wild-type Col-0 (*C*) versus the 20rDNA L6F6 (*D*). TEs and genes are annotated to illustrate the occurrence of euchromatin and heterochromatin, respectively. (*E*) Ratio between Hi-C contact frequencies from wild-type and 20rDNA L6F6. Negative ratios correspond to more contacts in the wild type, whereas positive ratios correspond to more contacts in the 20rDNA L6F6. (*F*) Eigenvector of the wild-type Co-0 Hi-C data set and annotation of the TDDO4 affecting the knob hk4s. Note the central duplication breakpoint exactly coincides with a change between LSD and a CSD.

principal component analysis (PCA) to retrieve the eigenvector, which is characteristic of 3D folding patterns of a Hi-C data set (Grob et al. 2014; Lieberman-Aiden et al. 2009). Sign changes in the eigenvector delineate basic 3D folding domains, known as loose structural domains (LSDs) and closed structural domains (CSDs), which are analogous to animal A and B compartments (Lieberman-Aiden et al. 2009). We could not identify significant changes in the eigenvectors between wild-type Col-0 and the 20rDNA L6F6. Moreover, outside the duplicated regions, no changes in genomic bin contact frequencies could be observed. We therefore focused on the duplicated regions and analyzed duplication breakage points with the eigenvector obtained by the PCA analysis of the wild-type Col-0 Hi-C data (Supplemental Fig. S11). We observed that in a majority of the TDDOs in L6F6, at least one of the breakage points coincides with sign changes (CSDs to LSDs) or directional changes (valleys and peaks within a structural domain) in the eigenvector, with the exception of TDDO3. Hence, the changes in 3D conformation may have facilitated the occurrence of the TDDOs. This was most prominent for TDDO4, where the more central breakpoint exactly colocalizes with the change between the CSD and the LSD, which defines the ancient inversion breakpoint that gave rise to the knob (Fig. 3F; Zapata et al. 2016). This suggests the existence of continuously fragile chromosomal regions, the borders between structural domains being diagnostic for these regions.

Duplication events create chimeric genes

Most of the time, TDDOs keep genes intact and do not lead to gene loss. However, truncated genes can be generated at the breakpoint junction, while keeping intact genes on the edges of duplication (Newman et al. 2015). Besides, when breakpoints are located in two different genes in the same orientation, gene fusion can take place if the open reading frame (ORF) is preserved. In 20rDNA L6F6, we systematically analyzed the TDDO breakpoint junctions (Supplemental Figs. S7, S12). Of the seven cases of TDDO identified in this line, three potentially created fused or truncated proteins (Fig. 4A,F). On Chromosome 1, TDDO1 fused the first exon of gene *AT1G55325* that encodes the N-terminal domain of the MEDIATOR 13-like with four of the five exons of *AT1G54770* that encodes the FCF2 pre-rRNA processing factor. On Chromosome 2, although genes are in the opposite orientation, TDDO3 creates a shorter ORF of the *AT2G38460* gene that potentially produces a truncated FERROPORTIN 1 protein. Finally, on Chromosome 4, TDDO4 fused the *AT4G05475* gene to a transposable element (TE) (*AT4G02960*), leading to the potential expression of three new ORFs, including one that encodes a protein with two leucin rich repeats (LRR) (Fig. 4A,F). We then systematically analyzed the presence of these chimeric genes in the genome of the parental *fas* mutant lines and in 20rDNA L6 and L9, respectively (Fig. 4G). The TDDO1-derived chimeric gene can be detected in both L6 and L9, which confirm the appearance of TDDO1 after the cross between *fas1-4* and *fas2-4* (Figs. 2, 4G). The chimeric gene generated from TDDO3 was specifically detected in L6, whereas the chimeric gene generated by TDDO4 was detected in *fas2-4*, L6, and L9 plants, confirming the results obtained earlier (Fig. 2) but also suggesting that some generations of 20rDNA L9 inbreeds plants may still segregate TDDO4.

We finally investigated whether these chimeric genes were transcribed. A first analysis of our RNA-seq data revealed that these genes were all able to accumulate transcripts. Using RT-qPCR, we confirmed the expression of the TDDO1- and TDDO3-derived chimeric genes, as well as the ability of the TDDO1-derived chimeric gene to be properly spliced (Supplemental Fig. S13). However, although reads could be detected in the RNA-seq data, we did not detect any signals for the TDDO4-derived chimeric gene by RT-

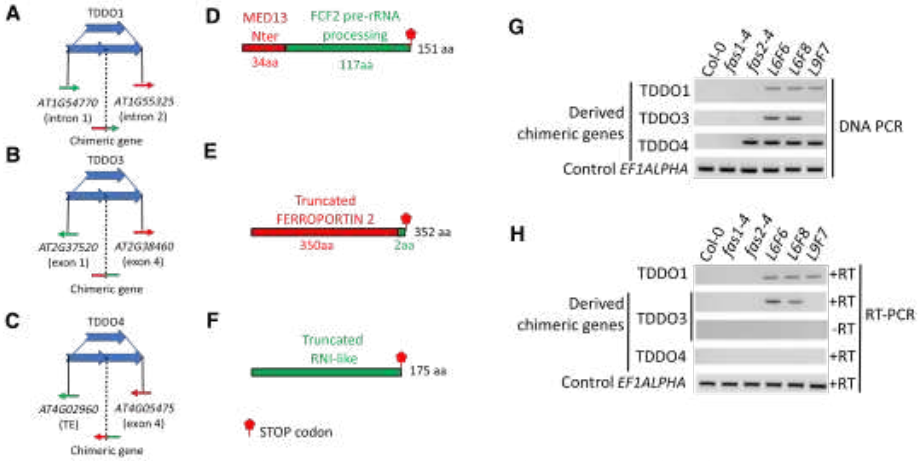


Figure 4. TDDOs provoke chimeric genes formation. (*A–C*) Schematic representation of TDDO1, TDDO3, and TDDO4 that provoked TE and/or gene fusion in the 20rDNA L6F6. Genes or TEs present in the breaking junction and their orientation are shown. (*D–F*) Open reading frames (ORFs) potentially generated at the breaking points. TDDO1 provokes the fusion of the first exon of *AT1G55325* that encodes for an ATPase motif of MEDIATOR 13 and the last four exons of *AT1G54770* that contain an RNA processing domain (*D*). The chimeric gene created between *AT2G37520* and *AT2G38460* potentially encodes for a truncated FERROPORTIN protein (*E*). The breaking points at TDDO4 fuse the 5′ sequence of a TE (*AT4G02960*) with the second and last exon of the gene *AT4G05475*, which sequence encodes two Leucine Rich Repeats (LRR) (*F*). (*G*,*H*) PCR was performed with primers flanking the breaking junctions of TDDO1, TDDO3, and TDDO4 in the wild-type Col-0, the two mutants *fas1-4* (G5) and *fas2-4* (G5), and in 20rDNA lines L6 (generations F6 and F8) and L9 (generation F7). All PCR products were confirmed by Sanger sequencing. Genomic DNA (*G*) and cDNA (*H*) were used as templates. Amplicons from the locus encoding the elongation factor *EF1ALPHA* was used as a loading control.

PCR (Fig. 4H). In conclusion, our data show that TDDOs can promote the expression of chimeric genes.

Characterization and impact of duplication events on gene expression

All TDDOs gained in 20rDNA L6F6 correspond to a gain of 2.31 Mb per haploid genome and induce CNVs of 626 genes and 851 transposable elements (TEs) (Supplemental Table S1). Changes in the 3D genome organization and CNVs can have an impact on chromatin marks and gene expression. We therefore analyzed the global gene expression pattern by poly(A)+ RNA-seq and the methylome by whole-genome bisulfite sequencing (WGBS) in wild-type Col-0 versus 20rDNA L6F6. We analyzed four replicates per samples by RNA-seq and identified differentially accumulating transcripts: 321 up-regulated genes and 14 up-regulated TEs, as well as 37 down-regulated genes but no down-regulated TEs in 20rDNA L6F6 compare to the wild-type Col-0 (with an adjusted *P*-value < 0.01 and $\log_2$[fold change] > 1.5 or <1.5) (Fig. 5A; Supplemental Table S1). We confirmed these results by quantitative RT-PCR (RT-qPCR) on nine randomly chosen genes and TEs (Supplemental Fig. S14). We did not find any correlation between differentially expressed genes and genes located in the newly arisen NADs of 20rDNA L6F6 (Supplemental Fig. S15A,C).

However, we found that duplicated genes and TEs were significantly more expressed (Supplemental Fig. S15D,E). Of the up-regulated TEs, 57% (8) are also duplicated. If we consider the 321 up-regulated genes with a $\log_2$ fold change enrichment of 1.5, we found that 22% of these genes (71) belonged to duplicated genes, but the TDDOs only represent 2% of the genome. Conversely, no genes present in TDDO are down-regulated. Higher expression can only be observed from initially expressed genes in wild-type plants. Only 286 duplicated genes are actually expressed, and 160 of them are at least twice more expressed in 20rDNA L6F6 than in wild-type Col-0 (Fig. 5B). Depending on their genomic location, TDDOs perform differently. For instance, most of the TDDO3-derived genes produced at least twice as many transcripts in 20rDNA L6F6 (71 up-regulated genes of the 80 expressed genes), whereas genes present in TDDO4, enriched in genomic regions with heterochromatic features, were less up-regulated (61 fold change > 2 genes of the 142 expressed genes) (Fig. 5B). Finally, box-plot analyses of all genes versus the duplicated genes indeed revealed their overall ability to overaccumulate more transcripts in 20rDNA L6F6 (Fig. 5C). Thus, our data strongly suggest that gene duplication often leads to an increased expression, often higher than the twofold change expected in the hypothesis of additive expression.

To analyze the impact of CNVs at the DNA methylation level, we performed triplicate WGBS in wild-type Col-0 versus 20rDNA L6F6 lines. At the genome-scale, we observed a modest increase in CG, CHG, and CHH methylation in 20rDNA L6F6 at genes (Fig. 5E,F; Supplemental Fig. S16). However, methylation at TEs was affected in both CHG and CHH contexts, but not in the CG context (Fig. 5G,H; Supplemental Fig. S17). This observation is also true if we only analyze duplicated or up-regulated genes, with the exception of gene body methylation that is unaffected for up-regulated genes (Fig. 5D; Supplemental Fig. S16). Finally, differentially methylated regions (DMRs) identified in 20rDNA L6F6 compared to wild-type Col-0 did not show a potential overlap between up-regulated genes and hypomethylated regions.

Duplication events are linked to higher pathogen resistance

In the pool of up-regulated genes in 20rDNA L6F6, genes implicated in biotic and stress responses are particularly enriched (Fig. 6A).

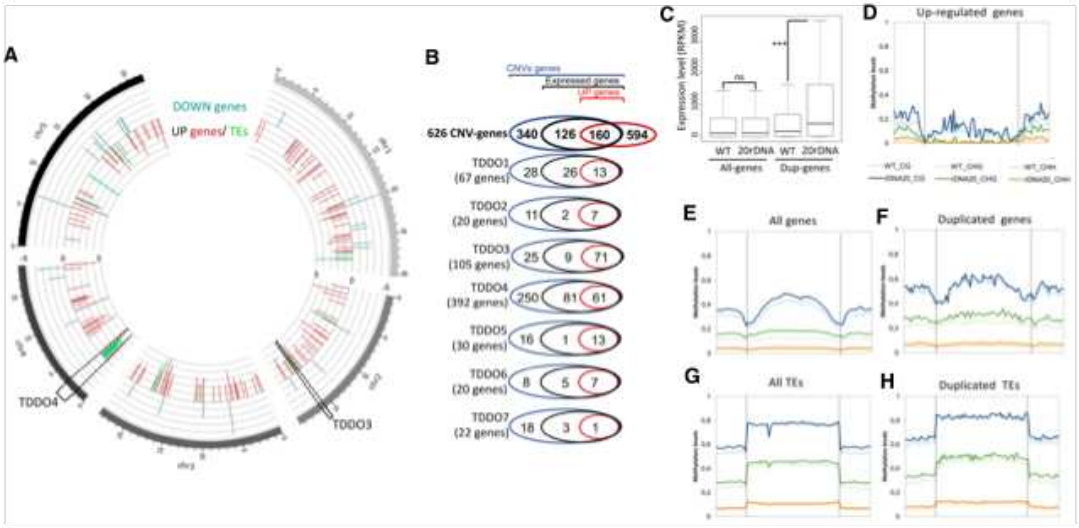


Figure 5. Impact of TDDO on global gene expression and cytosine methylation in L6F6. (*A*) Representation of the chromosomal position of genes (red bars) and TEs (green bars) of differentially accumulated transcripts in the 20rDNA L6F6 versus WT Col-0 with an adjusted *P*-value < 0.01 and a $\log_2$(fold change) > 2. Data are displayed using Circos (Krzywinski et al. 2009). The brackets display the position of TDDO3 and TDDO4. (*B*) Venn diagrams representing the proportion of expressed genes (containing at least two reads/genes in wild-type) and up-regulated genes (*P*-value < 0.01, FC > 2) among all the duplicated genes or in each TDDO in L6F6. (*C*) Dot plot revealing the relative expression of all genes or duplicated (DUP) genes in leaves of 3-wk-old plants in WT Col-0 or in 20rDNA L6F6. (***) *P*-value = 0.0005 was calculated using a Wilcoxon test. (*D–H*) Global DNA methylation analyses from genome-wide bisulfite sequencing experiments in WT Col-0 versus the 20rDNA L6F6. Global CG, CHG, and CHH methylation are shown for up-regulated genes with an adjusted *P*-value < 0.01 and a $\log_2$(fold change) > 1.5 (*D*), all genes (*E*), duplicated genes (*F*), all TEs (*G*), and duplicated TEs (*H*).

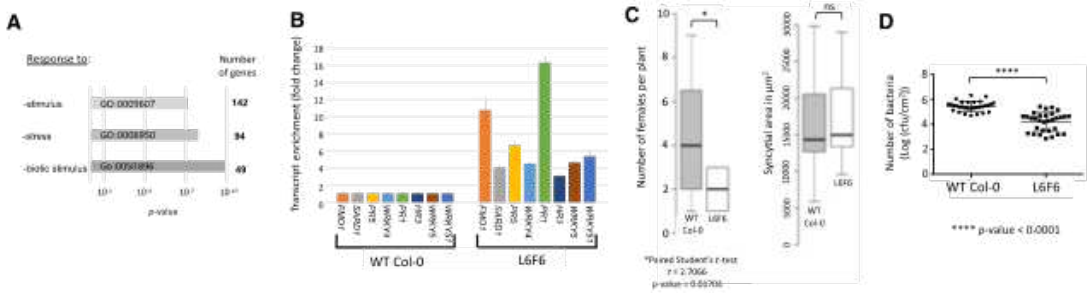


Figure 6. Biotic stress genes are overexpressed in L6F6, which is more resistant to nematode and bacterial pathogens. (*A*) GO term enriched in the pool of up-regulated genes identified in 20rDNA L6F6. (*B*) Histogram displaying the relative transcript enrichment for eight genes implicated in the biotic stress response using quantitative RT-PCR in WT Col-0 versus 20rDNA L6F6. (*C*) Wild-type (WT) Col-0 and 20rDNA L6F7 were inoculated with the sugar beet cyst nematode (*Heterodera schachtii*). Four weeks after inoculation, the number of adult females per plant was determined. Data are the average number of adult females ± SE ($n = 35 \times 3$). Data from the three independent experiments were pooled and are shown (*left*). The relative size of the syncytium cells was measured between both lines but no significant changes were noticed (*right*). (*D*) Wild-type (WT) Col-0 and 20rDNA L6F7 were inoculated with *Pseudomonas* strain DC3000 at 5×10^7 cfu/mL. Relative bacterial growth was determined 3 d after infection and is shown on the plot.

We performed RT-qPCR experiments and confirmed the overexpression of key genes involved in the plant–pathogen response (Fig. 6B). Among these genes are *PATHOGENESIS-RELATED GENE 1* (*PR1*) and *PATHOGENESIS-RELATED GENE 5* (*PR5*), whose higher expression levels are usually correlated with increased resistance against bacteria and nematodes (Wubben et al. 2008). Some of these genes were found in TDDO3 and TDDO4. Their higher expression rate could therefore be a consequence of the duplication events (Supplemental Fig. S17). Among them, *ASYMETRIC LEAVES 1* (*AS1*), present in TDDO3, has an evolutionarily conserved role in plant–pathogen interactions (Yang et al. 2008). AS1 indeed acts as a positive regulator of extracellular defenses against bacterial pathogens in a salicylic acid-independent manner (Nurmberg et al. 2007). In addition, genes encoding four cysteine-rich receptor-like kinases (CRKs), located in TDDO4, also overaccumulate transcripts in the L6F6 (Supplemental Fig. S18). Among these genes is *CRK36*, whose overexpression is sufficient to enhance pattern-triggered immunity response and bacterial pathogen resistance (Yeh et al. 2015).

A. thaliana is susceptible to various pathogens, from prokaryotes to multicellular organisms. To test their resistance capabilities, we first infected both the wild-type Col-0 and 20rDNA L6F6 with the sugar beet cyst nematode *Heterodera schachtii* (Fig. 6C). We observed that only half the number of females was able to develop on 20rDNA L6F6 plants in comparison with wild-type Col-0 plants. However, we did not observe a change in the syncytium feeding site size, that is, the plant feeding structure induced by these nematodes (Fig. 6C). Secondly, we tested the ability of 20rDNA L6F6 to be infected by the virulent bacteria *Pseudomonas syringae* strain DC3000. Three days after inoculation, bacterial growth was significantly lower in 20rDNA L6F6 (Fig. 6D) than in wild-type Col-0. Single-nucleotide polymorphisms (SNPs) could also explain changes in plant–pathogen responses, but our analyses revealed that among the 196 SNPs found in genes in L6F6 compared to wild-type Col-0, none correspond to genes implicated in biotic stress response (Supplemental Fig. S19). Considering that *fas2-4* mutant is hyper-resistant to *P. syringae* (Mozgová et al. 2015), and we identified TDDO4 in some lines of this mutant, one hypothesis is that the overexpression of pathogen response genes present in TDDO4 rather than *FAS2* gene mutation is directly implicated in the resistance against *P. syringae*. However, we cannot exclude that *fas2-4* and L6F6 pathogen resistance is mediated independently of TDDO4, which could also explain why very little overlap can be observed among the up-regulated genes in both lines (Supplemental Fig. S20).

In conclusion, we showed that higher accumulation of transcripts from genes implicated in the plant–pathogen response correlate with the plant's ability to resist against at least two types of distinct pathogens.

Discussion

Genomic structural variations shape animal and plant genomes (Krasileva 2019). Within a period of several millions of years, numerous rearrangements have occurred to shape the *Arabidopsis thaliana* genome, including duplications, translocations, inversions, and deletions (Blanc et al. 2000; Henry et al. 2006). Recently, genome analysis of seven accessions of *A. thaliana* revealed that they contain, on average, 15 Mb of rearranged sequences, generating CNVs for thousands of genes (Jiao and Schneeberger 2020). In this case, deletions, gain, or loss of copies are considered as important sources of CNVs and have potentially occurred in tens of thousands of years of evolution (Fulgione and Hancock 2018). CNVs occurring in the context of tandem duplication events represent between 3 and 4 Mb of genomic sequences in each of the seven accessions sequenced (Jiao and Schneeberger 2020). In our case, only a few generations were necessary to gain up to several megabases of genomic sequences by tandem duplications.

The rapid occurrence of these rearrangements is particularly intriguing. The relative sensitivity to genotoxic stress and the detection of a higher rate of spontaneous DSB in our 20rDNA lines is certainly one source of their appearance (Fig. 1D; Supplemental Fig. S2), but the precise mechanisms remain to be determined. One possibility is the implication of nonallelic homologous recombination (NAHR), usually responsible for TDDO (Zhang et al. 2013; Krasileva 2019). This mechanism can generate segmental duplications or deletions. In the 20rDNA L6F6, we detected duplications but no deletions, probably because of their deleterious effects.

Two other particular aspects of the detected TDDOs are their large sizes and locations, ranging from 57 kb to 1.44 Mb (Fig. 2; Supplemental Fig. S7). The TDDO borders do not share any genetic feature, and breakpoint junctions are not enriched in repetitive elements or particular genes. However, our Hi-C data revealed that

sign changes in the eigenvector seem to be overrepresented at breaking junctions, suggesting a potential link between the 3D genome folding and the occurrence of TDDOs. The systematic identification and characterization of additional TDDOs would be necessary to strengthen this hypothesis.

It is also intriguing that more than half of the 15 TDDOs are located on NOR-bearing chromosomes (Fig. 2; Supplemental Fig. S7). Because of their tandemly repeated nature, NORs are indeed subjected to an inherent instability. Therefore, the existence of a sensing system monitoring their abundance has been proposed (Nelson et al. 2019), potentially via unequal sister chromatid exchange (Tartof 1974a,b). The 20rDNA lines derive from the cross between *fas1* and *fas2* mutants, whose mutations provoked a gradual loss of rRNA genes copies (Mozgová et al. 2010). Importantly, L6 and L9 are the only siblings in which the number of rRNA genes remained stable at a low level, whereas all other lineages quickly acquired rRNA genes (Pavlištová et al. 2016). However, our data actually show that rRNA gene copies are increasing progressively throughout the inbreeding of 20rDNA L6F6 (Supplemental Fig. S6), suggesting that the CNVs are found not only at the level of the TDDOs, but also at the level of the NORs. It remains to be elucidated whether a link between the rRNA gene gains and the appearance of TDDO exists and if the same mechanisms are involved. Nevertheless, a loss of rRNA gene copies also associates with genomic instability and hypersensitivity to DNA damage in cancer cells (Wang and Lemos 2017; Xu et al. 2017). Moreover, DNA damage sensitivity and rDNA replication defects also occur in budding yeast low rDNA copy strains (Ide et al. 2010).

Short-term consequences of gene duplications have been studied in animals, especially in cancer cells, where multiple de novo tandem duplication events induce gene CNVs (Quigley et al. 2018; Wee et al. 2018). The 20rDNA L6F6 line is an unprecedented opportunity to study the transcriptional behavior of newly duplicated genes. Globally, duplicated genes tend to be more expressed (Fig. 5C). Previous observations suggest that the expression of tandem genes recently duplicated is often greater than twofold (Loehlin and Carroll 2016). Although we cannot exclude that the detected transcripts come from only one of the duplicated genes, it is more likely that equivalent additive expression occurs for the duplicated genes. During evolution, duplicated gene expression can quickly lead to specialized expression patterns, often in a tissue-specific manner, although a significant number retain correlated transcriptional profiles (Blanc and Wolfe 2004; Guschanski et al. 2017). In our case, we were able to correlate this change in gene expression with the acquisition of increased resistance to different pathogens (Fig. 6). Analyzing gene expression in the future generation will allow us to evaluate if rapid transcriptional regulation occurs.

Plant genomes are rapidly evolving and their capacity to adapt to environmental changes is crucial. Like genome hybridization and TE mobilization, CNV is one important tool of genome evolution (Kondrashov 2012; Gabur et al. 2019; Quadrana et al. 2019). Together with previous observation, our data show the importance of systematically detecting CNVs. CNVs can indeed associate with adaptive traits (Kondrashov 2012; Gabur et al. 2019; Alonge et al. 2020). In our case, we found a potential link between CNV and pathogen resistance (Fig. 6). CNVs were already shown to be implicated in nematode resistance in soybean (Cook et al. 2012), but also in potato cultivar genome heterogeneity (Pham et al. 2017). We showed that the CNVs in the 20rDNA lines occurred only in a few generations in controlled growing conditions. This last point is particularly interesting in the context of plant breeding. In addition, TDDOs have the potential to create chimeric genes (Fig. 4). TDDO events can promote cancer cell formation, via the activation of oncogenes (Quigley et al. 2018). In that case, breaking junctions can affect the expression of an oncogene by modifying its regulation by enhancers, for example. In our study, the chimeric genes created are expressed and properly spliced. Although we do not have evidence concerning their potential ability to be translated or if the resultant protein would be functional, it is tempting to speculate that TDDO-mediated chimeric genes can lead to gene novelty as previously described (Chen et al. 2013). Studying the consequences of TDDOs in future generations will certainly shed light on their potential impact on genome evolution and plant adaptation.

Methods

Plant materials

Seeds corresponding to the *fas1-4* (SAIL_662_D10) and *fas2-4* (SALK_033228) were previously reported (Exner et al. 2006). All 20rDNA seeds that include *fas1-4* and *fas2-4* parental lines, as well as L6 and L9 lines used in this study correspond to stock previously reported (Pavlištová et al. 2016). For NADs identification, wild-type Col-0 expressing the FIB2:YFP fusion protein was described in Pontvianne et al. (2013). The 20rDNA L6F5 line was transformed by agroinfiltration to insert a transgene expressing FIB2:YFP fusion protein as described previously (Pontvianne et al. 2013).

Nanopore sequencing and data analyses

Genomic DNA preparation was performed as previously described (Debladis et al. 2017). After Qubit dosage (dsDNA High Sensitivity, Thermo Fisher Scientific), a second step of DNA purification was performed with the Genomic DNA Clean and Concentrator kit (Zymo Research) and precipitated. A last Qubit dosage was performed before library preparation using the 1D Genomic DNA by ligation kit SQK-LSK109 (Oxford Nanopore Technologies), following the manufacturer's instructions. The R9.5 ONT flow-cell FLO-MIN106D (Oxford Nanopore Technologies) was used. We obtained 6.4 Gb of sequences for L6F6, 0.7 Gb for L9F6, 5.9 Gb for *fas1-4*, and 11.4 Gb for *fas2-4*.

ONT reads were mapped on the TAIR10 reference genome using minimap2 with -a -Q -map-ont options (Li 2018). The alignment files were converted into BED files using BEDTools, and the coverage per 100-kb window was calculated using coverageBED (Quinlan and Hall 2010). For each 100-kb window, the ratio r = 20%rDNA coverage/wild-type Col-0 coverage was calculated. The mean (m) and standard error (SE) were calculated across the entire genome. Differentially covered regions in the 20%rDNA line were defined as regions for which $r \geq m+2SE$ or $r \leq m-2SE$.

Additional methods can be found in the Supplemental Material.

Data access

All raw and processed sequencing data generated in this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under accession number PRJEB35832. Code used to produce Hi-C figures is available as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Rémy Merret and Michèle Laudié for Illumina sequencing and laboratory members for fruitful discussions. We also thank the flow cytometry facility, the microscopic facility, and the sequencing facility of Perpignan University *Via Domitia* Bioenvironnement (Perpignan, France). This work and A.P.-P.'s PhD fellowship are supported by the Agence Nationale de la Recherche (ANR), JCJC NucleoReg (ANR-15-CE12-0013-01) to F.P. F.P. was supported by the French Laboratory of Excellence project TULIP (ANR-10-LABX-41 and ANR-11-IDEX-0002-02). Work conducted at Iowa State University was supported by Hatch Act and State of Iowa funds. M.M. is a member of the European Training Network "*EpiDiverse*" that receives funding from the European Union Horizon 2020 program under Marie Skłodowska-Curie grant agreement No. 764965. M.D. and M.F. were supported by the Czech Science Foundation project 19-11880Y; by Ministry of Education, Youth and Sports of the Czech Republic INTER-COST (LTC18048); and by European Regional Development Fund, Project "SINGING PLANT" (CZ.02.1.01/0.0/0.0/16_026/0008446). A.P.-P., S.G., N.P., M.F., M.D., and F.P. are part of the European Cooperation in Science and Technology COST ACTION CA16212 INDEPTH.

Author contributions: T.J.B., V.P., and T.R.M. performed nematode infection assays. L.N. and T.H. performed *Pseudomonas DC3000* infection assays. M.F. and M.D. performed DNA-FISH analyses. S.G. performed the Hi-C analyses. F.P. and A.P.-P. conceived and designed all other experiments. A.P.-P., S.G., C.L., E.J., J.D., and F.P. collected the data. P.Z. and M.M. contributed data or analysis tools. A.P.-P., N.P., M.M., and F.P. performed the analysis. F.P. wrote the paper. A.P.-P., S.G., and M.M. edited the paper. F.P. acquired main funding.

References

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al. 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182:** 145–161.e23. doi:10.1016/j.cell.2020.05.021

Bersaglieri C, Santoro R. 2019. Genome organization in and around the nucleolus. *Cells* **8:** 579. doi:10.3390/cells8060579

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16:** 1679–1691. doi:10.1105/tpc.021410

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* **12:** 1093–1101. doi:10.1105/tpc.12.7.1093

Carpentier MC, Picart-Picolo A, Pontvianne F. 2018. A method to identify nucleolus-associated chromatin domains (NADs). *Methods Mol Biol* **1675:** 99–109. doi:10.1007/978-1-4939-7318-7_7

Chandrasekhara C, Mohannath G, Blevins T, Pontvianne F, Pikaard CS. 2016. Chromosome-specific NOR inactivation explains selective rRNA gene silencing and dosage control in *Arabidopsis*. *Genes Dev* **30:** 177–190. doi:10.1101/gad.273755.115

Charbonnel C, Gallego ME, White CI. 2010. Xrcc1-dependent and Ku-dependent DNA double-strand break repair kinetics in Arabidopsis plants. *Plant J Cell Mol Biol* **64:** 280–290. doi:10.1111/j.1365-313X.2010.04331.x

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet* **14:** 645–660. doi:10.1038/nrg3521

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, et al. 2012. Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* **338:** 1206–1209. doi:10.1126/science.1228746

Copenhaver GP, Pikaard CS. 1996. RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. *Plant J* **9:** 259–272. doi:10.1046/j.1365-313X.1996.09020259.x

Debladis E, Llauro C, Carpentier M-C, Mirouze M, Panaud O. 2017. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. *BMC Genomics* **18:** 537. doi:10.1186/s12864-017-3753-z

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci* **104:** 19920–19925. doi:10.1073/pnas.0709888104

Exner V, Taranto P, Schonrock N, Gruissem W, Hennig L. 2006. Chromatin assembly factor CAF-1 is required for cellular differentiation during plant development. *Dev Camb Engl* **133:** 4163–4172. doi:10.1242/dev.02599

Fulgione A, Hancock AM. 2018. Archaic lineages broaden our view on the history of *Arabidopsis thaliana*. *New Phytol* **219:** 1194–1198. doi:10.1111/nph.15244

Gabur I, Chawla HS, Snowdon RJ, Parkin IAP. 2019. Connecting genome structural variation with complex traits in crop plants. *Theor Appl Genet* **132:** 733–750. doi:10.1007/s00122-018-3233-0

Gibbons JG, Branco AT, Godinho SA, Yu S, Lemos B. 2015. Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proc Natl Acad Sci* **112:** 2485–2490. doi:10.1073/pnas.1416878112

Grob S, Schmid MW, Grossniklaus U. 2014. Hi-C analysis in *Arabidopsis* identifies the *KNOT*, a structure with similarities to the *flamenco* locus of *Drosophila*. *Mol Cell* **55:** 678–693. doi:10.1016/j.molcel.2014.07.009

Grummt I, Längst G. 2013. Epigenetic control of RNA polymerase I transcription in mammalian cells. *Biochim Biophys Acta* **1829:** 393–404. doi:10.1016/j.bbagrm.2012.10.004

Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. *Genome Res* **27:** 1461–1474. doi:10.1101/gr.215566.116

Henry Y, Bedhomme M, Blanc G. 2006. History, protohistory and prehistory of the *Arabidopsis thaliana* chromosome complement. *Trends Plant Sci* **11:** 267–273. doi:10.1016/j.tplants.2006.04.002

Himmelbach A, Ruban A, Walde I, Šimková H, Doležel J, Hastie A, Stein N, Mascher M. 2018. Discovery of multi-megabase polymorphic inversions by chromosome conformation capture sequencing in large-genome plant species. *Plant J Cell Mol Biol* **96:** 1309–1316. doi:10.1111/tpj.14109

Ide S, Miyazaki T, Maki H, Kobayashi T. 2010. Abundance of ribosomal RNA gene copies maintains genome integrity. *Science* **327:** 693–696. doi:10.1126/science.1179044

Jiao WB, Schneeberger K. 2020. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun* **11:** 989. doi:10.1038/s41467-020-14779-y

Kobayashi T. 2011. Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. *Cell Mol Life Sci* **68:** 1395–1403. doi:10.1007/s00018-010-0613-2

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc Biol Sci* **279:** 5048–5057. doi:10.1098/rspb.2012.1108

Krasileva KV. 2019. The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. *Curr Opin Plant Biol* **48:** 18–25. doi:10.1016/j.pbi.2019.01.004

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645. doi:10.1101/gr.092759.109

Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293. doi:10.1126/science.1181369

Loehlin DW, Carroll SB. 2016. Expression of tandem gene duplicates is often greater than twofold. *Proc Natl Acad Sci* **113:** 5988–5992. doi:10.1073/pnas.1605886113

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjalmsson BJ, Korte A, Nizhynska V, et al. 2013. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat Genet* **45:** 884–890. doi:10.1038/ng.2678

Mozgová I, Mokroš P, Fajkus J. 2010. Dysfunction of Chromatin Assembly Factor 1 induces shortening of telomeres and loss of 45S rDNA in *Arabidopsis thaliana*. *Plant Cell* **22:** 2768–2780. doi:10.1105/tpc.110.076182

Mozgová I, Wildhaber T, Liu Q, Abou-Mansour E, L'Haridon F, Métraux JP, Gruissem W, Hofius D, Hennig L. 2015. Chromatin assembly factor CAF-1 represses priming of plant defence response genes. *Nat Plants* **1:** 15127. doi:10.1038/nplants.2015.127

Nelson JO, Watase GJ, Warsinger-Pepe N, Yamashita YM. 2019. Mechanisms of rDNA copy number maintenance. *Trends Genet* **35:** 734–742. doi:10.1016/j.tig.2019.07.006

Németh A, Conesa A, Santoyo-Lopez J, Medina I, Montaner D, Peterfia B, Solovei I, Cremer T, Dopazo J, Langst G. 2010. Initial genomics of the human nucleolus. *PLoS Genet* **6:** e1000889. doi:10.1371/journal.pgen.1000889

Newman S, Hermetz KE, Weckselblatt B, Rudd MK. 2015. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am J Hum Genet* **96:** 208–220. doi:10.1016/j.ajhg.2014.12.017

Nurmberg PL, Knox KA, Yun BW, Morris PC, Shafiei R, Hudson A, Loake GJ. 2007. The developmental selector *AS1* is an evolutionarily conserved regulator of the plant immune response. *Proc Natl Acad Sci* **104:** 18795–18800. doi:10.1073/pnas.0705586104

Paredes S, Maggert KA. 2009. Ribosomal DNA contributes to global chromatin regulation. *Proc Natl Acad Sci* **106:** 17829–17834. doi:10.1073/pnas.0906811106

Paredes S, Branco AT, Hartl DL, Maggert KA, Lemos B. 2011. Ribosomal DNA deletions modulate genome-wide gene expression: "rDNA-sensitive" genes and natural variation. *PLoS Genet* **7:** e1001376. doi:10.1371/journal.pgen.1001376

Pavlištová V, Dvořáčková M, Jež M, Mozgová I, Mokroš P, Fajkus J. 2016. Phenotypic reversion in *fas* mutants of *Arabidopsis thaliana* by reintroduction of *FAS* genes: variable recovery of telomeres with major spatial rearrangements and transcriptional reprogramming of 45S rDNA genes. *Plant J Cell Mol Biol* **88:** 411–424. doi:10.1111/tpj.13257

Pham GM, Newton L, Wiegert-Rininger K, Vaillancourt B, Douches DS, Buell CR. 2017. Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *Plant J Cell Mol Biol* **92:** 624–637. doi:10.1111/tpj.13706

Picart-Picolo A, Picault N, Pontvianne F. 2019. Ribosomal RNA genes shape chromatin domains associating with the nucleolus. *Nucleus* **10:** 67–72. doi:10.1080/19491034.2019.1591106

Picart-Picolo A, Picart C, Picault N, Pontvianne F. 2020. Nucleolus-associated chromatin domains are maintained under heat stress, despite nucleolar reorganization in *Arabidopsis thaliana*. *J Plant Res* **133:** 463–470. doi:10.1007/s10265-020-01201-3

Pontvianne F, Grob S. 2020. Three-dimensional nuclear organization in *Arabidopsis thaliana*. *J Plant Res* **133:** 479–488. doi:10.1007/s10265-020-01185-0

Pontvianne F, Liu C. 2020. Chromatin domains in space and their functional implications. *Curr Opin Plant Biol* **54:** 1–10. doi:10.1016/j.pbi.2019.11.005

Pontvianne F, Abou-Ellail M, Douet J, Comella P, Matia I, Chandrasekhara C, Debures A, Blevins T, Cooke R, Medina FJ, et al. 2010. Nucleolin is required for DNA methylation state and the expression of rRNA gene variants in *Arabidopsis thaliana*. *PLoS Genet* **6:** e1001225. doi:10.1371/journal.pgen.1001225

Pontvianne F, Blevins T, Chandrasekhara C, Feng W, Stroud H, Jacobsen SE, Michaels SD, Pikaard CS. 2012. Histone methyltransferases regulating rRNA gene dose and dosage control in *Arabidopsis*. *Genes Dev* **26:** 945–957. doi:10.1101/gad.182865.111

Pontvianne F, Blevins T, Chandrasekhara C, Mozgova I, Hassel C, Pontes OMF, Tucker S, Mokros P, Muchova V, Fajkus J, et al. 2013. Subnuclear partitioning of rRNA genes between the nucleolus and nucleoplasm reflects alternative epiallelic states. *Genes Dev* **27:** 1545–1550. doi:10.1101/gad.221648.113

Pontvianne F, Boyer-Clavel M, Sáez-Vásquez J. 2016a. Fluorescence-activated nucleolus sorting in Arabidopsis. *Methods Mol Biol* **1455:** 203–211. doi:10.1007/978-1-4939-3792-9_15

Pontvianne F, Carpentier MC, Durut N, Pavlištová V, Jaške K, Schořová S, Parrinello H, Rohmer M, Pikaard CS, Fojtová M, et al. 2016b. Identification of nucleolus-associated chromatin domains reveals a role for the nucleolus in 3D organization of the *A. thaliana* genome. *Cell Rep* **16:** 1574–1587. doi:10.1016/j.celrep.2016.07.016

Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui M-A, Guy J, Bortolini Silveira A, Engelen S, Baillet V, Wincker P, et al. 2019. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nat Commun* **10:** 3421. doi:10.1038/s41467-019-11385-5

Quigley DA, Dang HX, Zhao SG, Lloyd P, Aggarwal R, Alumkal JJ, Foye A, Kothari V, Perry MD, Bailey AM, et al. 2018. Genomic hallmarks and structural variation in metastatic prostate cancer. *Cell* **174:** 758–769.e9. doi:10.1016/j.cell.2018.06.039

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA, Bhat P, Takei Y, et al. 2018. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174:** 744–757.e24. doi:10.1016/j.cell.2018.05.024

Rabanal FA, Nizhynska V, Mandáková T, Novikova PY, Lysak MA, Mott R, Nordborg M. 2017. Unstable inheritance of 45S rRNA genes in *Arabidopsis thaliana*. *G3 (Bethesda)* **7:** 1201–1209. doi:10.1534/g3.117.040204

Ramírez-Parra E, Gutierrez C. 2007. The many faces of chromatin assembly factor 1. *Trends Plant Sci* **12:** 570–576. doi:10.1016/j.tplants.2007.10.002

Santos AP, Gaudin V, Mozgová I, Pontvianne F, Schubert D, Tek AL, Dvořáčková M, Liu C, Fransz P, Rosa S, et al. 2020. Tidying-up the plant nuclear space: domains, function and dynamics. *J Exp Bot* **71:** 5160–5179. doi:10.1093/jxb/eraa282

Takeuchi Y, Horiuchi T, Kobayashi T. 2003. Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes Dev* **17:** 1497–1506. doi:10.1101/gad.1085403

Tartof KD. 1974a. Unequal mitotic sister chromatid exchange and disproportionate replication as mechanisms regulating ribosomal RNA gene redundancy. *Cold Spring Harb Symp Quant Biol* **38:** 491–500. doi:10.1101/SQB.1974.038.01.053

Tartof KD. 1974b. Unequal mitotic sister chromatin exchange as the mechanism of ribosomal RNA gene magnification. *Proc Natl Acad Sci* **71:** 1272–1276. doi:10.1073/pnas.71.4.1272

van Koningsbruggen S, Gierliński M, Schofield P, Martin D, Barton GJ, Ariyurek Y, den Dunnen JT, Lamond AI. 2010. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol Biol Cell* **21:** 3735–3748. doi:10.1091/mbc.e10-06-0508

Varas J, Santos JL, Pradillo M. 2017. The absence of the Arabidopsis chaperone complex CAF-1 produces mitotic chromosome abnormalities and changes in the expression profiles of genes involved in DNA repair. *Front Plant Sci* **8:** 525. doi:10.3389/fpls.2017.00525

Wang M, Lemos B. 2017. Ribosomal DNA copy number amplification and loss in human cancers is linked to tumor genetic context, nucleolus activity, and proliferation. *PLoS Genet* **13:** e1006994. doi:10.1371/journal.pgen.1006994

Wee Y, Wang T, Liu Y, Li X, Zhao M. 2018. A pan-cancer study of copy number gain and up-regulation in human oncogenes. *Life Sci* **211:** 206–214. doi:10.1016/j.lfs.2018.09.032

Wubben MJE, Jin J, Baum TJ. 2008. Cyst nematode parasitism of *Arabidopsis thaliana* is inhibited by salicylic acid (SA) and elicits uncoupled SA-independent pathogenesis-related gene expression in roots. *Mol Plant-Microbe Interact* **21:** 424–432. doi:10.1094/MPMI-21-4-0424

Xu B, Li H, Perry JM, Singh VP, Unruh J, Yu Z, Zakari M, McDowell W, Li L, Gerton JL. 2017. Ribosomal DNA copy number loss and sequence variation in cancer. *PLoS Genet* **13:** e1006771. doi:10.1371/journal.pgen.1006771

Yang JY, Iwasaki M, Machida C, Machida Y, Zhou X, Chua NH. 2008. βC1, the pathogenicity factor of TYLCCNV, interacts with AS1 to alter leaf development and suppress selective jasmonic acid responses. *Genes Dev* **22:** 2564–2577. doi:10.1101/gad.1682208

Yeh YH, Chang YH, Huang PY, Huang JB, Zimmerli L. 2015. Enhanced Arabidopsis pattern-triggered immunity by overexpression of cysteine-rich receptor-like kinases. *Front Plant Sci* **6:** 322. doi:10.3389/fpls.2015.00322

Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, Patel V, Velikkakam James G, Koornneef M, Ossowski S, et al. 2016. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci* **113:** E4052–E4060. doi:10.1073/pnas.1607532113

Zhang J, Zuo T, Peterson T. 2013. Generation of tandem direct duplications by reversed-ends transposition of maize ac elements. *PLoS Genet* **9:** e1003691. doi:10.1371/journal.pgen.1003691

Received January 23, 2020; accepted in revised form September 15, 2020.



Chapter 7

Identification of Extrachromosomal Circular Forms of Active Transposable Elements Using Mobilome-Seq

Sophie Lanciano, Panpan Zhang, Christel Llauro, and Marie Mirouze

Abstract

Active transposable elements (TEs) generate insertion polymorphisms that can be detected through genome resequencing strategies. However, these techniques may have limitations for organisms with large genomes or for somatic insertions. Here, we present a method that takes advantage of the extrachromosomal circular DNA (eccDNA) forms of actively transposing TEs in order to detect and characterize active TEs in any plant or animal tissue. Mobilome-seq consists in selectively amplifying and sequencing eccDNAs. It relies on linear digestion of genomic DNA followed by rolling circle amplification of circular DNA. Both active DNA transposons and retrotransposons can be identified using this technique.

Key words Extrachromosomal circular DNA (eccDNA), Plasmid safe, Rolling circle DNA amplification (RCA), Phi29, Transposable elements (TEs), Retrotransposons, Mobilome sequencing

1 Introduction

Transposable elements (TEs) represent a main source of genomic diversity and an evolutionary force in both plants and animals [1]. However, the mobile part of the genome or mobilome, comprising active TEs of a given organism is difficult to characterize. To establish an unbiased repertoire of mobile TEs, we have developed a simple strategy based on high-throughput sequencing to detect TEs in their extrachromosomal circular DNA (eccDNA) forms. These forms are presumably by-products of their life cycle [2]. They are most likely produced by the TE host, in the nucleus, through homologous recombination or non-homologous end-joining of linear extrachromosomal molecules. Recently, several methods to analyze extrachromosomal circular DNA have been reported [3–6]. Here, we describe an easy procedure that can be used starting from any genomic DNA extraction and that does not require any special equipment. Using this method, we have successfully identified active TEs in different plant models (Arabidopsis [7], rice [8], peanut [9], potato [10], poplar [11]) and animal

Jungnam Cho (ed.), *Plant Transposable Elements: Methods and Protocols*, Methods in Molecular Biology, vol. 2250,
https://doi.org/10.1007/978-1-0716-1134-0_7, © Springer Science+Business Media, LLC, part of Springer Nature 2021

models (Drosophila [12]). We have further contributed to identify viral eccDNA in Drosophila [13]. The mobilome-seq represents a novel approach to understand and evaluate the extent and impact of TE mobility on eukaryotic genomes.

2 Materials

Prepare all solutions using ultrapure water (e.g., Gibco water) and analytical grade reagents. Prepare and store all reagents at room temperature. Use filter tips to prevent cross-contaminations (*see* **Note 1**). Carefully follow all waste disposal regulations when disposing waste materials.

2.1 Genomic DNA Extraction and Purification

1. Qiagen DNeasy DNA extraction kit or CTAB buffer.
2. Qiagen PCR purification kit.

2.2 Linear DNA Digestion and Rolling Circle Amplification

1. Plasmid-Safe™ ATP-Dependent DNase (EpiCENTRE, Tebubio) at 10 U/μL.
2. Illustra TempliPhi Amplification Kit (GE healthcare).
3. Glycogen (Fisher).
4. NaOAc 3 M (μL).
5. Ethanol 100° and 70°.
6. PCR machine.
7. 28 °C and 37 °C incubators.
8. −20 °C freezer.
9. 4 °C centrifuge (14,462 × *g*).
10. Restriction enzyme such as HindIII or EcoRI.
11. Gel loading buffer.
12. 1% agarose gel.

2.3 Library Preparation

1. DNA PicoGreen kit (Invitrogen) or Quantus (Promega) for DNA quantification.
2. Nextera XT library kit (Illumina).
3. DNA High Sensitivity Bioanalyzer chip (Agilent Technologies).
4. MiSeq sequencer (Illumina).

3 Methods

The preparation of eccDNA sequencing consists in three different steps: (1) the selection, (2) amplification of eccDNA molecules, and (3) the library preparation and sequencing. After DNA extraction,

linear DNA molecules are digested and eccDNA are randomly amplified using rolling circle amplification (RCA). Then, this DNA material is used for high-throughput sequencing.

3.1 Genomic DNA Extraction and Purification

1. Use DNeasy Qiagen kit or CTAB extraction to extract genomic DNA from plant or animal tissue (*see* **Note 2**).
2. Use around 1–5 μg of genomic DNA (obtained in **step 1**, *see* **Note 3**) to perform PCR purification that will remove large genomic fragments using Qiagen PCR purification kit. Follow manufacturer's indications but elute in 30 μL of water.

3.2 Linear DNA Digestion and Rolling Circle Amplification

1. Digest linear DNA with the Plasmid Safe enzyme. In Eppendorf tubes, in a final volume of 50 μL add: 25 μL of purified genomic DNA, 17 μL of water, 2 μL of ATP (25 mM), 5 μL of Plasmid Safe Buffer, 1 μL of Plasmid Safe enzyme. Incubate for 17 h in a 37 °C incubator.
2. Perform ethanol precipitation by adding to the Plasmid Safe reaction tube: 50 μL of water, 10 μL of NaOAc (3 M, pH 5.2), 250 μL of Ethanol 100°, 1 μL of Glycogen. Precipitate overnight in a −20 °C freezer (*see* **Note 4**).
3. Centrifuge the tubes at 4 °C for 1 h at 14,462 × *g*. Wash the pellet by adding 500 μL of ice-cold Ethanol 70°. Centrifuge at 4 °C for 10 min. Let the tubes air-dry (*see* **Note 5**). Resuspend the pellet directly in 5 μL of TempliPhi Sample Buffer (*see* **Note 6**). Transfer in PCR tubes for the TempliPhi reaction.
4. Perform the TempliPhi reaction for random rolling circle amplification (RCA) of the circular DNA as follows. Denaturate the DNA for 3 min at 95 °C in a PCR machine. Perform the reaction by adding to the PCR reaction tubes: 5 μL of TempliPhi Reaction Buffer and 0.2 μL of TempliPhi Enzyme (*see* **Note 7**). Incubate for 65 h in a 28 °C incubator (*see* **Note 8**).
5. Inactivate the enzyme by incubating the tubes for 30 min at 70 °C in a PCR machine. Store indefinitely at −20 °C (*see* **Note 9**).
6. Perform quality control of the amplification by restriction digestion (*see* **Note 10**) and gel electrophoresis. Transfer 2 μL of amplified DNA in a new PCR tube and perform the digestion in a final volume of 10 μL by adding 6 μL of water, 1 μL of enzyme buffer, and 1 μL of restriction enzyme. Incubate overnight in a 37 °C incubator.
7. Add loading buffer to the digestion product. Run the digestion product on a 1% agarose gel with a size ladder (*see* **Note 11**). *See* Fig. 1 for an example of TempliPhi products after digestion.

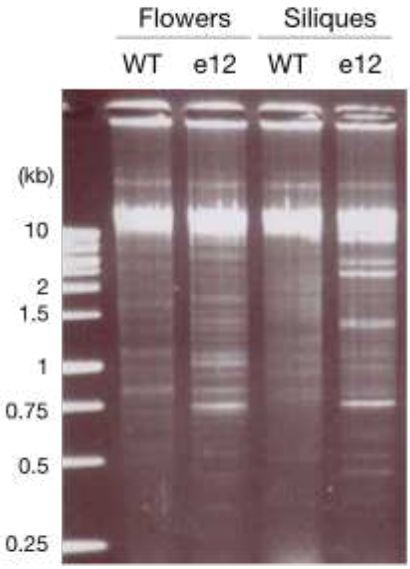


Fig. 1 Example of eccDNA amplification using TempliPhi. Genomic DNA was extracted from *Arabidopsis thaliana* tissues as indicated (WT: wild-type Columbia-0, e12: epiRIL12 [16]) and processed as described in Subheading 3. TempliPhi products were digested with XbaI and loaded on a 1% agarose gel

3.3 Library Preparation

1. Measure DNA concentration with DNA PicoGreen reagent or with a Quantus machine and dilute samples to 0.2 ng/μL.
2. Use 1–2 ng of DNA to prepare libraries with the Nextera XT library kit following the manufacturer's instructions (*see* **Note 12**).
3. After the 12 PCR cycles performed with index primers to amplify libraries, perform a quality control on a High Sensitivity DNA Bioanalyzer chip following the manufacturer's instructions. If necessary, normalize the libraries following the manufacturer's recommendations (*see* **Note 13**).
4. Sequence on a MiSeq sequencer (*see* **Note 14**).
5. Analyze reads as described in Lanciano et al. [7] (*see* **Note 15**).

4 Notes

1. We have found useful to dedicate a bench and a set of pipettes to library preparation if possible. This bench should be carefully cleaned with Ethanol 70° after each use.
2. We have used genomic DNA extractions from various sources and various DNA extraction methods without any impact on the following steps. We nevertheless recommend a classical CTAB extraction as it usually results in a higher amount of genomic DNA. Circular DNA is quite stable and samples can

be sent at this stage in dry ice or as precipitated pellets in Ethanol 100° or dry pellets at room temperature.

3. We routinely use 1 µg of genomic DNA (quantified with Qubit or Quantus spectrophotometers) but the procedure can be used with 500 ng.
4. Glycogen is used to facilitate the precipitation of tiny amounts of DNA at this step (when linear DNA is degraded and circular DNA not yet amplified) and to allow the visualization of the pellet. Here, the protocol can be stopped at −20 °C for at least several weeks.
5. We usually let the tubes air-dry on the bench but laminar fluxes can be used to accelerate the process. However, be careful to prevent over-drying the pellet as it might become difficult to resuspend.
6. It is crucial at this step to properly collect the pellet but resuspension can take long (a few minutes per tube). As the pellet is transparent, we estimate that the pellet is correctly resuspended when no adhesion is visible when pipetting in the bottom of the tube (where the pellet was located).
7. We usually prepare a reaction mix for this step if several tubes are processed at the same time.
8. We use this long incubation time (2 days and 3 nights) to maximize the RCA reaction. The temperature can be increased to 29 °C.
9. We have successfully stored RCA products at −20 °C for several years without affecting the following steps.
10. RCA products (produced by the bacteriophage Phi29 enzyme [14]) are typically complex molecules. The purpose of this step is to visualize on gel the efficiency of the amplification step. For this reason, a restriction digestion is necessary. Any frequent 6 bp cutter can be used, we routinely use HindIII or EcoRI.
11. The digestion product can appear as a smear or with some discrete bands. Both patterns suggest a correct amplification and might depend on the sample (population of eccDNA, mitochondrial DNA, and chloroplastic DNA). Note that these plastid circular DNA molecules are not eliminated at this stage; they will be depleted by downstream computational filtering. However, if the presence of these plastid circular molecules is known to be very abundant (e.g., the presence of mitochondrial plasmids in maize) they could decrease the analysis sensitivity. To avoid this undesirable effect, these molecules can be digested using the CRISPR-DASH technique [15] directly after the linear DNA digestion (Subheading 3.2, **step 1**). A single guide RNA (sgRNA) is designed and synthesized to target unwanted DNA. For the CRISPR/Cas9 reaction, a

ratio of 20/20/1(sgRNA/Cas9/DNA) is recommended to completely eliminate targeted DNA. This CRISPR/Cas9 digestion product is then used to repeat the linear DNA digestion (Subheading 3.2, **step 1**) and all steps described in Subheading 3.2.

12. We generally use a 1/20 dilution of or RCA products that corresponds to the desired concentration.
13. We estimate that 1 Gb of sequencing reads is generally sufficient to identify active TE candidates, given that only a fraction of the genome leading to eccDNA is sequenced. Therefore, in a MiSeq run we generally pool 12 libraries.
14. Use 300 nt paired-end reads in order to decrease the ambiguity of subsequent mapping.
15. For a detailed computational analysis, please follow our depository at https://github.com/njaupan.

Acknowledgments

This work was funded by IRD, a French ANR grant (ANR-13-JSV6-0002 "*ExtraChrom*") and a EU Horizon 2020 programme under the Marie Skłodowska-Curie grant agreement No. 764965 ("*EpiDiverse*") to M.M.

References

1. Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14:49–61
2. Lanciano S, Mirouze M (2018) Transposable elements: all mobile, all different, some stress responsive, some adaptive? Curr Opin Genet Dev 49:106–114
3. Møller HD, Parsons L, Jørgensen TS, Botstein D, Regenberg B (2015) Extrachromosomal circular DNA is common in yeast. Proc Natl Acad Sci U S A 112(24): E3114–E3122
4. Møller HD, Bojsen RK, Tachibana C, Parsons L, Botstein D, Regenberg B (2016) Genome-wide purification of extrachromosomal circular DNA from eukaryotic cells. J Vis Exp 110:e54239
5. Mehta D, Hirsch-Hoffmann M, Were M, Patrignani A, Shan-e-Ali Zaidi S, Were H, Gruissem W, Vanderschuren H (2019) A new full-length circular DNA sequencing method for viral-sized genomes reveals that RNAi transgenic plants provoke a shift in geminivirus populations in the field. Nucleic Acids Res 47 (2):e9
6. Shoura MJ, Gabdank I, Hansen L, Merker J, Gotlib J, Levene SD, Fire AZ (2017) Intricate and cell- type-specific populations of endogenous circular DNA (eccDNA) in *Caenorhabditis elegans* and *Homo sapiens*. G3 7:3295–3303
7. Lanciano S, Carpentier MC, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, Ghesquière A, Panaud O, Mirouze M (2017) Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. PLoS Genet 13(2):e1006630
8. Thieme M, Lanciano S, Balzergue S, Daccord N, Mirouze M, Bucher E (2017) Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding. Genome Biol 18(1):134
9. Bertioli DJ et al (2019) The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. Nat Genet 51(5):877–884
10. Esposito S, Barteri F, Casacuberta J, Mirouze M, Carputo D, Aversano R (2019) LTR-TEs abundance, timing and mobility in *Solanum commersonii* and S. *tuberosum* genomes following cold-stress conditions. Planta 250(5):1781–1787

11. Sow MD, Le Gac AL, Fichot R, Lanciano S, Delaunay A, Le Jan I, Lesage-Descauses MC, Citerne S, Caius J, Brunaud V, Soubigou-Taconnat L, Cochard H, Segura V, Chaparro C, Grunau C, Tost J, Brignolas F, Strauss SH, Mirouze M, Maury S (Manuscript submitted) Hypomethylated poplars show higher tolerance to water deficit and highlight dual role of DNA methylation in shoot meristem: regulation of stress response and of genome integrity. https://doi.org/10.1101/2020.04.16.045328
12. Barckmann B, El-Barouk M, Pélisson A, Mugat B, Li B, Franckhauser C, Fiston Lavier AS, Mirouze M, Fablet M, Chambeyron S (2018) The somatic piRNA pathway controls germline transposition over generations. Nucleic Acids Res 46(18):9524–9536
13. Poirier EZ, Goic B, Tomé-Poderti L, Frangeul L, Boussier J, Gausson V, Blanc H, Vallet T, Loyd H, Levi LI, Lanciano S, Baron C, Merkling SH, Lambrechts L, Mirouze M, Carpenter S, Vignuzzi M, Saleh MC (2018) Dicer-2-dependent generation of viral DNA from defective genomes of RNA viruses modulates antiviral immunity in insects. Cell Host Microbe 23(3):353–365.e8
14. Johne R, Müller H, Rector A, van Ranst M, Stevens H (2009) Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends Microbiol 17:205–211
15. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL (2016) Depletion of abundant sequences by hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol 17:41
16. Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D, Paszkowski J, Mathieu O (2009) Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461:427–430

Plant Biotechnology Journal

Plant Biotechnology Journal (2022), pp. 1–20 doi: 10.1111/pbi.13775

Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates

Adam Nunn[1,2], Isaac Rodríguez-Arévalo[3,4], Zenith Tandukar[5], Katherine Frels[5,6], Adrián Contreras-Garrido[7], Pablo Carbonell-Bejerano[7], Panpan Zhang[8,9], Daniela Ramos Cruz[3,4], Katharina Jandrasits[3,4], Christa Lanz[7], Anthony Brusa[5], Marie Mirouze[8,9], Kevin Dorn[10,11], David W Galbraith[12], Brice A. Jarvis[13], John C. Sedbrook[13], Donald L. Wyse[5], Christian Otto[1], David Langenberger[1], Peter F. Stadler[2,14], Detlef Weigel[7], M. David Marks[10], James A. Anderson[5], Claude Becker[3,4,*] and Ratan Chopra[5,10,*]

[1]*ecSeq Bioinformatics GmbH, Leipzig, Germany*
[2]*Department of Computer Science, Leipzig University, Leipzig, Germany*
[3]*Genetics, Faculty of Biology, Ludwig Maximilians University, Martinsried, Germany*
[4]*Gregor Mendel Institute of Molecular Plant Biology GmbH, Austrian Academy of Sciences (ÖAW), Vienna BioCenter (VBC), Vienna, Austria*
[5]*Department of Agronomy and Plant Genetics, University of Minnesota, Saint Paul, MN, USA*
[6]*Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE, USA*
[7]*Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany*
[8]*Institut de Recherche pour le Développement, UMR232 DIADE, Montpellier, France*
[9]*Laboratory of Plant Genome and Development, University of Perpignan, Perpignan, France*
[10]*Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, MN, USA*
[11]*USDA-ARS, Soil Management and Sugarbeet Research, Fort Collins, CO, USA*
[12]*BIOS Institute, Arizona Cancer Center, Department of Biomedical Engineering, University of Arizona, School of Plant Sciences, Tucson, AZ, USA*
[13]*School of Biological Sciences, Illinois State University, Normal, IL, USA*
[14]*Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*

Received 17 August 2021;
revised 28 November 2021;
accepted 23 December 2021.
**Correspondence* (Tel +49 89 2180 74740; fax +49 89 2180 74702; email: claude.becker@biologie.uni-muenchen.de (CB); Tel +1 551-689-5299; fax +1 612-625-1268; email: rchopra@umn.edu (RC))

Keywords: pennycress, genome assembly, genome annotations, genetic mapping, comparative genomics.

Summary

Thlaspi arvense (field pennycress) is being domesticated as a winter annual oilseed crop capable of improving ecosystems and intensifying agricultural productivity without increasing land use. It is a selfing diploid with a short life cycle and is amenable to genetic manipulations, making it an accessible field-based model species for genetics and epigenetics. The availability of a high-quality reference genome is vital for understanding pennycress physiology and for clarifying its evolutionary history within the Brassicaceae. Here, we present a chromosome-level genome assembly of var. MN106-Ref with improved gene annotation and use it to investigate gene structure differences between two accessions (MN108 and Spring32-10) that are highly amenable to genetic transformation. We describe non-coding RNAs, pseudogenes and transposable elements, and highlight tissue-specific expression and methylation patterns. Resequencing of forty wild accessions provided insights into genome-wide genetic variation, and QTL regions were identified for a seedling colour phenotype. Altogether, these data will serve as a tool for pennycress improvement in general and for translational research across the Brassicaceae.

Introduction

Native to Eurasia, field pennycress (*Thlaspi arvense* L.) is a member of the Brassicaceae family and is closely related to the oilseed crop species rapeseed (*Brassica rapa* and *Brassica napus* L.), camelina (*Camelina sativa* L.) and the wild plant *Arabidopsis thaliana* (Beilstein *et al.*, 2010; Warwick *et al.*, 2002). It is an emerging oil feedstock species with the potential to improve sustainability of cold climate cropping systems through use as a cash cover crop (Boateng *et al.*, 2010; Chopra *et al.*, 2018; Sedbrook *et al.*, 2014). Pennycress is extremely winter hardy (Warwick *et al.*, 2002) and can be planted in traditional fallow periods following summer annuals such as wheat, maize or soya bean (Cubins *et al.*, 2019; Johnson *et al.*, 2015; Ott *et al.*, 2019; Phippen and Phippen, 2012). By providing a protective living cover from the harvest of the previous summer annual crop through early spring, pennycress prevents soil erosion and nutrient loss, which in turn protects surface and below-ground

Please cite this article as: Nunn, A., Rodríguez-Arévalo, I., Tandukar, Z., Frels, K., Contreras-Garrido, A., Carbonell-Bejerano, P., Zhang, P., Ramos Cruz, D., Jandrasits, K., Lanz, C., Brusa, A., Mirouze, M., Dorn, K., Galbraith, D.W., Jarvis, B.A., Sedbrook, J.C., Wyse, D.L., Otto, C., Langenberger, D., Stadler, P.F., Weigel, D., Marks, M. D., Anderson, J.A., Becker, C. and Chopra, R. (2022) Chromosome-level *Thlaspi arvense* genome provides new tools for translational research and for a newly domesticated cash cover crop of the cooler climates. *Plant Biotechnol J.*, https://doi.org/10.1111/pbi.13775.

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

water sources, suppresses early-season weed growth, and provides a food source for pollinators (Del Gatto *et al.*, 2015; Johnson *et al.*, 2015; Weyers *et al.*, 2019, 2021). The short life cycle allows for harvest in May or June in temperate regions, with reported seed yields ranging from 750 to 2400 kg/ha (Cubins *et al.*, 2019; Moore *et al.*, 2020). Following harvest, an additional crop of summer annuals can be grown in a double-crop system that provides increased total seed yields and beneficial ecosystem services (Johnson *et al.*, 2015; Phippen and Phippen, 2012; Thomas *et al.*, 2017). The pennycress seed contains an average of 30%–35% oil, and the fatty acid profile is conducive to producing biofuels (Fan *et al.*, 2013; Moser, 2012; Moser *et al.*, 2009). Seed oil also has the potential to be converted into an edible oil and protein source (Chopra *et al.*, 2020b; Claver *et al.*, 2017; McGinn *et al.*, 2019).

Thlaspi arvense is a homozygous diploid species ($2n = 2x = 14$) (Mulligan, 1957) and is predominantly self-pollinating (Mulligan and Kevan, 1973), suggesting that breeding efforts could proceed with relative ease and speed. It is amenable to genetic transformation using the floral dip method (McGinn *et al.*, 2019), and its diploid nature with many one-to-one gene correspondence with *A. thaliana* (Chopra *et al.*, 2018) could provide an avenue for gene discovery followed by field-based phenotypic validation. Indeed, several agronomic and biochemical traits have already been identified in pennycress using this translational approach, including traits crucial for *de novo* domestication of *T. arvense* such as transparent testa phenotypes (Chopra *et al.*, 2018), early flowering (Chopra *et al.*, 2020b), reduced shatter (Chopra *et al.*, 2020b) and seed oil composition traits (Chopra *et al.*, 2020b; Esfahanian *et al.*, 2021; Jarvis *et al.*, 2021; McGinn *et al.*, 2019). Field pennycress could thus serve as a *de novo*-domesticated oilseed crop for the cooler climates of the world and at the same time as a new dicotyledonous model for functional genetics studies. Its amenability for translational research constitutes a clear advantage vis-a-vis *A. thaliana*. However, to establish *T. arvense* as a genetic model and a crop, it is important to develop genomic resources that will help explore the spectrum of genetic diversity, the extent and patterns of gene expression, genetic structure and untapped genetic potential for crop improvement.

Here, we describe a set of new resources developed for research and breeding communities, including a high-quality, chromosome-level genome assembly of *T. arvense* var. MN106-Ref, representing ~97.5% of the estimated genome size of 539 Mbp. We provide robust annotations of both protein-coding and non-coding genes, including putative transfer RNA (tRNA), ribosomal RNA (rRNA) and small nucleolar RNA (snoRNA) predictions, alongside small RNA-producing loci, transposable element (TE) families and predicted pseudogenes. From transcriptome data based on a panel of eleven different tissues and life stages, we built a gene expression atlas. In combination with whole-genome DNA methylation profiles of both roots and shoots, this provides a basis for exploring gene regulatory and/or epigenetic mechanisms within pennycress. A comprehensive analysis of forty resequenced pennycress accessions highlights the nucleotide diversity in these collections, alongside gene variants and population structure. Finally, by means of modified bulked-segregant analysis (BSA), we identified quantitative trait loci (QTL) associated with seedling colour phenotype, exemplifying the usefulness of this resource. The genome and resequencing information presented in this study will increase the value of pennycress as a model and as tool for translational research and accelerate pennycress breeding through the discovery of genes affecting important agronomic traits.

Results

An improved reference genome sequence

The genome of *T. arvense* var. MN106-Ref was assembled *de novo* from 476X (256 Gb) depth PacBio Sequel II continuous long reads (CLRs) (38 kb N50). The initial assembly attempts exceeded the genome size by ~53% with respect to the range of 459–540 Mbp total size estimated from flow cytometry and k-mer analyses (Table S1). Reducing the duplicated fraction, polishing and scaffolding/rescaffolding using several approaches resulted in a final assembly of ~526 Mbp, corresponding to ~97.5% of the upper limit of the flow cytometry-based estimate and representing an improvement of ~20% relative to the original assembly size. Scaffolding/rescaffolding of the genome assembly was achieved using Bionano optical, Hi-C contact, genetic linkage and comparative synteny maps. The final genome contains 964 scaffolds, with ~83.6% of the total estimated size represented by seven large scaffolds, in agreement with the haploid chromosome number, demonstrating a vast improvement in overall contiguity and bringing the assembly to chromosome level. The coding space is 98.7% complete on the basis of conserved core eukaryotic single-copy genes (BUSCO), with 92.1% being single copy and 6.6% duplicated. Full descriptive statistics of the final version in comparison with T_arvense_v1 are given in Table 1; intermediary versions are summarized in Table S2.

The seven largest scaffolds are all characterized by high gene density towards both telomeres and a high density of repeats and TEs in the pericentromeric and centromeric regions (Figure 1, Figure S1). While the protein-coding gene fraction of the genome is similar in size to other closely related Brassicaceae (Wang *et al.*, 2011), the large repetitive fraction suggests an increased genome size driven by TE expansion (Beric *et al.*, 2021). In addition, the spatial distribution of sRNA loci followed the gene density but

Table 1 Full descriptive statistics comparing the previously published T_arvense_v1 assembly with the present version T_arvense_v2

Assembly category	T_arvense_v1	T_arvense_v2
No. of contigs	44 109	4714
Largest contig	–	41.6 Mbp
contig N50	0.02 Mbp	13.3 Mbp
No. of scaffolds	6768	964
No. of scaffolds (≥50 000 bp)	1807	607
Largest scaffold	2.4 Mbp	70.0 Mbp
Total length	343 Mbp	526 Mbp
Total length (≥50 000 bp)	276 Mbp	514 Mbp
GC (%)	37.99	38.39
N50	0.14 Mbp	64.9 Mbp
NG50	0.05 Mbp	64.9 Mbp
N75	0.06 Mbp	61.0 Mbp
NG75	–	55.2 Mbp
L50	561	4
LG50	1678	4
L75	1469	6
LG75	–	7
No. of Ns per 100 kbp	5165.00	0.51

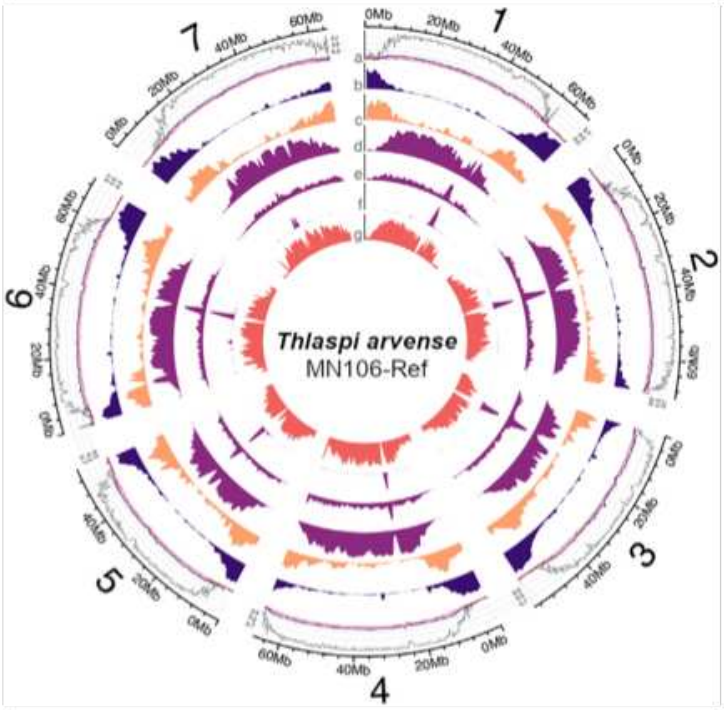


Figure 1 Overview of the seven largest scaffolds representing chromosomes in *T. arvense* var. MN106-Ref. The tracks denote (a) DNA methylation level in shoot tissue (CG: grey; CHG: black; CHH: pink; 200 kbp window size), and density distributions (1 Mbp window size) of (b) protein-coding loci, (c) sRNA loci, (d) Gypsy retrotransposons, (e) Copia retrotransposons, (f) LTR retrotransposons and (g) pseudogenes.

was concentrated predominantly at the boundary between genes and TEs.

In addition to the duplicate-containing contigs, alignments of the raw CLR reads to the new genome revealed the presence of what appeared to be a small number of collapsed repeats in scaffolds 1, 3, 5 and 7, which were typically larger than 25 kbp and indicative of misassembly in these loci (Figure S2). Further investigation revealed an overlap with tandem repeat clusters of 18S and 28S rRNA annotations at those loci on scaffolds 3 and 5, and a large supersatellite of 5S rRNA on scaffold 1. In addition, there were corresponding genes associated with organellar DNA at those loci on scaffolds 3 and 7, indicating either erroneous incorporation of plastome sequence during assembly or genuine nuclear integrations of plastid DNA (NUPTs) (Michalovova *et al.*, 2013).

Comparative genomics

Exploiting information from the genome of *Eutrema salsugineum* (Yang *et al.*, 2013), a closely related species (Franzke *et al.*, 2011) with a much smaller genome (241 Mbp) but the same karyotype ($n = 7$), aided during rescaffolding (see methods; Figure S3) and confirmed synteny of the seven largest scaffolds in the two species (Figure S4). There is a large-scale synteny between the two genomes, with the exception of some regions on scaffolds 2, 3, 6 and 7. This could be due to the low gene density observed in the *T. arvense* genome towards the centre of each chromosome and/or the high presence of dispersed repeats in those regions.

Chromosome evolution in the Brassicaceae has been studied through chromosome painting techniques, and 24 chromosome blocks (A-X) have been defined from an ancestral karyotype of $n = 8$ (Murat *et al.*, 2015; Schranz *et al.*, 2006). We identified the 24 blocks in *T. arvense* based on gene homology and synteny between *T. arvense* and *A. thaliana* (Figure 2). While in general the distribution of the chromosomal blocks resembles that in the close relatives *E. salsugineum* and *S. parvula*, some blocks are rearranged in a small section at the end of the scaffold representing chromosome 1 and at the beginning of chromosome 6. The first case involves the transposition of a small part of block C in between A and B, while chromosome 6 has a possible inversion between the blocks O and W when compared to *E. salsugineum* and *S. parvula*. Overall, despite having an increase in genome size compared with *E. salsugineum* and *S. parvula*, *T. arvense* conserves all the ancestral Brassicaceae karyotype blocks. The synteny analysis also revealed intra-chromosomal rearrangements, but no obvious inter-chromosomal rearrangements.

Genome annotation

Transcriptome assembly

We sequenced total cDNA with strand-specific RNA-seq from eleven tissues, including rosette leaves, cauline leaves, inflorescences, open flowers, young green siliques, old green siliques, green seeds, mature seeds, seed pods, roots of 1-week-old seedlings and shoots of 1-week-old seedlings (Table S3). Reads from each tissue sample were aligned to the genome with unique mapping rates between 76% and 91%, with the exception of old green silique (19%), green seed (59%) and mature seed (12%). The majority of unmapped reads in each case were due to insufficient high-quality read lengths. We constructed independent tissue-specific transcriptome assemblies and combined them

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

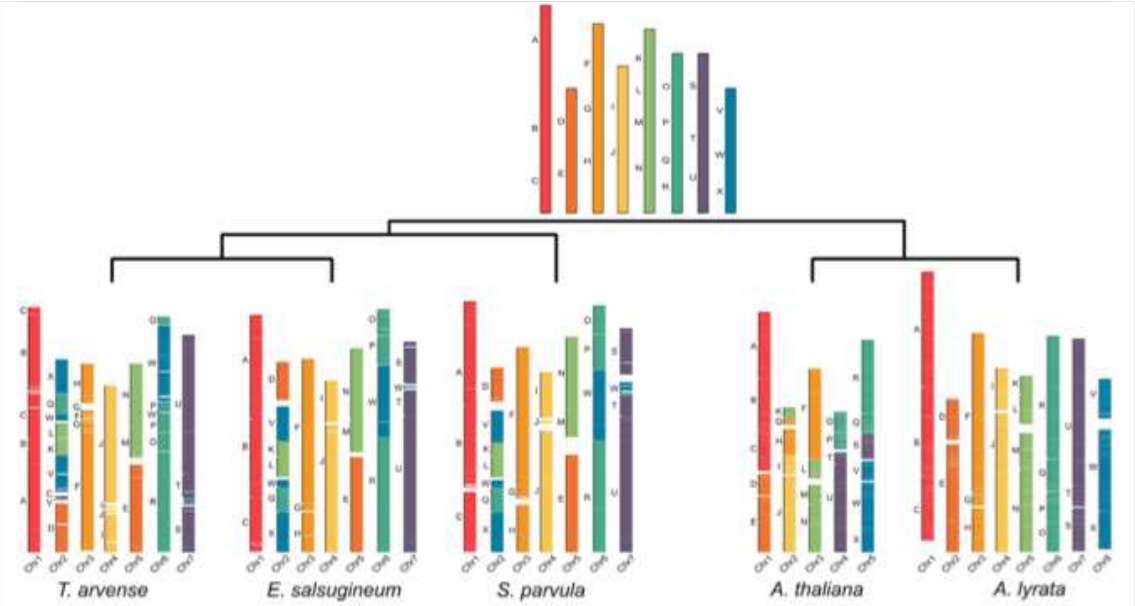


Figure 2 Distribution of ancestral genomic blocks (top panel) along the seven largest scaffolds of *T. arvense* MN106-Ref (T_arvense_v2), and a comparison of these genomic blocks with *Eutrema salsugineum*, *Schrenkiella parvula*, *Arabidopsis thaliana* and *Arabidopsis lyrata*.

into a multi-sample *de novo* assembly, yielding 30 650 consensus transcripts. These were further refined by prioritizing isoforms supported by Iso-seq data, resulting in 22 124 high-quality consensus transcripts to inform gene models.

Protein-coding genes

In addition to the expression data, gene models were informed by protein homology using a combined database of Viridiplantae from UniProtKB/Swiss-Prot (Boutet *et al.*, 2007) and selected Brassicaceae from RefSeq (Pruitt *et al.*, 2012). Following initial training and annotation by *ab initio* gene predictors, protein-coding loci were further annotated with InterPro to provide PFAM domains, which were combined with a BLAST search to the UniProtKB/Swiss-Prot Viridiplantae database to infer gene ontology (GO) terms. In accordance with MAKER-P recommendations (Campbell *et al.*, 2014), the final set of 27 128 protein-coding loci was obtained by filtering out those with an annotation edit distance (AED) score of 1 unless they also contained a PFAM domain. Approximately 95% of loci had an AED score <0.5 (Figure S5), demonstrating a high level of support with the available evidence, and 21 171 (~78%) were annotated with a PFAM domain. Analysis of gene orthologs and paralogs among related Brassicaceae confirmed the close relationship with *E. salsugineum*, with the protein-coding fraction occupying a genome space comparable to related species (Figure 3a). A total of 4433 gene duplication events were recorded with OrthoFinder, comparable to *E. salsugineum* (5108), but fewer than in *B. rapa* (11 513), for example.

The full descriptive statistics are given in Table 2, in comparison with the original T_arvense_v1 annotation (Dorn *et al.*, 2015) lifted over to the new genome with Liftoff v1.5.2 (Shumate and Salzberg, 2020), where applicable. Gene feature distributions are comparable between T_arvense_v1 and the present assembly of MN106-Ref (hereafter referred to as T_arvense_v2; Figure S6). Unique genes that were successfully lifted over from the previous version were included as a separate fraction in the final annotation (source: T_arvense_v1), resulting in 32 010 annotated genes in total. Up to ~95.2% completeness can be obtained by combining the full set of both the current and previous annotations according to a BUSCO evaluation of 2121 conserved, single-copy orthologs. The improved contiguity of the genome space allowed for the resolution of genes such as the tandem duplicated *MYB29* and *MYB76*, which were concatenated in the previous version (Figure 3b).

Non-coding loci

In addition to the protein-coding gene annotations, we annotated non-coding RNA (ncRNA) genes, pseudogenes, and TEs. Descriptive annotation statistics are summarized in Table 2. While many of these annotation features in *T. arvense* were similar to those found in other plant species, we observed several unique patterns, which we will describe in detail below. ncRNA annotations were inferred from sequence motifs (tRNA, rRNA, snoRNA) or from sequencing data (siRNA, miRNA). We predicted clusters of both 5S rRNA and tandem repeat units of 18S and 28S rRNA with RNAmmer (Lagesen *et al.*, 2007), often in relative proximity to loci identified with Tandem Repeats Finder v4.09.1 (Benson, 1999) and putatively associated with centromeric repeat motifs (not shown). Of the largest seven scaffolds, only scaffolds 4 and 7 carried no such annotations. Notably, several large clusters of 5S rRNA genes were interspersed throughout the pericentromeric region of scaffold 1, whereas the remaining four scaffolds contained 18S and 28S rRNA gene annotations. Finally, we identified 243 homologs from 114 snoRNA families.

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

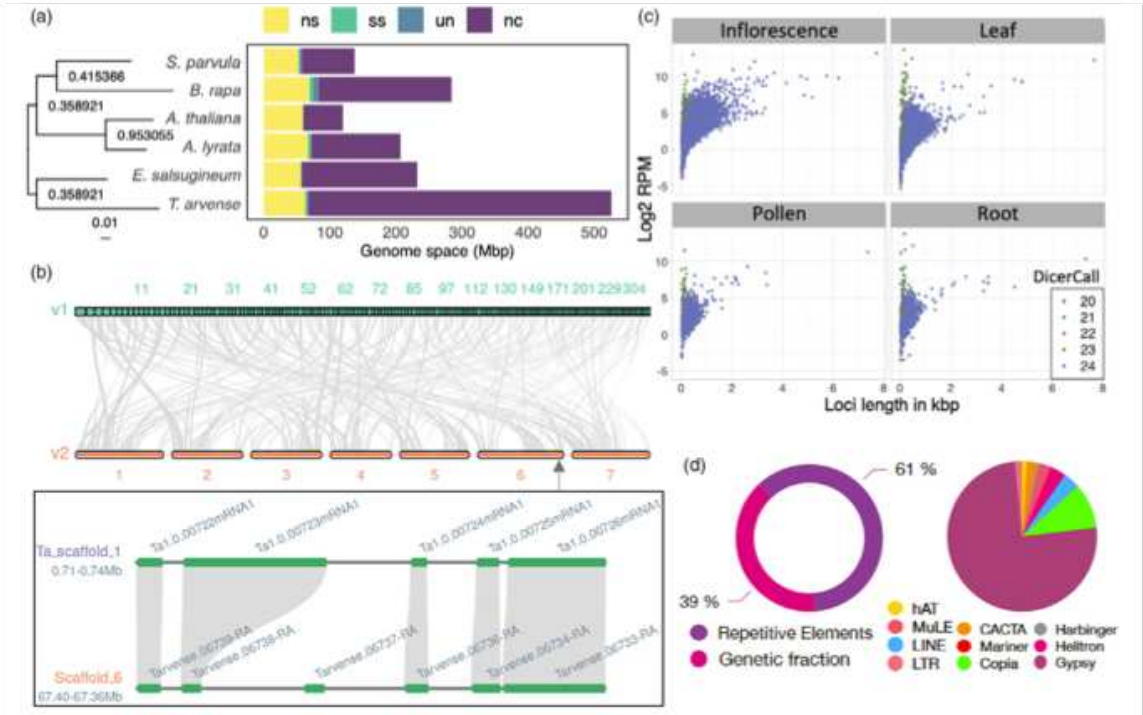


Figure 3 Feature annotations within *T. arvense* var MN106-Ref. (a) Rooted species tree inferred from all genes, denoting node support and branch length in substitutions per site, and horizontal stacked bar chart comparing the genetic fraction in pennycress with other *Brassicaceae* sp. (ns = nonspecific orthologs, ss = species-specific orthologs, un = unclassified genes, nc = non-coding/intergenic fraction). (b) Comparison of gene macrosynteny between v1 and v2 of the genome, and a microsynteny example of genes *MYB29* and *MYB76*, which are resolved in the v2 annotation. (c) Small RNA biogenesis locus length and expression values in each of four tissues. (d) Overall repetitive content in the genome as discovered by RepeatMasker2, and relative abundance of TEs within the fraction of repetitive elements.

sRNA annotation

We identified 19 386 siRNA loci. More than 98% of these loci corresponded to heterochromatic 23- to 24-nt siRNA loci, with only 196 producing 20- to 22-nt siRNAs. The sRNA loci were expressed unevenly across tissues, as inferred from prediction with data from different tissues. Only 2938 loci were shared across all four tissues studied (rosette leaves, roots, inflorescences and pollen). Inflorescences were the major contributor with 6728 private loci. Despite these differences between tissues, we observed similar overall patterns in terms of locus length, expression (Figure 3c) and complexity (Figure S7).

Altogether, sRNA loci accounted for ~8 Mbp or ~1.5% of the assembled genome. Of the seven largest scaffolds, where the majority of genes are located, the total coverage of siRNA loci ranged between 1.5% and 2% and the loci appeared to be preferentially concentrated at the boundary between TEs and the protein-coding gene fraction of the genome. To further explore this, we partitioned the seven largest scaffolds into gene-enriched and gene-depleted regions, based on a median of 14 genes per Mbp and a mean of 54.2 genes per Mbp. We defined gene-enriched loci as those above and gene-depleted loci as those below the mean. At the chromosomal level, sRNA loci correlated with gene-enriched regions and were scarce in regions with high TE content. This trend is in contrast to that observed in *A. thaliana* (Hardcastle *et al.*, 2018) but resembles what has been observed, for example, in maize (He *et al.*, 2013) and tomato (Tomato Genome Consortium, 2012).

Phased secondary siRNAs (phasiRNAs) are a class of secondary sRNAs that, due to the way they are processed, produce a distinct periodical pattern of accumulation (Axtell, 2013b). In the *T. arvense* genome, we observed 139 loci with such phased patterns. In contrast to the general notion that phasiRNAs are typically 21 nt long (Lunardon *et al.*, 2020), we found 24-nt siRNAs to be dominant in 133 of these loci.

MicroRNAs

MicroRNA (miRNA)-encoding genes were predicted using a combination of ShortStack and manual curation (see Methods). We identified 72 miRNA-producing loci, with 53 that were already known from other species, and 19 appeared to be species-specific. Most of the identified families were produced from only one or two loci, with miR156 and miR166 being produced by the most loci, with eight and five family members, respectively. A total of 21 out of 25 families in *T. arvense* are found in other rosids, and three (miR161, miR157 and miR165) only in other Brassicaceae. One family, miR817, is also present in rice. There is a strong preference for 5′-U at the start of both unique and conserved miRNAs (Figure S8), in line with previous reports (Voinnet, 2009). The expression level of both conserved

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

Table 2 Summary of feature annotations in comparison with the original version T_arvense_v1

Type	T_arvense_v1	T_arvense_v2	diff.
(A) Protein-coding genes			
Total number of loci	27 390	27 128	-262
Total number of unique loci	4780	5034	+254
Total number of transcript isoforms	–	30 650	+30 650
Number of matching loci with changes in CDS	–	–	+14 102
Number of matching loci with changes in UTR(s)	–	–	+22 559
Loci containing one or more PFAM domain	–	21 171	+21 171
Loci annotated with one or more GO term	–	13 074	+13 074
(B) Non-coding genes			
tRNA	–	1148	+1148
rRNA clusters (<25 kbp)	–	63	+63
snoRNA	–	243	+243
Small interfering RNA (siRNA)	–	19 373	+19 373
MicroRNA (miRNA)	–	72	+72
(C) Other gene types			
Pseudogenes (set II Ψs)	–	44 490	+44 490
Transposable element genes	–	423 251	+423 251

and novel miRNA families was compared between tissues, showing that the ten most highly expressed across all tissues are conserved families, whereas novel miRNA demonstrates a marginal tendency to be more lowly expressed or with potential for differential expression (Figure S9).

sRNA loci

When we overlaid the sRNA loci with our annotated genomic features, most sRNAs localized to the intergenic space, but a substantial fraction, especially 20- to 22-nt sRNAs, were produced from intronic sequences (Figure S10a). Helitrons make up only 1.5% of the genome space, yet more than 5% of sRNA biogenesis loci overlap with this type of TE. Most sRNA loci (93.0%) fell within 1.5 kbp of annotated genes or TEs (Figure S10b,c). As expected, 23- to 24-nt sRNAs were more frequently associated with TEs, whereas 20- to 22-nt sRNAs were more often produced by coding genes (Axtell, 2013a).

Pseudogenes

In accordance with the MAKER-P protocol, pseudogenes (Ψ) were predicted in intergenic DNA with the ShiuLab pseudogene pipeline (Zou *et al.*, 2009). A total of 44 490 set II pseudogenes were annotated, exceeding those in *A. thaliana* (~3700) or rice (~7900) by one order of magnitude. We identified 35 818 pseudogenes overlapping with TEs, and 8672 pseudogenes that were either concentrated in intergenic space or more towards the protein-coding gene complement of the genome, and thus perhaps less likely to have arisen from retrotransposition. Approximately 59.2% of these contained neither a non-sense nor a frameshift mutation, indicating either (i) that the regulatory sequences of the pseudogenes were silenced first, (ii) a pseudoexon that may be linked to another non-functional exon, or (iii) a possible undiscovered gene.

Transposable elements

In total, we identified 423 251 TEs belonging to 10 superfamilies and covering ~61% of the genome (Figure 3d). Retrotransposons (75% of all TEs are Gypsy elements; 10% Copia; 4% LINE) by far outnumbered DNA transposons (3% Helitrons; 1% hAT; 2% CACTA; 1% Pif-Harbinger; 2% MuLE). A detailed breakdown of repeats is shown in Table S4. As the most abundant retrotransposon superfamily, Gypsy elements accounted for 46% of the total genome space, which is consistent with a high abundance observed in the pericentromeric heterochromatin of *E. salsugineum*, where centromere expansion is thought to have been caused by Gypsy proliferation (Zhang *et al.*, 2020). In addition, we identified 359 protein-coding genes located fully within TE bodies that could represent Pack-TYPE elements and contribute to gene shuffling (Catoni *et al.*, 2018). Among these elements, 153 were intersecting with mutator-like elements, suggesting they correspond to Pack-MULE loci. TEs were located primarily in low gene density regions, while the fraction of TE-contained genes was randomly distributed.

Expression atlas

With cDNA sequences from 11 different tissues or developmental stages, we could annotate tissue-specific expression patterns. The complete expression atlas is provided in Data S1. We evaluated the relative extent of tissue-specific gene expression using the Tau (τ) algorithm (Yanai *et al.*, 2005), from the normalized trimmed mean of *M*-value (TMM) counts in all tissues (Robinson and Oshlack, 2010). To preclude potential biases caused by substantial differences in library size, we excluded low-coverage samples from mature seeds and old green siliques. In total, 4045 genes had high or even complete tissue specificity (τ = 0.8–1.0), while 5938 genes had intermediate specificity (0.2–0.8) and 6107 had no or low specificity (0–0.2); the remaining genes were ignored due to missing data. The relative breakdown of each specificity fraction by tissue type is shown in Figure 4a, with 'roots', 'green seeds' and 'inflorescences' representing the tissues with the greatest proportion of high or complete specificity genes. The relative $\log_2$(TMM) expression values of the top 30 most highly expressed genes in each tissue, given a high or complete specificity score, are plotted in Figure 4b with respect to the overall mean expression per gene across all included tissues. These include, for example, genes with homology to *EXTENSIN 2* (*EXT2; A. thaliana*) in 'roots', *CRUCIFERIN* (*BnC1; B. napus*) in 'green seeds', and *PECTINESTERASE INHIBITOR 1* (*PMEI1; A. thaliana*) in 'inflorescences' and 'open flowers' (Data S2).

DNA methylation

Cytosine methylation (also commonly referred to as DNA methylation) is a prevalent epigenetic mark in plant genomes and is often associated with heterochromatin and transcriptional inactivation of TEs and promoters, but also with higher and more stable expression when present in gene bodies (Zhang *et al.*, 2018). In plants, DNA methylation occurs in three cytosine contexts, CG, CHG and CHH (where H is any base but G), with the combined presence of CG, CHG and CHH methylation usually indicative of heterochromatin formation and TE silencing, while gene body methylation consists only of CG methylation (Bewick and Schmitz, 2017). In the light of the high TE density in *T. arvense*, we analysed genome-wide DNA methylation by whole-genome bisulphite sequencing (WGBS) in shoots and roots of 2-week-old seedlings. Genome-wide, 70% of cytosines were

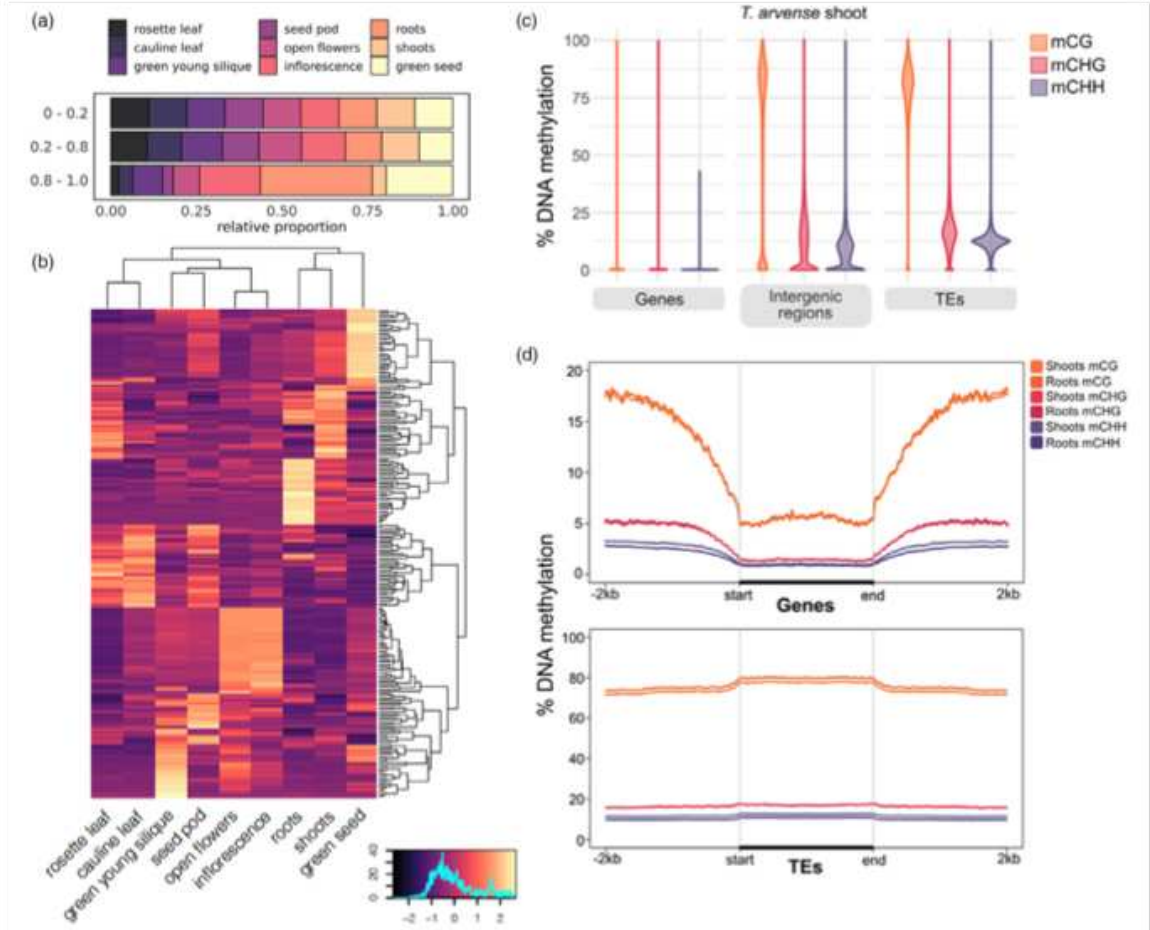


Figure 4 Regulatory dynamics in pennycress. (a) Relative fraction of genes in each tissue for low (0–0.2), intermediate (0.2–0.8) and high/absolute specificity (0.8–1.0) subsets. (b) Log_2(TMM) expression values of the top 30 most highly expressed genes in each tissue, relative to the mean across all tissues, from the subset of genes with a high/absolute tau specificity score. (c) Distribution of average DNA methylation for different genomic features, by cytosine sequence context. (d) DNA methylation along genes (top) and TEs (bottom), including a 2-kb flanking sequence upstream and downstream. DNA methylation was averaged in non-overlapping 25-bp windows.

methylated in the CG context, 47% in the CHG context and 33% in the CHH context. In line with findings in other Brassicaceae, methylation at CG sites was consistently higher than at CHG and CHH (Figure 1a; Figure S11). When we compared the WGBS data against the genome annotation, high levels of DNA methylation (mostly ^{m}CG) colocalized with regions of dispersed repeats and TEs in the centre of the chromosomes. Conversely, methylation was depleted in gene-rich regions (Figure 1a,b). In line with this, DNA methylation was consistently high along TEs, particularly in the CG context (Figure 4c). In contrast to *E. salsugineum* (Bewick *et al.*, 2016; Niederhuth *et al.*, 2016), DNA methylation dropped only slightly in regions flanking TEs, which might be related to the overall dense TE content in *T. arvense*.

In contrast to TE and promoter methylation, gene body methylation (gbM) is generally associated with medium-to-high gene expression levels (Zhang *et al.*, 2006; Zilberman *et al.*, 2006). gbM occurs in ~30% of protein-coding genes in *A. thaliana*, with DNA methylation increasing towards the 3′-end of the gene (Zhang *et al.*, 2006). The *T. arvense* relative *E. salsugineum* lacks gbM (Bewick *et al.*, 2016; Niederhuth *et al.*, 2016). gbM was also largely absent in *T. arvense* (Figure 4d), suggesting that gbM was lost at the base of this clade.

Applications towards crop improvement

Genetic variation in a pennycress collection

Knowledge of genetic diversity within wild populations is an essential process for improvement and domestication of new crop species. We analysed a geographically broad sample of forty accessions (Figure S12) using whole-genome resequencing to characterize population structure and variation in germplasm available for breeding. We identified a total of 13 224 528 variants with QD value of ≥2000. Of these, 12 277 823 (92.8%) were SNPs, 426 115 (3.2%) were insertions, and 520 590 (3.9%) were deletions relative to the reference genome. Across all variants, 661 156 (2.9%) were in exons, with 340 132

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

synonymous, 314 075 nonsynonymous and 6949 non-sense changes. STRUCTURE analysis of both indel and SNP data sets resulted in optimal models of $k = 3$ populations (Figure S13). Both data sets assigned the three lines of Armenian descent, which were highly distinct and had the largest genetic distance to the other accessions, to a single discrete population with limited to no gene flow to the other populations. These results are consistent with previous reports in pennycress (Frels *et al.*, 2019) and were further supported by whole-genome dendrograms (Figure 5a). We also calculated linkage disequilibrium (LD) among 2 518 379 genome-wide markers and chromosome-specific markers using TASSEL v5.2.75 (Bradbury *et al.*, 2007) with a sliding window of 40 markers. The *r*-squared values were plotted against the physical distance with a LOESS curve fitted to the data to show LD decay (Figure S14). Genome-wide, LD decayed to an *r*-squared value (r^2) of 0.2 over 6.2 kbp (Hill and Weir, 1988), which is comparable to LD decay reported in related Brassica species at $r^2 = 0.3$, including *B. rapa* (2.1 kbp) (Wu *et al.*, 2019) and *B. napus* (12.1 kbp) (Lu *et al.*, 2019).

Gene structure variation in pennycress accessions

The natural variation present in germplasm is an important source of alleles to facilitate breeding efforts and presents an opportunity to understand the evolution of gene families and adaptation within a species. To understand these in a more targeted approach, we sequenced on the PacBio Sequel platform the transcriptomes of two accessions, MN108 and Spring32-10, that are amenable to transformation and gene editing (McGinn *et al.*, 2019), using RNA from leaves, roots, seeds, flowers and siliques. We constructed *de novo* reference transcriptomes using the Isoseq3 pipeline, resulting in 25 296 and 26 571 accession-specific isoforms for MN108 and Spring32-10, respectively. These transcriptomes were then polished using the raw reads and processed through the SQANTI3 pipeline (Tardaguila *et al.*, 2018) to characterize the genes and isoforms identified in each of the accessions. We identified 212 of 220 unique genes and 3780 of 3857 unique isoforms for MN108 and Spring32-10 respectively compared with the new reference. Transcripts mapping to the known reference denoted by 'Full Splice Match' (FSM) and 'Incomplete Splice Match' (ISM) accounted for 28.7% and 30.6% of all transcript models in MN108 and Spring32-10, respectively (Figure 5b, c). Transcripts of the antisense, intergenic and genic intron categories collectively accounted for a total of 12.0% (MN108) and 11.2% (Spring32-10). About ~15% of all identified transcripts were novel isoforms when compared to the reference transcriptome for T_arvense_v2.

Mapping a pale seedling phenotype

From a segregating population with a high oleic pennycress (*fae-1/rod1-1*) background (Chopra *et al.*, 2020b), we identified pale seedling lines (Figure 5d). This phenotype segregated in a Mendelian fashion. To determine the genetic control for this phenotype, we separately pooled genomic DNA from 20 wild-type and 20 pale plants. We processed sequence data obtained from each of these pools through the MutMap pipeline (Sugihara *et al.*, 2020) and discovered a putative genomic interval (63.85–63.95 Mbp) on scaffold 6 linked to the pale phenotype. SnpEff (Cingolani *et al.*, 2012) identified polymorphisms that might have deleterious effects on function of genes in this region (Table S5). The most obvious candidate is *MEX1*, encoding a maltose transporter located in the chloroplast, knockout of which causes a pale seedling phenotype in *A. thaliana* (Niittylä *et al.*, 2004).

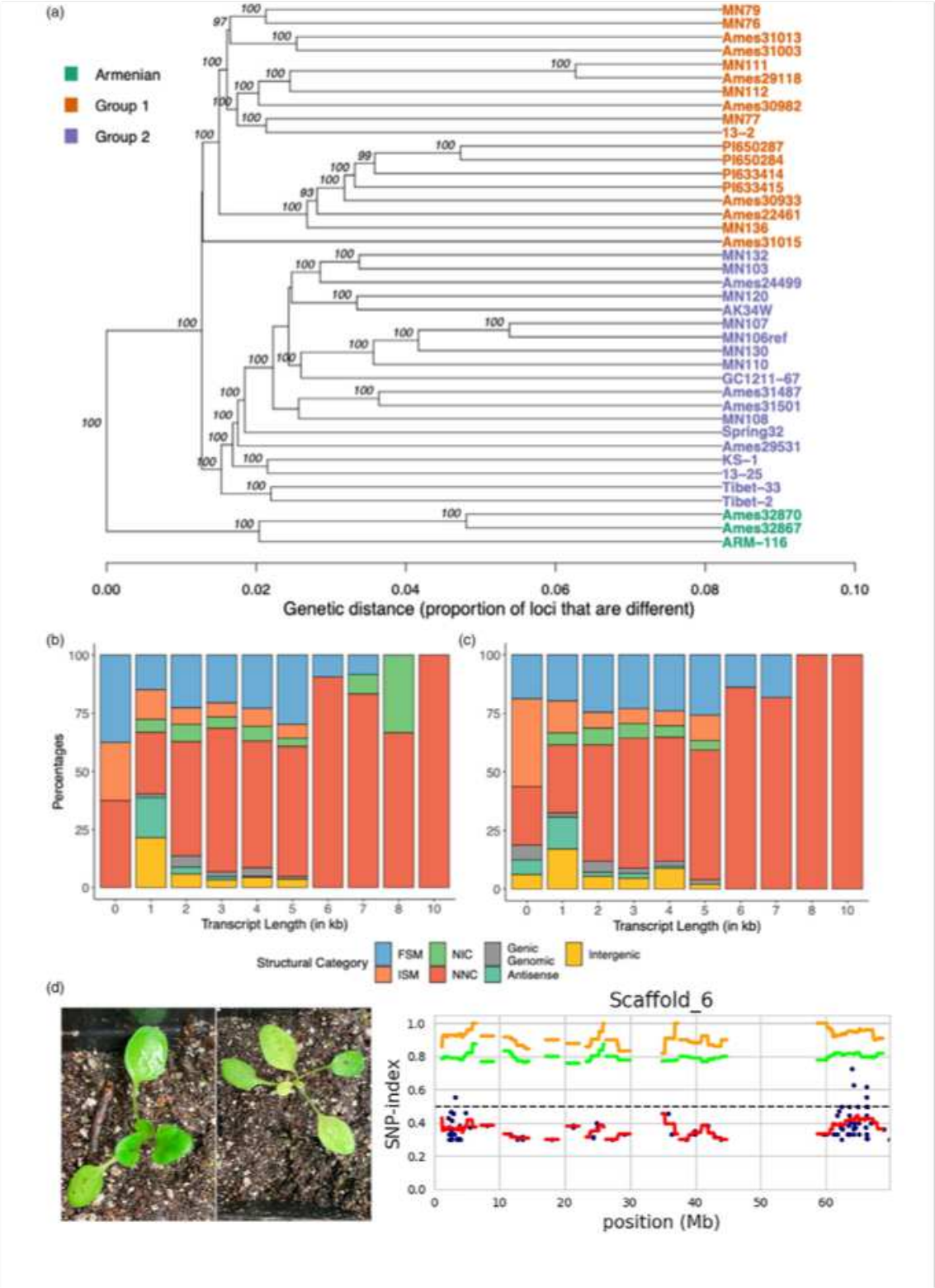
Discussion

In this study, we report a high-quality reference genome assembly and annotation for *T. arvense* (var. MN106-Ref), a newly domesticated oilseed crop for the cooler climates of the world. The improved genome assembly, containing seven chromosome-level scaffolds, revealed two main features: a landscape characterized by a large repetitive fraction populated with TEs and pseudogenic loci in pericentromeric regions, and a gene complement similar in size to other Brassicaceae and densely concentrated towards the telomeres (Figure 1). Previous annotations were enriched with additional gene models for protein-coding loci, and now include non-coding genes for tRNAs, rRNAs, snoRNAs, siRNAs and miRNAs, alongside predicted pseudogenes and TEs (Table 2). These newly improved assembly features will allow for efficient combining of traits and help accelerate future breeding as it would provide knowledge about the gene localization and the linkage of genes of interest. For example, the improved genome assembly has revealed that multiple domestication syndrome genes (*ALKENYL HYDROXALKYL PRODUCING 2-like, TRANSPARENT TESTA 8, EARLY FLOWERING 6*) (Figure S1) are located on a single chromosome.

Improved genomic resources can facilitate general understanding of plant biology and evolutionary biology while aiding plant breeding and crop improvement (Scheben *et al.*, 2016). For example, pennycress and *Arabidopsis* share many key features that made *Arabidopsis* the most widely studied model plant system (Meinke *et al.*, 1998). The use of *Arabidopsis* for translational research and for identifying potential gene targets in *T. arvense* is possible and has been extensively validated (Chopra *et al.*, 2018, 2020a, 2020b; Jarvis *et al.*, 2021; McGinn *et al.*, 2019). Previous studies have suggested that over a thousand unique genes in *T. arvense* are represented by multiple genes in *Arabidopsis* and vice versa. Our comparative genomics by way of synteny with *E. salsugineum* (Yang *et al.*, 2013) revealed a high level of agreement, particularly between the protein-coding fraction of the genome, represented as conserved blocks in the largest seven scaffolds relative to the ancestral karyotype in Brassicaceae (Murat *et al.*, 2015; Figure 2). The detailed description of gene synteny between *T. arvense* and other Brassicaceae provides insights into the evolutionary relevance of *T. arvense* within lineage II of Brassicaceae. In addition, the difference in genome size between *T. arvense* and other species, despite the reduced level of gene duplication and the 1:1 gene relationship, can be explained by the large repetitive fractions present throughout both the centromeric and pericentromeric regions. In the absence of whole-genome duplication events, these repetitive fractions indicate that the increased

Figure 5 (a) Dendrogram representing the forty wild accessions in our study showing three distinct subpopulations, inferred from STRUCTURE analysis (Figure S13). (b,c) Variation of transcript isoforms for MN108 (b) and Spring32-10 (c) accessions based on SQANTI3 analysis. (d) A pale phenotype segregating in an improved pennycress line (*fae-1-1/rod1-1*) was analysed with a modified bulked-segregant analysis, and the QTL region associated with this phenotype was mapped using the MutMap approach.

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20



© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

genome size may be a consequence of active TE expansion. This is therefore suggestive of a mechanism by which deleterious retrotransposon insertions must be mitigated in *T. arvense*. This could be explained by the high proportion of *Gypsy* retrotransposons in this species, usually located in heterochromatic regions, or by integration site selection (Sultana *et al*., 2017), or otherwise by silencing by small RNA activity and/or DNA methylation (Bucher *et al*., 2012; Sigman and Slotkin, 2016). Given the relatively high error rate of PacBio CLR reads (~10% before correction) with respect to circular consensus sequencing (CCS), the repetitive fraction would also help to explain the initial overestimation of the assembly size as a result of duplicated contigs. We also detected several loci with highly overrepresented read coverage indicative of repeat collapsing during the assembly process, often intersecting with 5S, 18S and 28S rRNA annotations. Such regions are difficult even for current long read technologies due to the large size of the tandem repeat units.

With the availability of improved genomic resources, increasing interest has turned towards understanding tissue-specific gene regulation to reduce pleiotropic effects upon direct targeting of genes during crop improvement. In this study, we have generated a resource using mRNA-seq, sRNA-seq and WGBS to gain insights into genes and their associated regulatory landscape. These data sets help elucidate the extent of tissue specificity and provide useful information for gene modification targets. For example, fatty acid desaturase 2 gene (FAD2; *Ta12495* – T_arvense_v1) is involved in the oil biosynthesis pathway and is expressed in many different tissues analysed in this study (Data S1). *FAD2* gene knockout should result in higher levels of oleic acid in the seed oil and provide an opportunity for pennycress oil to be used in food applications. It has been observed, however, that knockout mutants in pennycress display delayed growth and reduced seed yields in spring types (Jarvis *et al*., 2021), and reduced winter survival in the winter types (Chopra *et al*., 2019), as a purported consequence of its broad expression profile. Similarly, genes such as *AOP2-LIKE* (*Tarvense_05380* – T_arvense_v2) have been targeted to reduce glucosinolates in pennycress seed meal for food and animal feed applications (Chopra *et al*., 2020b). However, *AOP2-LIKE*, too, is expressed in many tissues during development, which might explain why knockout plants with reduced glucosinolate content are reportedly more susceptible to insect herbivores such as flea beetles feeding on rosette leaves and root tissues (Marks *et al*., 2021). Our tissue-specific expression data suggest that, to overcome this challenge, one could alternatively target genes such as *Glucosinolate Transporter 1* (GTR1; *Tarvense_14683*), which is expressed specifically in reproductive tissues (Data S1). This might achieve the desired reductions of seed glucosinolates while avoiding developmental defects. Such approaches have been effectively used in *Arabidopsis* and many *Brassica* species (Andersen and Halkier, 2014; Nour-Eldin *et al*., 2012).

Finally, the forty resequenced accessions described here provide a rich source of variants that reflect the genetic diversity and population structure of the species in the collection (Figure 5a). Further evaluations of transcriptome sequences showed ample variation in the transcripts from two separate lines – MN108 and Spring32-10 – that are highly amenable to transformation and highlighted the potential for developing pan-genomes in the future. These genomic resources will facilitate genetic mapping studies in pennycress in both natural populations and mutant panels. We have identified genomic regions associated with a pale leaf mutant in pennycress seedlings using a modified BSA-Seq approach in this study (Figure 5d).

Over the last few years, significant efforts have been made towards the discovery of crucial traits and translational research in pennycress, centring on MN106-Ref and the gene space information generated by Dorn *et al*. (2013) and Dorn *et al*. (2015). In this study, we continued to generate genomic tools for this accession, with improved contiguity and high-quality annotations to make *T. arvense* var. MN106-Ref more accessible as a field-based model species for genetics and epigenetics studies and to provide tools for this new and extremely hardy winter annual cash cover crop. However, the assembly of additional accessions can only help to further enrich the resources available for the study of pennycress. In parallel to this study, a Chinese accession of *T. arvense* (YUN_Tarv_1.0) was assembled using Oxford Nanopore, Illumina HiSeq and Hi-C sequencing (Geng *et al*., 2021). This timely availability of an additional frame of reference opens the door to a pan-genomic approach in evolutionary research and allows for the better characterization of structural variants moving forward. Furthermore, the use of different sequencing technologies and assembly software provides an additional avenue to correct misassemblies and base calling errors in either case. The overall longer contigs assembled with PacBio CLR, for example, and the consideration of various genetic map data in addition to Hi-C provides a greater resolution of scaffolds particularly throughout the centromere and pericentromeric regions (Figure S15). The reduced error rate of PacBio CCS (used for polishing) is also reflected in the overall k-mer content, which is measured with a two-order magnitude higher consensus quality over scaffolds representing chromosomes and ~99% overall completeness for T_arvense_v2 (Tables S6-S8), indicative of high-quality, error-free sequences more appropriate for variant calling, for instance. Geng *et al*. (2021) also reported WGS analysis on forty Chinese accessions and reported an LD decay of 150 kbp at an *r*-squared value (r^2) of 0.6, which is considerably higher than the values determined on the forty accessions in this study, as well as those reported for related *Brassica* species (Lu *et al*., 2019; Wu *et al*., 2019). We believe the combination of resources will allow us to investigate the differences that might exist between accessions originating from different geographic locations around the world and help provide further insight into structural variations and evolutionary dynamics.

In conclusion, the T_arvense_v2 assembly offers new insights into the genome structure of this species and of lineage II of Brassicaceae more generally, and it provides new information and resources relevant for comparative genomic studies. The tools presented here provide a solid foundation for future studies in an alternative model species and an emerging crop.

Methods

Seeds for the reference genome development

Seeds from a small natural population of *T. arvense* L. were collected near Coates, MN by Dr. Wyse, and the accession number MN106 was assigned to this population. We propagated a single plant for ten generations from this population, and we refer to this line as MN106-Ref.

Sample collection, library preparation and DNA sequencing for assembly

PacBio CLR library

Plants were cultivated, sampled and prepared at the Max Planck Institute for Developmental Biology (Tübingen, Germany). Plant

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

seeds were stratified in the dark at 4 °C for 4–6 day prior to planting on soil. Samples were collected from young rosette leaves of *T. arvense* var. MN106-Ref seedlings, cultivated for 2 weeks under growth chamber conditions of 16–23 °C, 65% relative humidity and a light/dark photoperiod of 16 h:8 h under 110–140 μmol/m^2/s light. High molecular weight (HMW) DNA was obtained following nucleus isolation and DNA extraction with the Circulomics Nanobind Plant Nuclei Big DNA Kit according to the protocol described in Workman *et al.* (2018) and (Workman *et al.*, 2019). A total of 11 extractions from 1.5–2 g frozen leaves each were processed in that way, yielding a pooled sample with a total of 12 μg of DNA by Qubit® 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA) estimation, and high DNA purity with a mean absorbance ratio of 1.81 at 260/280 nm absorbance and 2.00 at 260/230 nm absorbance, as measured by NanoDrop 2000/2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA). HMW DNA was sheared by one pass through a 26G needle using a 1-mL syringe, resulting in an 85-kb peak size sample as estimated by FEMTO Pulse Analyzer (Agilent Technologies, Santa Clara, CA). A large insert gDNA library for PacBio Sequel II CLR sequencing was prepared using the SMRTbell® Express Template Preparation Kit 2.0. The library was size-selected for >30 kb using BluePippin with a 0.75% agarose cassette (Sage Science) and loaded into one Sequel II SMRT cell at a 32 pM concentration. This yielded a genome-wide sequencing depth of approximately 476X over ~6.9 million polymerase reads with a subread N50 of ~38 kbp.

PacBio CCS library

MN106-Ref plants were grown in growth chambers at the University of Minnesota. Individual plants were grown to form large rosettes for isolating DNA. Approximately 25 g of tissue was harvested and submitted to Intact Genomics (Saint Louis, MO) for high molecular weight DNA extraction. This yielded a pooled sample with a total of 269 ng of DNA by Qubit® (Thermo Fisher Scientific, Waltham, MA) estimation, and high DNA purity with a mean absorbance ratio of 1.87 at 260/280 nm and 2.37 at 260/230 nm, as measured by Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA). To further clean up the high molecular weight DNA, we used Salt:Chloroform Wash protocol recommended by PacBio. This yielded a total of 12.1 ng/uL of high-quality DNA for library preparation. A large insert gDNA library was prepared, and 15 kb High Pass Size Selection on Pippin HT was performed at the University of Minnesota Genomics Center (Minneapolis, MN). These libraries were sequenced on 4 SMRT cells using PacBio Sequel II (Pacific Biosciences, Menlo Park).

Bionano library

High molecular weight DNA was isolated from young leaves and nicking endonuclease – BspQI was chosen to label high-quality HMW DNA molecules. The nicked DNA molecules were then stained as previously described (Lam *et al.*, 2012). The stained and labelled DNA samples were loaded onto the NanoChannel array (Bionano Genomics, San Diego, CA) and automatically imaged by the Irys system (Bionano Genomics, San Diego, CA).

Hi-C library

The MN106-Ref plant tissue used for PacBio CCS was submitted to Phase Genomics (San Diego, CA). The Hi-C library was prepared following the proximo Hi-C plant protocol (Phase Genomics, San Diego, CA), and the libraries were sequenced to 116X depth on an Illumina platform with the paired-end mode and read length of 150 bp.

Illumina PCR-free library

Libraries for PCR-free short read sequencing were prepared from MN106-Ref genomic DNA using the TruSeq DNA PCR-Free Low Throughput Library Prep Kit (Illumina, San Diego, CA) in combination with TruSeq DNA Single Indexes Set A (Illumina, San Diego, CA) according to the manufacturer's protocol. We prepared two libraries, with average insert sizes of 350 bp and 550 bp, respectively. Samples were sequenced to 125X depth (~66 Gb) on an Illumina HiSeq 2500 (Illumina, San Diego, CA) instrument with 125-bp paired-end reads.

Genome assembly and construction of chromosome-level scaffolds

The initial assembly was performed using Canu v1.9 (Koren *et al.*, 2017) with default options, aside from cluster runtime configuration and the settings `corOutCoverage=50`, `minReadLength=5000`, `minOverlapLength=4000`, `correctedErrorRate=0.04` and `genomeSize=539m`, which were selected based on the characteristics of the library. Canu performs consensus-based read correction and trimming, resulting in a curated set of reads that were taken forward for assembly (Figure S16).

The resulting assembly overestimated the genome size by approximately 53% (Table S2), which we surmised was likely due to uncorrected sequencing errors in the remaining fraction of reads, in which Canu was able to assemble into independent, duplicated contigs. Analysis of single-copy orthologs from the *Eudicotyledons odb10* database with BUSCO v3.0.2 (Simão *et al.*, 2015) revealed a high completeness of 98.4% and a duplication level of 23.6% (Table S6). Subsequent alignment of the reads to the assembly using minimap2 v2.17 (Li, 2018) and purge_dups v1.0.1 (Guan *et al.*, 2020) presented bimodal peaks in the read depth distribution, indicative of a large duplicated fraction within the assembly (Figure S17). As efforts to collapse this duplicated fraction using assembly parameters were unsuccessful, and purge_dups is intended to correct duplication arising from heterozygosity (which does not apply in *T. arvense*), the fraction was reduced by manual curation instead. Contigs starting from the left-hand side of the read depth distribution were consecutively removed until reaching an approximation of the estimated genome size, with any contigs containing non-duplicated predicted BUSCO genes kept preferentially in favour of discarding the next contig with lower read depth in the series.

The deduplicated assembly from Canu was polished with the PacBio Sequel II HiFi CCS reads using two iterations of RACON v1.4.3 (Vaser *et al.*, 2017), prior to repeat reassembly. Bionano maps were used to build *de novo* scaffolds using the polished assembly; hybrid scaffolds were generated using the *de novo* Bionano maps and the assembly (https://bionanogenomics.com/support-page/data-analysis-documentation/). To further resolve repetitive regions and improve assembly contiguity, the bionano-scaffolded assembly was integrated into the HERA pipeline (Du and Liang, 2019). The Hi-C data were aligned with bwa-mem v0.7.17 (Li and Durbin, 2009), PCR duplicates were marked with picard tools v1.83 (http://broadinstitute.github.io/picard), and the quality was assessed with the `hic_qc.py` tool of Phase Genomics (https://github.com/phasegenomics/hic_qc). The assembly was then scaffolded with the Hi-C alignments using SALSA v2.2 (Ghurye *et al.*, 2017) and subsequently polished with

the PCR-free Illumina data using two iterations of PILON v1.23 (Walker *et al.*, 2014). The final assembly was the result of a meta-assembly with quickmerge v0.3 (Chakraborty *et al.*, 2016), which combined the current assembly with an earlier draft version assembled using Canu 1.8 (Koren *et al.*, 2017) directly from the PacBio CCS reads and polished only with the Illumina PCR-free short-reads, following an almost identical workflow, in order to help address the possibility of misassembly arising from technical sources and improve overall contiguity. This resulting assembly was evaluated with BUSCO (Simão *et al.*, 2015) and QUAST v5.0.2 (Gurevich *et al.*, 2013). Intermediate assembly statistics are given in comparison with (i) immediately after Canu, and (ii) the final version after rescaffolding (Table S2).

Genome size estimation using flow cytometry and k-mer-based approach

The nuclei of field pennycress line MN106-Ref, *Arabidopsis thaliana*, maize and tomato were stained with propidium iodide, and fluorescent signals were captured using a Becton-Dickinson FACSCanto flow cytometer (https://www.bdbiosciences.com/). DNA content for all four species that corresponded to $G_{0/1}$ nuclei is listed in Table S1. The genome size of Arabidopsis is 135 Mb, and therefore, the genome size of pennycress was calculated to be 501 ± 33 Mb. Using the Illumina HiSeq2500 platform, we obtained ~100× PCR-free reads, which were used for subsequent K-mer analysis using Jellyfish (Marçais and Kingsford, 2011). The 101-mer frequency distribution curve exhibited a peak at 22 k-mer, and analysis showed that the total number of K-mers was 11 403 836 319. Using the formula of genome size = total K-mer number/peak depth, the genome size of this sequencing sample was estimated to be 518 356 196 bp. Similarly, the single-copy content of the genome was estimated to reach 79%. Using both methods of genome size estimation, we found the pennycress genome ranged from 459 to 540 Mb.

Development of genetic maps for rescaffolding

To improve the contiguity and correct misassemblies, we developed two genetic linkage maps using F_2 populations. The first linkage map was derived from a cross between a wild Minnesota accession 'MN106-Ref' and a genetically distant Armenian accession 'Ames32867'. The resulting F_1 plants were allowed to self-fertilize, and seeds from a single plant were collected and propagated to the F_2 generation. Approximately 500 mg fresh tissue was collected from 94 individuals in the F_2 population. The tissue was desiccated using silica beads and pulverized using a TissueLyser. DNA was isolated with the BioSprint DNA Plant Kit (Qiagen, Valencia, CA). The F_2 population along with the two parental genotypes was genotyped with genotyping by sequencing at the University of Minnesota Genomics Center (Minneapolis, MN). Each sample was digested with the *BtgI_BgLII* restriction enzyme combination, barcoded and sequenced on the Illumina NovaSeq S1 (single-end 101 bp) yielding 1 237 890 mean reads per sample. The raw reads were demultiplexed based on the barcode information and aligned to the most recent iteration of the pennycress genome using bwa. Sequence-aligned files were processed through samtools v1.9 (Li *et al.*, 2009) and picard tools to sort the files and remove group identifiers. Variants were called using GATK HaplotypeCaller v3.3.0. SNPs identified among these 94 lines were used for the development of genetic maps. The second linkage map was derived from a cross between MN106-Ref and a mutant line '2019-M2-111'. To identify the variant alleles in 2019-M2-111, we performed whole-genome resequencing using paired-end reads on the Illumina Platform. SNPs were identified using a similar approach as described above. Sixty-seven SNP markers were designed using the biallelic information from resequence data. DNA was extracted from 48 samples from the mutant F_2 population using the Sigma-Aldrich ready extract method, allele-specific and flanking primers synthesized from IDT (Iowa, USA) for each of the alleles were mixed (Data S3), and genotyping was performed using the methods described in Chopra *et al.* (2020a).

A total of 35 436 SNPs were identified among the population used for the first linkage map, SNP sites were selected with no-missing data, QD > 1000, and the segregation of the markers was 1:2:1. A total of 743 high-quality SNPs were retained for further analysis. A genetic map for the population was constructed using JoinMap 5 (Stam, 1993). Only biallelic SNPs were used in the analysis, and genetic maps were constructed with regression mapping based on default parameters of recombination frequency of <0.4 with only the first two steps. The Kosambi mapping function was chosen for map distance estimation, and the Ripple function was deployed to confirm marker order within each of the seven linkage groups. A total of 319 markers were mapped to seven linkage groups (Data S4). Similarly, 67 markers were genotyped on 48 individuals from the second population of linkage and 52 markers were mapped to six linkage groups (Data S5). Both of these linkage maps were used for reordering and correcting the scaffolds as described below.

Rescaffolding

Initial exploration regarding gene and TE distributions and methylation patterns pointed to potential misassemblies in the assembled genome. Further investigation by way of synteny comparison with a closely related species, *Eutrema salsugineum* (Yang *et al.*, 2013), revealed that several of these likely occurred during scaffolding as orientation errors. Some of these errors could also be supported in comparison with the recent assembly of a Chinese accession (YUN_Tarv1.0) of *T. arvense*. Consequently, we manually introduced breakpoints at selected loci in the assembled genome where they were supported by at least two sources of data from whole-genome alignments to YUN_Tarv1.0, synteny maps to *E. salsugineum* (derived from reciprocal best blast), genetic linkage maps (wild-derived and EMS mutation based) and Hi-C contact maps. These were cross-examined with minimap2 alignments of PacBio CLR reads to the genome, an overview of corresponding gene distributions produced by Liftoff v1.5.2 (Shumate and Salzberg, 2020) and the resulting synteny analysis to *E. salsugineum*. The resulting contigs were then rescaffolded with ALLMAPS v1.1.5 (Tang *et al.*, 2015) to produce the final assembly, integrating both the synteny map and genetic map data and manually discounting contigs that were supported only by single markers. The final assembly statistics in comparison with previous intermediary stages are given in Table S2.

Comparative genomics

Genome sequences

Arabidopsis thaliana (Araport 11), *Schrenkiella parvula* (v2.2) and *Arabidopsis lyrata* (v2.1) genome sequences and gene annotation were downloaded from Phytozome (Goodstein *et al.*, 2012). The *Eutrema salsugineum* gene annotation was obtained from Phytozome and lifted over the assembly GenBank GCA_000325905.2.

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

Genome alignments and synteny analysis

The genome alignments between the different versions of the *T. arvense* assembly to *E. salsugineum* were done using MUMmer v4.0.0 (Marçais *et al.*, 2018) with a minimal length of 200 nt and followed by filtering for 1:1 matches and removing alignments smaller than 1000 bp. To identify the interspecies gene orthologs and syntenic relationships between *T. arvense* and other species, we used MCScan in the JCVI utility library (https://github.com/tanghaibao/jcvi; Tang *et al.*, 2008). The ortholog relationships were obtained using the proteinic translation of the CDS and using the argument `--cscore=0.99`. To define the syntenic blocks and the corresponding genomic coordinates, we used the parameters `--minspan=15` and `--minsize=5`. The genomic coordinates from the syntenic blocks were parsed to draw the syntenic relationships using Circos v0.69-8 (Krzywinski *et al.*, 2009).

To determine the different ancestral Brassicaceae chromosomal blocks (ABKs), we took the ortholog relationship between each gene in *T. arvense* and *A. thaliana* from the synteny analysis, and compared it with a gene list derived from Murat *et al.* (2015) where each ortholog gene of *A. thaliana* had an assigned ABK block (Murat *et al.*, 2015).

Genome annotation

Tissue preparation for RNA sequencing

Thlaspi arvense var. MN106-Ref seeds were surface-sterilized with chlorine gas for 1 h and stratified for 3 day at 4 °C. For seedling-stage RNA extractions, seeds were plated on ½ MS medium supplemented with 1% plant agar and stratified for 3 day at 4 °C. For all other tissue collections, plants were sown on soil and grown in a climate-controlled growth chamber in long-day conditions (16/8-h light/dark at 21°/16 °C, light intensity 140 $\mu E/m^2*s$, with 60% relative humidity); plants were watered twice per week. Two weeks after germination, plants growing on soil were vernalized at 4 °C in the dark for 4 weeks, then moved back to the growth chamber. Samples were collected from 11 different tissues in three biological replicates (two in case of mature seeds); for each replicate, we pooled tissue from two individuals. Tissues included the following: one-week-old shoots (from plate culture), one-week-old roots (from plate culture), rosette leaves, cauline leaves, inflorescences, open flowers, young green siliques (about 0.5 × 0.5 cm), older green siliques (about 1 × 1 cm), seed pods, green seeds and mature seeds.

RNA extraction and sequencing

Total mRNA was extracted using the RNeasy Plant Kit (Qiagen, Valencia, CA) and treated with DNase I using the DNA-free Kit DNase Treatment and Removal Reagents (Ambion by Life Technologies, Carlsbad, CA), following the manufacturer's protocols. cDNA libraries were constructed using the NEBNext Ultra II Directional RNA Library Prep Kit (New England BioLabs, Ipswich, MA, USA Inc.) for Illumina following the manufacturer's protocol. Libraries were sequenced on a HiSeq 2500 instrument (Illumina, San Diego, CA) as 125-bp paired-end reads.

Transcriptome assembly

Following quality control and adapter clipping with cutadapt (Martin, 2011), biological replicates for each of eleven tissue types from Illumina mRNA-seq libraries were aligned independently using STAR v2.5.3a (Dobin *et al.*, 2013), then merged according to tissue type, prior to assembly by a reference-based approach. Each assembly was performed using Ryuto v1.3m (Gatter and Stadler, 2019), and consensus reconstruction was then performed using TACO v0.7.3 (Niknafs *et al.*, 2017) to merge tissue-specific transcriptome assemblies. PacBio Iso-seq libraries from MN106-Ref were refined, clustered and polished following the Iso-seq3 pipeline (https://github.com/PacificBiosciences/IsoSeq), prior to alignment with STARlong and isoform collapsing using the cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake) suite. The Iso-seq data were later leveraged together with the Illumina mRNA-seq data to prioritize convergent isoforms using custom in-house scripting.

Genome annotation

The final assembly was annotated using the MAKER-P v2.31.10 (Campbell *et al.*, 2014, 2014) pipeline on the servers provided by the EpiDiverse project, at ecSeq Bioinformatics GmbH (Leipzig, Germany). Plant proteins were obtained from the *Viridiplantae* fraction of UniProtKB/Swiss-Prot and combined with RefSeq sequences derived from selected Brassicaceae: *Arabidopsis thaliana*, *Brassica napus*, *Brassica rapa*, *Camelina sativa* and *Raphanus sativus*. TEs were obtained from RepetDB (Amselem *et al.*, 2019) for selected plant species: *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Arabis alpina*, *Brassica rapa*, *Capsella rubella* and *Schrenkiella parvula (Eutrema parvulum)*. Repeat library construction was carried out using RepeatModeler v1.0.11 (Smit and Hubley, 2008) following basic recommendations from MAKER-P (Campbell *et al.*, 2014). Putative gene fragments were filtered out following BLAST search to the combined Swiss-Prot + RefSeq protein plant database after exclusion of hits from RepetDB. The *de novo* library was combined with a manually curated library of plant sequences derived from repbase (Bao *et al.*, 2015). Genome masking is performed with RepeatMasker v4.0.9 (Smit, 2004) as part of the MAKER-P pipeline. Protein-coding genes, non-coding RNAs and pseudogenes were annotated with the MAKER-P pipeline following two iterative rounds under default settings, using (i) transcript isoforms from Illumina mRNA-seq and PacBio Iso-seq data, (ii) protein homology evidence from the custom Swiss-Prot + RefSeq plant protein database and (iii) the repeat library and TE sequences for masking. The initial results were used to train gene models for *ab initio* predictors SNAP v2006-07-28 (Korf, 2004) and Augustus v3.3.3 (Stanke *et al.*, 2006), which were fed back into the pipeline for the subsequent rounds. The final set of annotations was filtered based on Annotation Edit Distance (AED) < 1 except in cases with corresponding PFAM domains, as derived from InterProScan v5.45-80.0 (Jones *et al.*, 2014). The tRNA annotation was performed with tRNAscan-SE v1.3.1 (Lowe and Eddy, 1997) and the rRNA annotation with RNAmmer v1.2 (Lagesen *et al.*, 2007). The snoRNA homologs were derived using Infernal v1.1.4 (Nawrocki and Eddy, 2013) from plant snoRNA families described in Patra Bhattacharya *et al.* (2016). A small phylogeny based on gene orthologs and duplication events in comparison with selected Brassicaceae (*A. lyrata*, *A. thaliana*, *B. rapa*, *S. parvula* and *E. salsugineum*) was performed with OrthoFinder v2.5.2, and the resulting species tree is rooted using STRIDE (Emms and Kelly, 2017) and inferred from all genes using STAG (Emms and Kelly, 2018).

Transposable element annotation

Two *de novo* annotation tools, EDTA v1.7.0 (Ou *et al.*, 2019) and RepeatModeler v2.0 (Flynn *et al.*, 2020), were used to annotate

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

TEs independently. For EDTA, the following parameters were used in addition to defaults: `--species others`, `--step all`, `--sensitive 1`, `--anno 1`, and `--evaluate 1`. For RepeatModeler2, the additional parameters were `-engine ncbi` and `-LTRStruct`. The outputs of both tools were evaluated by manual curation. First, we used tblastn to align each TE consensus with the transposase database obtained from repbase, and the retrotransposon domains (GAG, Pol, Env, etc.) were viewed one by one with dotter (Sonnhammer and Durbin, 1995). Sequences with multiple paralogs were mapped back to the genome and manually extended to determine the full-length boundary of each TE. A total of 107 full-length, representative *Copia* and *Gypsy* families were successfully evaluated. The TE consensus from RepeatModeler2 was selected as the most accurate model based on full-length paralogs. RepeatMasker was then used to construct the GFF3-like file from the FASTA file from RepeatModeler2, with the optional settings: `-e ncbi -q -no_is -norna -nolow -div 40 -cutoff 225`. The perl script `rmOutToGFF3.pl` was used to generate the final GFF3 file.

sRNA plant material

Seeds were sterilized by overnight incubation at −80 °C, followed by 4 h of bleach treatment at room temperature (seeds in open 2 mL tube in a desiccator containing a beaker with 40 mL chlorine-based bleach (<5%; DanKlorix, Colgate-Palmolive, New York, NY) and 1 mL HCl (32%; Carl Roth, Karlsruhe, Germany)). For rosette, inflorescence and pollen, seeds were stratified in the dark at 4 °C for six days prior to planting on soil, then cultivated under growth chamber conditions of 16–23 °C, 65% relative humidity and a light/dark photoperiod of 16 h:8 h under 110–140 µmol/m^2/s light. Rosette leaves were harvested after two weeks of growth. For inflorescence and pollen, 6-week-old plants were vernalized for 4 weeks at 4 °C in a light/dark photoperiod of 12 h:12 h under 110–140 µmol/m^2/s light. Two weeks after bolting, inflorescence and pollen were collected. Pollen grains were collected by vortexing open flowers in 18% sucrose for 5 min followed by centrifugation at 3000***g*** for 3 min in a swinging bucket rotor. For root samples, seeds were stratified for 6 days at 4 °C in the dark on ½ MS media. Plants were grown in 3–4 mL ½ MS medium plates in long day (16 h) at 16 °C. Root samples were collected 12–14 days after stratification.

sRNA extraction and library preparation

Total RNA was extracted by freezing collected samples with liquid nitrogen and grinding with a mortar and pestle with TRIzol reagent (Life Technologies, Carlsbad, CA). Then, total RNA (1 µg) was treated with DNase I (Thermo Fisher Scientific, Waltham, MA) and used for library preparation. Small RNA libraries were prepared as indicated by the TruSeq Small RNA Library Prep Kit (Illumina, San Diego, CA), using 1 µg of total RNA as input, as described by the TruSeq RNA sample prep V2 guide (Illumina, San Diego, CA). Size selection was performed using the BluePippin System (SAGE Science, Massachusetts). Single-end sequencing was performed on a HiSeq 3000 instrument (Illumina, San Diego, CA).

sRNA locus annotation

Raw FASTQ files were processed to remove the 3′-adapter and quality-controlled with trim_galore v0.6.6 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) using `trim_galore -q 30 --small_rna`. Read quality was checked with FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reference annotation of sRNA loci was created following the steps indicated by Lunardon *et al.* (2020). In short, each library was aligned to the reference genome independently using ShortStack v3.8.5 (Axtell, 2013b), with default parameters, to identify clusters of sRNAs *de novo* with a minimum expression threshold of 2 reads per million (RPM). sRNA clusters from all libraries of the same tissue were intersected using BEDTools v2.26.0 `multiIntersectBed` (Quinlan and Hall, 2010) with default parameters, and only those loci present in at least three libraries were retained. For each tissue, sRNA clusters 25 nt apart were padded together with the `bedtools merge -d` option. sRNA loci whose expression was <0.5 RPM in all libraries of each tissue were also removed. Finally, sRNA loci for all different tissues were merged in a single file retaining tissue of origin information with `bedtools merge -o distinct` options. miRNAs predicted by the ShortStack tool were manually curated (Appendix S1) following the criteria of Axtell (2013b): maximum hairpin length of 300 nt; ≥75% of reads mapping to the hairpin must belong to the miRNA/miRNA* duplex; for the miRNA/miRNA* duplex, no internal loops allowed, two-nucleotide 3′ overhangs, maximum five mismatched bases and only three of which are nucleotides in asymmetric bulges; and mature miRNA sequence should be between 20 and 24 nt.

Expression atlas

Gene expression was measured from the same tissue-specific STAR alignments taken prior to merging biological replicates for transcript assembly, excluding coverage outliers 'mature seed' and 'green old silique'. A total of 27 samples from 9 tissues were therefore considered for gene expression analysis. Raw counts were generated using subread featureCounts v2.0.1 (Liao *et al.*, 2014) and subsequently normalized using the trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010) derived from edgeR v3.34 (Robinson *et al.*, 2010). Averaged expression counts by group were taken for tissue specificity evaluation using the Tau (τ) algorithm (Yanai *et al.*, 2005), as implemented in the R package tispec v0.99.0 (https://rdrr.io/github/roonysgalbi/tispec/), which provides a measure of τ in the range of 0 - 1, where 0 is non/low specificity, and 1 indicates high/absolute specificity.

DNA methylation

We extracted genomic DNA from roots and shoots of 2-week-old seedlings grown on ½ MS medium with 0.8% agar and 0.1% DMSO. Seedlings were grown vertically in 16-h/8-h light/dark cycle; at the time of sampling, roots were separated from shoot tissue with a razor blade and the plant tissue was flash-frozen in liquid nitrogen. Genomic DNA was extracted from ground tissue using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Libraries for WGBS were prepared using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs). Adapter-ligated DNA was treated with sodium bisulphite using the EpiTect Plus Bisulfite Kit (Qiagen, Hilden, Germany) and amplified using the Kapa HiFi Uracil + ReadyMix (Roche, Basel, Switzerland) in 10 PCR cycles. WGBS libraries were sequenced on an Illumina HiSeq2500 instrument with 125-bp paired-end reads.

The WGBS libraries were processed using the nf-core/methylseq v1.5 pipeline (10.5281/zenodo.2555454) combining bwa-meth v0.2.2 (Pedersen *et al.*, 2014) as an aligner and MethylDackel v0.5.0 (https://github.com/dpryan79/MethylDackel) for the methylation calling. The default parameters were used for the entire workflow with the exception of the methylation calling where the following arguments were used: –

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

`D 1000 --maxVariantFrac 0.4 --minOppositeDepth 5 --CHG --CHH --nOT 3,3,3,3 --nOB 3,3,3,3 -d 3`. Only cytosines with a minimum coverage of 3x were kept for the subsequent analysis. Further comparisons between the methylated cytosines and the genome annotation were performed using BEDtools v2.27.1 (Quinlan and Hall, 2010).

Population genomics

DNA from forty pennycress accessions was extracted from approximately 500 mg of leaf tissue pooled from five plants using a plant genomic DNA kit (Epoch Life Science). DNA was then subjected to whole-genome sequencing on an Illumina Novaseq sequencer (2 × 125 bp). Raw reads were then aligned to the new reference genome (T_arvense_v2) using bwa-mem (Li and Durbin, 2009). The aligned files were processed with Samtools and Picard tools, and variants were called using GATK HaplotypeCaller v3.3.0 (Ren *et al.*, 2018). Variants were annotated using SnpEff 5.0e (Cingolani *et al.*, 2012). Data sets for both Indel and SNP panels were trimmed based on LD prior to population genomic analysis using Plink v1.9 (Purcell *et al.*, 2007) with the parameter `--indep-pairwise 1000 5 0.5`. Population structure for both SNP and indel data was then characterized using the admixture model and independent allele frequencies in STRUCTURE v2.3.4 (Pritchard *et al.*, 2000). Dendrograms of both SNP and Indel data were generated under the UPGMA method using the R package poppr (Kamvar *et al.*, 2014).

The forty accessions were planted in a three replication, randomized complete block design in a greenhouse maintained at 21/20 °C and 16 hour days. Ten seeds per replicate were planted in 13.3-cm^2 pots in Sungrow propagation potting mix. Seedlings were thinned to one plant per pot after emergence. Winter annual accessions require vernalization to induce flowering, so all winter accessions were placed in a growth chamber maintained at 4 °C with 16-h light for a period of 21 days about 4 weeks after emergence. Spring annual accessions were planted approximately five weeks after winter accessions. Data for days to flowering were collected on 34 accessions that germinated as the number of days that elapsed from the date of emergence to the appearance of the first flower. The vernalization requirement for winter accessions explains the large differences in mean number of days to flowering between spring and winter accessions. Additional phenotypes and data associated with these sequenced accessions are available in Data S6.

Structural variants using Iso-seq data

Single-molecule real-time (SMRT) isoform sequencing (Iso-seq) based on PacBio (Pacific Biosciences, Menlo Park, CA) generated reads was used to investigate unambiguous full-length isoforms for two pennycress wild accessions, MN108 and Spring32-10. Total RNA extraction was performed on the green seed, hypocotyl, seedling root and flower tissues from pennycress plants grown in a climate-controlled growth chamber maintained 21/20 °C during 16-h:8-h day–night setting. Approximately 250 ng of total RNA was obtained and subjected to the Iso-seq Express Library Workflow (Pacific Biosciences, Menlo Park, CA). cDNA is synthesized from full-length mRNA with the NEBNext Single Cell/ Low Input RNA Prep Kit (New England Biolabs, Ipswich, MA) followed by PCR amplification. The amplified cDNA is converted into SMRTbell templates using the PacBio SMRTbell Express Template Prep Kit 2.0 for sequencing on the Sequel System. Sequencing was performed at the University of Minnesota Genomics Center Facility (Minneapolis, MN).

The polished high-quality FASTA file obtained from Iso-seq3 was aligned to pennycress version 2 (T_arvense_v2) with minimap2 (Li, 2018). The resulting SAM file was sorted and collapsed using the cDNA_Cupcake package to obtain an input GFF file such that each transcript has exactly one alignment and at most one ORF prediction. `Sqanti3_qc.py`, part of the SQANTI3 package (Tardaguila *et al.*, 2018), was deployed on the resulting GFF file along with the reference genome in the FASTA format and a GTF annotation file. This returned a reference corrected transcriptome, transcript-level and junction-level files with structural and quality descriptors, and a QC graphical report. Among the splice junction sites, SQANTI3 defines canonical junctions such as AT-AC, GC-AG and GT-AG, whereas all others are classified as non-canonical splice junctions.

Linkage disequilibrium analysis

Linkage disequilibrium (LD) among genome-wide markers and chromosome-specific markers was calculated with TASSEL v5.2.75 (Bradbury *et al.*, 2007) with a sliding window size of 40 markers with 100 734 460 total comparisons. The *r*-squared values obtained via the linkage disequilibrium function in TASSEL were plotted against the physical distance with a LOESS curve fitted to the data to show LD decay (Figure S14).

Bulked-segregation sequencing and MutMap analysis

Bulked-segregant analysis (BSA) (Michelmore *et al.*, 1991) coupled with whole-genome sequencing (BSA-Seq) was performed to locate genomic region harbouring the gene responsible for the pale mutant phenotype in pennycress (Figure 5d). Two pools were created with one pool containing leaf tissue from 20 individual pale mutants and the other pool consisting of wild-type individuals that did not exhibit the pale phenotype. DNA was extracted from fresh pennycress leaves using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA). Both pools were sequenced on an Illumina HiSeq 2000 instrument using 2 × 125 base-paired reads at the University of Minnesota Genomics Center (Minneapolis, MN). The reads were analysed using the MutMap pipeline (Sugihara *et al.*, 2020), and the QTL region was surveyed for candidate genes.

Comparison with YUN_Tarv_1.0

Synteny between T_arvense_v2 and YUN_Tarv_1.0 was assessed with minimap2 alignments and the resulting dotplot generated with the R package dotPlotly (https://github.com/tpoorten/dotPlotly). The k-mer analysis of quality and completeness was carried out for each assembly with Merqury v1.3 (Rhie *et al.*, 2020; Table S7, S8 and S9), using both the PCR-free Illumina HiSeq reads generated in this study and those obtained from Geng *et al.* (2021) under the accession SRR14757813 in the NCBI Sequence Read Archive.

Acknowledgements

We thank Win Phippen, Thomas Gatter, Prabin Bajgain, Korbinian Schneeberger, Raúl Wijfjes, MPI DB Genome Center, Vienna Biocenter Core Facilities (VBCF), University of Minnesota Genomics Center (UMGC), Minnesota Supercomputing Institute (MSI), and all EpiDiverse network members and beneficiaries. We acknowledge the hard work of many who contributed to this

study including Brett Heim, Krishan Rai, Nicole Folstad, Matthew A. Ott, Shweta Jain and many others.

Funding

This material is based upon work that is supported by the Minnesota Department of Agriculture (J.A., K.F., R.C.) and by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award numbers 2018-67009-27374 (J.A., R.C., K.F.), and 2019-67009-29004 (M.D.M, J.S.) and the Agriculture and Food Research Initiative Competitive Grant No. 2019-69012-29851 (M.D.M, R.C., J.S.). This research was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Science Program grant no. DE-SC0021286 (M.D.M, R.C.). This work was further funded by the Austrian Academy of Sciences (C.B., I.R.A., K.J., D.R.C.); the Max Planck Society (D.W., A.C.G., P.C.B., C.L.); the European Union's Horizon 2020 research and innovation programme by the European Research Council (ERC), Grant Agreement No. 716823 'FEAR-SAP' (I.R.A., C.B.), by the Marie Sklodowska-Curie ETN 'EpiDiverse', Grant Agreement No. 764965 (D.R.C., C.B.) and by Marie Sklodowska-Curie, Grant Agreement MSCA-IF No 797460 (P.C.B.); and the German Federal Ministry of Education and Research BMBF, Grant No. 031A538A, de.NBI-RBC (A.N., P.F.S.).

Conflicts of interest

The authors declare potential competing interests as intellectual property applications have been submitted on some of the genes discussed in this study.

Author contributions

RC, AN, KF and CB conceived the study. RC and AN led the genome assembly and evaluation, assisted by IRA and PCB. IRA performed the comparative genomics analysis of synteny during genome rescaffolding and in the final evaluation. AN led the genome annotation and performed analysis for protein-coding genes, non-coding genes (tRNA, rRNA, snoRNA) and pseudogenes. ACG performed small RNA library sequencing, annotation and analysis, supervised by DW. PZ and ACG performed the transposable element annotation, supervised by DW and MM. AN performed the gene expression analysis and evaluation of tissue specificity. CB and KJ provided PCR-free libraries. RC performed k-mer analysis for genome estimation. KF and RC provided the CCS libraries, which were prepared by the UMGC. CB and IRA provided the DNA methylation libraries and analysis. RC, ZT, MDM and KF developed linkage mapping populations, designed primers, performed genotyping and built genetic maps. KF, RC and KD generated resources for Hi-C, Bionano and resequencing of accessions. KF and ZT phenotyped resequenced accessions. RC performed SNP analysis of resequenced datasets. ZT performed the linkage disequilibrium decay analysis. AB performed population genomics. RC and BJ prepared samples for Iso-seq libraries. RC and ZT performed gene structure variation analysis. RC and MDM performed bulk-segregant analysis. The PacBio CLR library was prepared and sequenced by PCB and AN under the guidance of CL. DR prepared and sequenced mRNA-seq libraries. RC, AN, CB, ACG, IRA and ZT wrote the manuscript. All authors reviewed and approved the manuscript.

Data availability statement

The assembly and all NGS-based raw data are deposited in the ENA Sequence Read Archive Repository (www.ebi.ac.uk/ena/) under study accession number PRJEB46635. The summary of data provided by each institute and corresponding application is described in Table S10.

References

Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V. and Maumus, F. (2019) RepetDB: a unified resource for transposable element references. *Mob. DNA*, **10**, 6.

Andersen, T.G. and Halkier, B.A. (2014) Upon bolting the GTR1 and GTR2 transporters mediate transport of glucosinolates to the inflorescence rather than roots. *Plant Signal. Behav.* **9**, e27740.

Axtell, M.J. (2013a) Classification and comparison of small RNAs from plants. *Annu. Rev. Plant Biol.* **64**, 137–159.

Axtell, M.J. (2013b) ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, **19**, 740–751.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. and Mathews, S. (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA*, **107**, 18724–18728.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

Beric, A., Mabry, M.E., Harkess, A.E., Brose, J., Schranz, M.E., Conant, G.C., Edger, P.P. *et al.* (2021) Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales). *G3 Genes|genomes|genetics*, **11**(7), https://doi.org/10.1093/g3journal/jkab140

Bewick, A.J., Ji, L., Niederhuth, C.E., Willing, E.-M., Hofmeister, B.T., Shi, X., Wang, L. *et al.* (2016) On the origin and evolutionary consequences of gene body DNA methylation. *Proc. Natl. Acad. Sci. USA*, **113**, 9111–9116.

Bewick, A.J. and Schmitz, R.J. (2017) Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.* **36**, 103–110.

Boateng, A.A., Mullen, C.A. and Goldberg, N.M. (2010) Producing stable pyrolysis liquids from the oil-seed Presscakes of mustard family plants: pennycress (*Thlaspi arvense* L.) and Camelina (*Camelina sativa*). *Energy Fuels*, **24**, 6624–6632.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406**, 89–112.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Bucher, E., Reinders, J. and Mirouze, M. (2012) Epigenetic control of transposon transcription and mobility in Arabidopsis. *Curr. Opin. Plant Biol.* **15**, 503–510.

Campbell, M.S., Holt, C., Moore, B. and Yandell, M. (2014) Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinform.* **48**, 4.11.1–39.

Campbell, M.S., MeiYee, L., Carson, H., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Jikai, L. *et al.* (2014) MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**(2), 513–524. https://doi.org/10.1104/pp.113.230144

Catoni, M., Jonesman, T., Cerruti, E. and Paszkowski, J. (2018) Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling. *Nucleic Acids Res.* **47**, 1311–1320.

Chakraborty, M., Baldwin-Brown, J.G., Long, A.D. and Emerson, J.J. (2016) Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **44**, e147.

Chopra, R., Folstad, N., Lyons, J. and Ulmasov, T. (2019) The adaptable use of Brassica NIRS calibration equations to identify pennycress variants to facilitate the rapid domestication of a new winter oilseed crop. *Ind. Crops Prod.* **128**, 55–61.

Chopra, R., Folstad, N. and Marks, M.D. (2020a) Combined genotype and fatty-acid analysis of single small field pennycress (*Thlaspi arvense*) seeds increases

the throughput for functional genomics and mutant line selection. *Ind. Crops Prod.* **156**, 112823.

Chopra, R., Johnson, E.B., Daniels, E., McGinn, M., Dorn, K.M., Esfahanian, M., Folstad, N. *et al.* (2018) Translational genomics using Arabidopsis as a model enables the characterization of pennycress genes through forward and reverse genetics. *Plant J.* **96**(6), 1093–1105. https://doi.org/10.1111/tpj.14147

Chopra, R., Johnson, E.B., Emenecker, R., Cahoon, E.B., Lyons, J., Kliebenstein, D.J., Daniels, E. *et al.* (2020) Identification and stacking of crucial traits required for the domestication of pennycress. *Nature Food*, **1**(1), 84–91. https://doi.org/10.1038/s43016-019-0007-z

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms. *SnpEff. Fly*, **6**(2), 80–92. https://doi.org/10.4161/fly.19695

Claver, A., Rey, R., López, M.V., Picorel, R. and Alfonso, M. (2017) Identification of target genes and processes involved in erucic acid accumulation during seed development in the biodiesel feedstock Pennycress (*Thlaspi arvense* L.). *J. Plant Physiol.* **208**, 7–16.

Cubins, J.A., Wells, M.S., Frels, K., Ott, M.A., Forcella, F., Johnson, G.A., Walia, M.K. *et al.* (2019) Management of pennycress as a winter annual cash cover crop. A review. *Agron. Sustain. Dev.*, **39**, 5. https://doi.org/10.1007/s13593-019-0592-0

Del Gatto, A., Melilli, M.G., Raccuia, S.A., Pieri, S., Mangoni, L., Pacifico, D., Signor, M. *et al* (2015) A comparative study of oilseed crops (*Brassica napus* L. subsp. oleifera and Brassica carinata A. Braun) in the biodiesel production chain and their adaptability to different Italian areas. *Ind. Crops Prod.* **75**, 98–107. https://doi.org/10.1016/j.indcrop.2015.04.029

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dorn, K.M., Fankhauser, J.D., Wyse, D.L. and Marks, M.D. De novo assembly of the pennycress (*Thlaspi arvense*) transcriptome provides tools for the development of a winter cover crop and biodiesel feedstock. *Plant J.* 2013;**75**(6):1028-1038. https://doi.org/10.1111/tpj.12267

Dorn, K.M., Fankhauser, J.D., Wyse, D.L. and Marks, M.D. (2015) A draft genome of field pennycress (*Thlaspi arvense*) provides tools for the domestication of a new winter biofuel crop. *DNA Res.* **22**, 121–131.

Du, H. and Liang, C. (2019) Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nat. Commun.* **10**, 5360.

Emms, D.M. and Kelly, S. (2017) STRIDE: Species tree root inference from gene duplication events. *Mol. Biol. Evol.* **34**, 3267–3278.

Emms, D.M. and Kelly, S. (2018) STAG: Species tree inference from all genes. *bioRxiv*, 267914.

Esfahanian, M., Nazarenus, T.J., Freund, M.M., McIntosh, G., Phippen, W.B., Phippen, M.E., Durrett, T.P. *et al.* (2021) Generating Pennycress (*Thlaspi arvense*) seed triacylglycerols and acetyl-triacylglycerols containing medium-chain fatty acids. *Front. Energy Res.* **9**, https://doi.org/10.3389/fenrg.2021.620118

Fan, J., Shonnard, D.R., Kalnes, T.N., Johnsen, P.B. and Rao, S. (2013) A life cycle assessment of pennycress (*Thlaspi arvense* L.) -derived jet fuel and diesel. *Biomass Bioenergy*, **55**, 87–100.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. and Smit, A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA*, **117**, 9451–9457.

Franzke, A., Lysak, M.A., Al-Shehbaz, I.A., Koch, M.A. and Mummenhoff, K. (2011) Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci.* **16**, 108–116.

Frels, K., Chopra, R., Dorn, K.M., Wyse, D.L., Marks, M.D. and Anderson, J.A. (2019) Genetic diversity of field pennycress (*Thlaspi arvense*) reveals untapped variability and paths toward selection for domestication. *Agronomy*, **9**, 302.

Gatter, T. and Stadler, P.F. (2019) Ryūtō: network-flow based transcriptome reconstruction. *BMC Bioinform.* **20**, 190.

Geng, Y., Guan, Y., Qiong, L., Lu, S., An, M., Crabbe, M.J.C., Qi, J. *et al.* (2021) Genomic analysis of field pennycress (*Thlaspi arvense*) provides insights into mechanisms of adaptation to high elevation. *BMC Biol.* **19**(1), https://doi.org/10.1186/s12915-021-01079-0

Ghurye, J., Pop, M., Koren, S., Bickhart, D. and Chin, C.-S. (2017) Scaffolding of long read assemblies using long range contact information. *BMC Genom.* **18**, 527.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**(D1), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y. and Durbin, R. (2020) Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, **36**, 2896–2898.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Hardcastle, T.J., Müller, S.Y. and Baulcombe, D.C. (2018) Towards annotating the plant epigenome: the *Arabidopsis thaliana* small RNA locus map. *Sci. Rep.* **8**, 6338.

He, G., Chen, B., Wang, X., Li, X., Li, J., He, H., Yang, M. *et al.* (2013) Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. *Genome Biol.* **14**(6), https://doi.org/10.1186/gb-2013-14-6-r57

Hill, W.G. and Weir, B.S. (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* **33**, 54–78.

Jarvis, B.A., Romsdahl, T.B., McGinn, M.G., Nazarenus, T.J., Cahoon, E.B., Chapman, K.D. and Sedbrook, J.C. (2021) CRISPR/Cas9-induced fad2 and rod1 mutations stacked with fae1 confer high oleic acid seed oil in pennycress (*Thlaspi arvense* L.). *Front. Plant Sci.* **12**, 652319.

Johnson, G.A., Kantar, M.B., Betts, K.J. and Wyse, D.L. (2015) Field pennycress production and weed control in a double crop system with soybean in Minnesota. *Agron. J.* **107**, 532–540.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kamvar, Z.N., Tabima, J.F. and Grünwald, N.J. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2**, e281.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinform.* **5**, 59.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. *et al.* (2009) Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T. and Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108.

Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P. *et al.* (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**(8), 771–776. https://doi.org/10.1038/nbt.2303

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964.

Lu, K., Wei, L., Li, X., Wang, Y., Wu, J., Liu, M., Zhang, C. *et al.* (2019) Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. *Nat. Commun.* **10**(1), https://doi.org/10.1038/s41467-019-09134-9

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

Lunardon, A., Johnson, N.R., Hagerott, E., Phifer, T., Polydore, S., Coruh, C. and Axtell, M.J. (2020) Integrated annotations and analyses of small RNA–producing loci from 47 diverse plants. *Genome Res.* **30**, 497–513.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944.

Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Marks, M.D., Chopra, R. and Sedbrook, J.C. (2021) Technologies enabling rapid crop improvements for sustainable agriculture: example pennycress (*Thlaspi arvense* L.). *Emerg. Top Life Sci.* **5**, 325–335.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.

McGinn, M., Phippen, W.B., Chopra, R., Bansal, S., Jarvis, B.A., Phippen, M.E., Dorn, K.M. *et al.* (2019) Molecular tools enabling pennycress (*Thlaspi arvense*) as a model plant and oilseed cash cover crop. *Plant Biotechnol. J.* **17** (4), 776–788. https://doi.org/10.1111/pbi.13014

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. and Koornneef, M. (1998) *Arabidopsis thaliana*: a model plant for genome analysis. *Science*, **282**, 662, 679–82.

Michalovova, M., Vyskot, B. and Kejnovsky, E. (2013) Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity*, **111**, 314–320.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA*, **88**, 9828–9832.

Moore, S.A., Wells, M.S., Gesch, R.W., Becker, R.L., Rosen, C.J. and Wilson, M.L. (2020) Pennycress as a cash cover-crop: improving the sustainability of sweet corn production systems. *Agronomy*, **10**, 614.

Moser, B.R. (2012) Biodiesel from alternative oilseed feedstocks: camelina and field pennycress. *Biofuels*, **3**, 193–209.

Moser, B.R., Knothe, G., Vaughn, S.F. and Isbell, T.A. (2009) Production and evaluation of biodiesel from field pennycress (*Thlaspi arvense* L.) oil. *Energy Fuels*, **23**, 4149–4155.

Mulligan, G.A. (1957) Chromosome numbers of Canadian weeds. I. *Can. J. Bot.* **35**, 779–789.

Mulligan, G.A. and Kevan, P.G. (1973) Color, brightness, and other floral characteristics attracting insects to the blossoms of some Canadian weeds. *Can. J. Bot.* **51**, 1939–1952.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Crollius, H.R. *et al.* (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.* **16**(1), https://doi.org/10.1186/s13059-015-0814-y

Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Niederhuth, C.E., Bewick, A.J., Ji, L., Alabady, M.S., Kim, K.D., Li, Q., Rohr, N.A. *et al.* (2016) Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194.

Niittylä, T., Messerli, G., Trevisan, M., Chen, J., Smith, A.M. and Zeeman, S.C. (2004) A previously unknown maltose transporter essential for starch degradation in leaves. *Science*, **303**, 87–89.

Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M. and Iyer, M.K. (2017) TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods*, **14**, 68–70.

Nour-Eldin, H.H., Andersen, T.G., Burow, M., Madsen, S.R., Jorgensen, M.E., Olsen, C.E. and Dreyer, I. (2012) NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *Nature*, **488**, 531–534.

Ott, M.A., Eberle, C.A., Thom, M.D., Archer, D.W., Forcella, F., Gesch, R.W. and Wyse, D.L. (2019) Economics and agronomics of relay-cropping pennycress and Camelina with soybean in Minnesota. *Agron. J.* **111**, 1281–1292.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B. *et al.* (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**(1), https://doi.org/10.1186/s13059-019-1905-y

Patra Bhattacharya, D., Canzler, S., Kehr, S., Hertel, J., Grosse, I. and Stadler, P.F. (2016) Phylogenetic distribution of plant snoRNA families. *BMC Genom.* **17**, 969.

Pedersen, B.S., Eyring, K., De, S., Yang, I.V. and Schwartz, D.A. (2014) *Fast and accurate alignment of long bisulfite-seq reads. arXiv [q-bio.GN].*

Phippen, W.B. and Phippen, M.E. (2012) Soybean seed yield and quality as a response to field pennycress residue. *Crop Sci.* **52**, 2767–2773.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pruitt, K.D., Tatusova, T., Brown, G.R. and Maglott, D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J. *et al.* (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Human Genet.* **81**(3), 559–575. https://doi.org/10.1086/519795

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Ren, S., Bertels, K. and Al-Ars, Z. (2018) Efficient acceleration of the Pair-HMMs forward algorithm for GATK HaplotypeCaller on graphics processing units. *Evol. Bioinform. Online*, **14**, 1176934318760543.

Rhie, A., Walenz, B.P., Koren, S. and Phillippy, A.M. (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25.

Scheben, A., Yuan, Y. and Edwards, D. (2016) Advances in genomics for adapting crops to climate change. *Curr. Plant Biol.* **6**, 2–10.

Schranz, M.E., Lysak, M.A. and Mitchell-Olds, T. (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* **11**, 535–542.

Sedbrook, J.C., Phippen, W.B. and Marks, M.D. (2014) New approaches to facilitate rapid domestication of a wild plant to an oilseed crop: example pennycress (*Thlaspi arvense* L.). *Plant Sci.* **227**, 122–132.

Shumate, A. and Salzberg, S.L. (2020) Liftoff: accurate mapping of gene annotations. *Bioinformatics*, **37**, 1639–1643.

Sigman, M.J. and Slotkin, R.K. (2016) The first rule of plant transposable element silencing: location, location, location. *Plant Cell*, **28**, 304–313.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Smit, A.F.A. (2004) *Repeat-Masker Open-3.0*. http://www.repeatmasker.org

Smit, A.F.A. and Hubley, R. (2008) *RepeatModeler Open-1.0*.

Sonnhammer, E.L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–10.

Stam, P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J.* **3**, 739–744.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439.

Sugihara, Y., Young, L., Yaegashi, H., Natsume, S., Shea, D.J., Takagi, H., Booker, H. *et al.* (2020) *High-performance pipeline for MutMap and QTL-seq*. bioRxiv, 2020.06.28.176586.

Sultana, T., Zamborlini, A., Cristofari, G. and Lesage, P. (2017) Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* **18**, 292–308.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J.C., Schnable, P.S. *et al.* (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3.

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., del Risco, H., Ferrell, M. *et al.* (2018) SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* **28**(3), 396–411. https://doi.org/10.1101/gr.222976.117

Thomas, J.B., Hampton, M.E., Dorn, K.M., David Marks, M. and Carter, C.J. (2017) The pennycress (*Thlaspi arvense* L.) nectary: structural and transcriptomic characterization. *BMC Plant Biol.* **17**, 201.

Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Vaser, R., Sović, I., Nagarajan, N. and Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746.

Voinnet, O. (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell*, **136**, 669–687.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A. *et al.* (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9** (11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y. *et al.* (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**(10), 1035–1039. https://doi.org/10.1038/ng.919

Warwick, S.I., Francis, A. and Susko, D.J. (2002) The biology of Canadian weeds. 9. *Thlaspi arvense* L. (updated). *Can. J. Plant Sci.* **82**, 803–823.

Weyers, S.L., Gesch, R.W., Forcella, F., Eberle, C.A., Thom, M.D., Matthees, H.L., Ott, M. *et al.* (2021) Surface runoff and nutrient dynamics in cover crop–soybean systems in the Upper Midwest. *J. Environ. Qual.* **50**(1), 158–171. https://doi.org/10.1002/jeq2.20135

Weyers, S., Thom, M., Forcella, F., Eberle, C., Matthees, H., Gesch, R., Ott, M. *et al.* (2019) Reduced potential for nitrogen loss in cover crop-soybean relay systems in a cold climate. *J. Environ. Qual.* **48**(3), 660–669. https://doi.org/10.2134/jeq2018.09.0350

Workman, R., Fedak, R., Kilburn, D., Hao, S., Liu, K. and Timp, W. (2019) *High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing v1 (protocols.io.4vbgw2n). protocols.io.*.

Workman, R.E., Myrka, A.M., Wong, G.W., Tseng, E., Welch, K.C. and Timp, W. (2018) Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird Archilochus colubris. *GigaScience*, **7**(3), giy009. https://doi.org/10.1093/gigascience/giy009

Wu, D., Liang, Z., Yan, T., Xu, Y., Xuan, L., Tang, J., Zhou, G. *et al.* (2019) Whole-genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. *Molecular Plant*, **12**(1), 30–43. https://doi.org/10.1016/j.molp.2018.11.007

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21** (5), 650–659. https://doi.org/10.1093/bioinformatics/bti042

Yang, R., Jarvis, D.E., Chen, H., Beilstein, M.A., Grimwood, J., Jenkins, J., Shu, S. *et al.* (2013) The reference genome of the halophytic plant *Eutrema salsugineum*. *Front. Plant Sci.* **4**, https://doi.org/10.3389/fpls.2013.00046

Zhang, H., Lang, Z. and Zhu, J.-K. (2018) Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* **19**, 489–506.

Zhang, S.-J., Liu, L., Yang, R. and Wang, X. (2020) Genome size evolution mediated by gypsy retrotransposons in Brassicaceae. *Genom. Proteom. Bioinform.* **18**, 321–332.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R. *et al.* (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, **126**(6), 1189–1201. https://doi.org/10.1016/j.cell.2006.08.003

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2006) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* **39**, 61–69.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.-H. (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.* **151**, 3–15.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Karyotype plot of the seven largest scaffolds representing chromosomes in *T. arvense* MN106-Ref (T_arvense_v2), alongside a concatenation of all minor scaffolds.

Figure S2 Integrative Genome Viewer (IGV) snapshot of PacBio read coverage (top track) over the largest seven scaffolds of the genome, including distributions of genes (middle track) and transposable elements (bottom track).

Figure S3 Sequence dot plots showing the largest seven scaffolds of the closely-related species *E. salsugineum* and their equivalent in *T. arvense* var. MN106-Ref (T_arvense_v2), comparing the difference both (a) before and (b) after rescaffolding.

Figure S4 Synteny analysis between the largest seven scaffolds of *T. arvense* var. MN106-Ref (Ta) and (a) their equivalent in the closely-related species *E. salsugineum* (Es), and (b) *A. thaliana* (At).

Figure S5 The cumulative distribution of annotation edit distance (AED) scores from the final set of protein-coding loci, denoting that ~95% of annotated genes are supported with a score ≤ 0.5 overall.

Figure S6 An overview of annotated genomic feature distributions in comparison to T_arvense_v1 for (a) gene lengths, (b) CDS lengths, (c) per gene exon number, and (d) intron lengths.

Figure S7 Small RNA (sRNA) annotation in the T_arvense_v2 genome assembly.

Figure S8 Predicted miRNAs in the T_arvense_v2 genome assembly.

Figure S9 Relative expression level of novel and conserved miRNA families between tissue types.

Figure S10 sRNA types and their association with different genomic features.

Figure S11 Methylation rate frequency distribution by sequence context in shoot and root tissues.

Figure S12 Map showing original sampling sites of pennycress accessions used for resequencing analysis in this study.

Figure S13 Structure plot showing inferred population membership for SNP data (top) and Indel data (bottom) at $k = 3$ for the resequenced accessions.

Figure S14 Genome-wide linkage disequilibrium decay plotted against physical distance for MN106-Ref (T_arvense_v2) at an r-squared value of 0.2 and chromosome level LD decay described in the right. Linkage disequilibrium (LD) was calculated using 2 518 379 genome-wide markers with a sliding window of 40 markers.

Figure S15 Synteny between T_arvense_v2 (*x*-axis) and YUN_-Tarv_1.0 (*y*-axis).

Figure S16 Read length distribution of trimmed PacBio Sequel II HiFi CLR reads taken forward for assembly with Canu v1.9.

Figure S17 Distribution of PacBio Sequel II HiFi CLR read mapping depth frequency over assembled contigs, with bimodal peaks due to contig regions with lower depth than the average indicating that they are duplicated.

Table S1 Estimation of the genome size of *T. arvense* using flow cytometry with *Arabidopsis thaliana*, tomato (*Solanum lycopersicum*), and maize (*Zea mays*) as references.

Table S2 Full descriptive statistics for intermediate versions of the assembly starting with correction, trimming and initial assembly

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

of PacBio reads (Canu), further polishing and scaffolding using optical maps and contact maps (Bionano + HiC), and the final version following manual curation and rescaffolding with the help of genetic linkage and synteny maps (ALLMAPS).

Table S3 Alignment statistics of mRNA-seq reads prior to merging by tissue type.

Table S4 Detailed per-class statistics of the transposable element fraction of the *T. arvense* genome.

Table S5 Description of genes identified in the QTL region (Scaffold_6: 63.85–63.95 Mbp) of the BSA analysis of pale seedling phenotype in pennycress.

Table S6 BUSCO statistics on (a) initial assembly, immediately after CANU, and (b) final assembly. Both are derived from orthologs to the *Eudicotyledons odb10* database.

Table S7 Merqury k-mer (k = 21) analysis of Illumina HiSeq reads sequenced from the accession in YUN_Tarv_1.0, showing greater QV scores in T_arvense_v2 for the equivalent top 7 scaffolds based on k-mers found uniquely in each assembly and those shared with the read set.

Table S8 Merqury k-mer (k = 21) analysis of Illumina HiSeq reads (PCR-free) sequenced from the accession MN106-Ref, showing greater QV scores in T_arvense_v2 for the equivalent top 7 scaffolds based on k-mers found uniquely in each assembly and those shared with the read set.

Table S9 Merqury k-mer (k = 21) analysis of each total assembly showing relative completeness of k-mers present in each read set from Illumina HiSeq.

Table S10 Summary of data provided by each institute and corresponding application.

Appendix S1 Manual curation of predicted miRNAs.

Data S1 Normalized read counts for the genes expressed in each of the tissues analysed (See excel file). Tau values are incorporated in each of the genes to highlight the specificity.

Data S2 Top 30 most-expressed genes in each tissue, relative to the mean across all tissues, from the subset of genes with a high/absolute tau specificity score.

Data S3 Location of SNPs and the primers used in the genotyping of EMS-based population for development of linkage map.

Data S4 Genetic map developed using an F2 population derived from MN106 and Ames32867.

Data S5 Genetic map developed using an F2 population derived from MN106 and 2019-M2-111.

Data S6 Phenotypes, total reads, and coverage associated with the accessions used for GWAS.

© 2022 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–20

The end