Characterization of the pangenome variability and its role in domestication and adaptation in African rice

Christine Tranchant Dubreuil

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

En Génétique et Génomique École doctorale GAIA

Unité de recherche DIADE

Caractérisation de la variabilité du pangénome et son rôle dans la domestication et l'adaptation au sein du riz Africain

Présentée par Christine TRANCHANT-DUBREUIL

le 22 juin 2023

Sous la direction de François SABOT et Yves VIGOUROUX

Devant le jury composé de

Karine ALIX, Professeure AgroParisTech **UMR Génétique Quantitative et Evolution, Gif-sur-Yvette**	**Rapportrice**
Hadi QUESNEVILLE, Directeur de Recherche **Unité de Recherche Génomique Info, INRAE, Versailles**	**Rapporteur**
Leandro QUADRANA, Chargé de Recherche **Quantitative Genomics and Epigenomics - Institut of Plant Science Paris-Saclay, Paris**	**Examinateur**
Joëlle RONFORT, Directrice de Recherche **UMR AGAP, INRAE, Montpellier**	**Examinatrice**
Alain GHESQUIERE, Directeur de Recherche **UMR DIADE, Montpellier**	**Invité**

UNIVERSITÉ DE MONTPELLIER

Résumé en Français

En préambule

Débutée il y a environ 12 000 ans, la domestication des plantes a révolutionné la trajectoire des sociétés humaines, façonné nos sociétés en société agricole et modifié nos paysages (Fuller et al. 2014). Les plantes domestiquées fournissent directement ou indirectement la majeure partie de notre alimentation. elles ont été à la base du développement de notre système agricole et, plus largement, de nos villes et de notre civilisation. Sur plus de 250 000 espèces d'angiospermes dans le monde, seul 1% des espèces ont été domestiquées et seulement 103 espèces fournissent 90% de l'apport énergétique alimentaire mondial. Le riz, le maïs et le blé en représentent les deux tiers (Dirzo and Raven 2003). Cette domestication est associée à une série de modifications de traits communs à de nombreuses plantes domestiquées qui permet de les distinguer de leurs apparentés naturels (Doebley, Gaut, and Smith 2006) (Figure 1) tels que le non-égrenage ou des fruits plus gros.

De nombreuses études ont examiné l'impact de la sélection humaine sur la diversité génétique des espèces cultivées en comparant les génomes des plantes cultivées et de leurs ancêtres sauvages (Berger et al. 2012; Eyre-Walker et al. 1998; Hyten et al. 2006). Elles ont permis de mieux comprendre comment la domestication a "bricolé" une plante cultivée, à la fois en filtrant certaines allèles parmi l'ensemble des variations alléliques existantes et en sélectionnant de nouvelles mutations responsables de traits physiologiques ou phénotypiques intéressants.

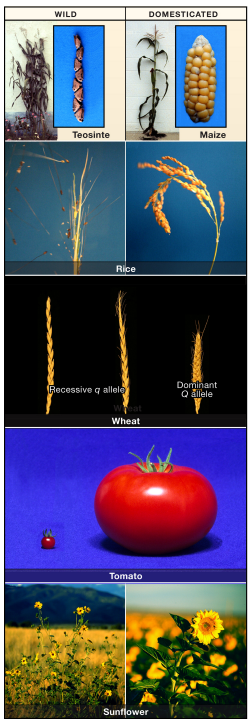

Figure 1: Phénotype de quelques plantes domestiquées et de leur ancêtre sauvage.
Figure issue de Doebley, Gaut, and Smith 2006.

La diversité génétique peut être observée au niveau de l'ADN à travers un large continuum de

mutations allant de la variation d'une seule base (SNP), à de petites insertions-délétions (indels, < 50pb) et à des variations structurales plus importantes. Ces dernières comprennent une multitude de types tels que les duplications, les inversions, les translocations, les insertions/délétions, les transpositions ou des variations en nombre de copies (Copy Number Variation, CNV) ou la présence/absence d'une unique variation (Presence/Absence Variation, PAV) (Gabur et al. 2018). Ces mutations génèrent de la diversité génétique sur laquelle d'autres mécanismes évolutifs vont pouvoir agir. Ainsi la fréquence de ces variations au sein d'une population d'individus pourra être modifiée comme l'est la fréquence des SNP sous l'impact de processus tels que les nouvelles mutations, les flux de gènes (ou migration), la dérive génétique ou la sélection naturelle. Cette dernière peut être négative, positive ou balancée.

Après la publication de la première séquence de référence d'une plante, *Arabidopsis thaliana* (Initiative 2000), il y a maintenant 20 ans, plus de 1000 génomes de plantes provenant de 788 espèces ont été publiés. Les progrès spectaculaires des technologies de séquençage à haut débit et des algorithmes d'assemblage optimisés accélèrent aujourd'hui ce phénomène (Sun, Shang, et al. 2022, Marks et al. 2021). Ces progrès ont permis de séquencer et assembler des génomes même complexes et de grande taille, comme celui de *Pinus lambertiana*, d'une taille de 31 Gb (Stevens et al. 2016). S'appuyant sur les génomes de référence, de nombreuses études sur la diversité génétique ont été réalisées, d'abord basées sur le polymorphisme de type "SNP" (Atwell et al. 2010; Lai et al. 2012; Xu, Liu, et al. 2012; Zhou, Jiang, et al. 2015), puis, progressivement sur des variations structurales plus larges (Muñoz-Amatriaín et al. 2013; Springer et al. 2009; Swanson-Wagner et al. 2010).

En comparant plusieurs génomes au sein d'une même espèce, il est devenu de plus en plus évident qu'un seul génome de référence ne peut capturer toute la diversité présente au sein d'une espèce. D'importantes variations du nombre de gènes, et plus largement des variations structurales, entre les individus d'une même espèce ont été observées. Cela a conduit à un changement progressif de paradigme, du concept de génome à celui de pangénome. Le pangénome désigne l'ensemble des séquences (géniques ou non) présentes au sein d'un groupe d'individus et il se compose du "core genome" regroupant les séquences présentes chez tous les individus et (ii) le "dispensable genome" contenant des séquences uniquement partagées par une partie des individus (Figure 2, Morgante, De Paoli, and Radovic 2007; Tettelin et al. 2005). Au travers de plusieurs études, la pangénomique est rapidement apparue comme une

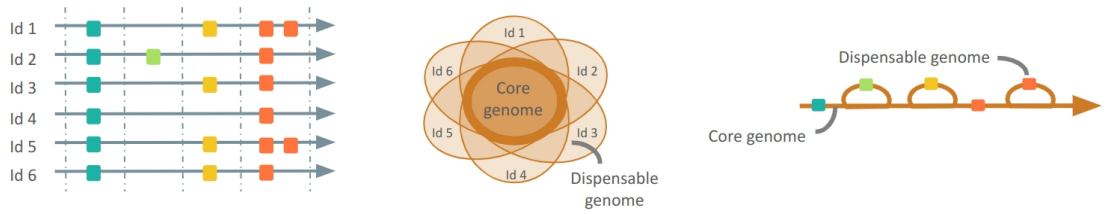

Figure 2: Trois représentations du pangénome.
Le pangénome est représenté (à gauche) comme un inventaire d'éléments génomiques partagés ou non au sein d'un groupe d'individus; (au centre) sous la forme d'un diagramme de Venn dans lequel le génome central est l'ensemble commun de séquences partagées par tous les individus du groupe, le reste appartient au génome dispensable; (à droite) sous la forme d'un graphe orienté, dans lequel des chemins alternatifs représentent les variantes structurales.

nouvelle approche particulièrement intéressante pour explorer le large contenu des variations structurales à l'échelle d'une espèce ou plus largement. Ainsi, plusieurs études ont mis en évidence qu'un grand nombre de séquences, dont des gènes, font partie du génome accessoire, en anglais "dispensable". Gordon et al. ont estimé que jusqu'à 8 Mo de séquences présentes dans chaque individu de *Brachypodium distachyon* étaient absentes du génome de référence (Gordon et al. 2017). Chez la tomate, Gao et al. ont détecté 4 873 gènes absents du génome de référence mais présents dans d'autres accessions cultivées et sauvages séquencées (Gao et al. 2019). De plus, de nombreuses études ont également montré que les variations structurales présentes dans ce génome accessoire ("dispensable") peuvent être associées à

des variations phénotypiques telles que le temps de floraison (Gordon et al. 2017; Song et al. 2020) ou, plus spécifiquement, à des variations entre des accessions cultivées et sauvages telles que la couleur de l'enveloppe des grains de soja (Song et al. 2020) ou la couleur des fruits chez la fraise (Qiao et al. 2021).

Les principaux objectifs de ce projet doctoral ont été d'explorer la diversité génétique au sein d'une céréale, le riz africain, et la manière dont la diversité de cette espèce a été remodelée au cours de sa domestication. Les deux principales questions que nous souhaitions aborder étaient les suivantes : quels rôles jouent ces variations structurales dans la composition en gènes et l'adaptation, et comment les forces évolutives ont façonné l'organisation et la dynamique du (pan)génome. Tout d'abord, nous avons réalisé un état de l'art des connaissances sur le concept émergent de pangénome, afin d'identifier les défis, les opportunités et les limites. Nous avons ensuite développé une approche et un outil pour créer un pangénome d'eucaryote à partir de données de séquençage "short-read" de génomes d'individus. Profitant du reséquençage de 247 génomes de riz africains, nous avons appliqué notre approche comme preuve de concept pour construire le premier pangénome de riz africain. Enfin, après avoir caractérisé ces pangénomes aux niveaux inter- et intra-espèces, nous avons étudié comment la domestication a façonné le pangénome du riz africain.

Pangenome : du concept à des études de diversité à large échelle au sein d'espèces

Un état de l'art a tout d'abord été réalisé pour définir comment appliquer une approche pangénomique à notre problématique. En 2019, au début de ce travail, le concept de pangénome était largement utilisé au sein des espèces bactériennes (Medini et al. 2005; Tettelin et al. 2005, Vernikos et al. 2015) et commençait à être étendu aux organismes supérieurs, bien que l'analyse pangénomique ait été (et soit encore aujourd'hui) un défi, en raison de la grande taille et de la complexité de leur génome (contenu en séquences répétées ou polyploïdie). A cette époque, le nombre d'analyses pangénomiques sur les plantes ne cessait d'augmenter (Contreras-Moreira et al. 2017; Golicz, Bayer, et al. 2016; Gordon et al. 2017; Li, Zhou, Ma, et al. 2014; Schatz et al. 2014; Wang, Mauleon, et al. 2018; Yao et al. 2015; Zhao et al. 2018, Figure 3). Cependant, il y avait peu de revue sur ce concept émergent chez les plantes et l'article de Golicz, Batley, and Edwards, publié en 2016, en est l'un des rares exemples (Golicz, Batley, and Edwards 2016). Par conséquent, rédiger une revue sur la pangénomique chez les plantes a été une opportunité pour faire une synthèse de la littérature, identifier les défis, les limites et les idées de recherche sur ce sujet émergent. Nous avons également pensé que cette revue pourrait être une ressource utile pour d'autres chercheurs qui se posaient les mêmes questions que nous, comme par exemple :

- Pourquoi ce changement de paradigme du concept de génome à pangenome?
- Qu'est-ce qu'un pangénome? Implique-t-il uniquement des gènes?
- Comment construire un pangénome?
- Quels sont les facteurs et les forces qui influencent la capacité du pangénome à augmenter sa taille?

Des études antérieures avaient également montré l'absence de gènes spécifiques dans un génome de référence, tels que les gènes conférant une tolérance à la submersion (Submergence 1, Sub1) (Xu, Xu, et al. 2006) ou l'immersion (SNORKEL1, SNORKEL2) (Hattori et al. 2009) chez le riz asiatique. Cependant, les études pangénomiques ont mis en évidence que cela pouvait impliquer un grand nombre de gènes. Par exemple, parmi les premiers articles sur la pangénomique des plantes, deux études sur le blé (Montenegro et al. 2017) et le riz (Yao et al. 2015) ont montré que 12 150 gènes et 8 000 gènes, respectivement, manquaient dans le génome de référence. En 2017, Gordon et al. ont identifié des gènes absents du génome de référence de *Brachypodium distachyon* qui étaient impliqués dans l'adaptation à l'environnement en influençant la variation du temps de floraison (Gordon et al. 2017). Ainsi, au cours des dernières décennies, il est apparu de plus en plus clairement qu'un seul génome de référence était insuffisant pour capturer toutes les séquences présentes dans une espèce et que le concept de génome de référence devait être repensé. C'est l'un des points de départ du passage du paradigme du génome de référence au pangénome.

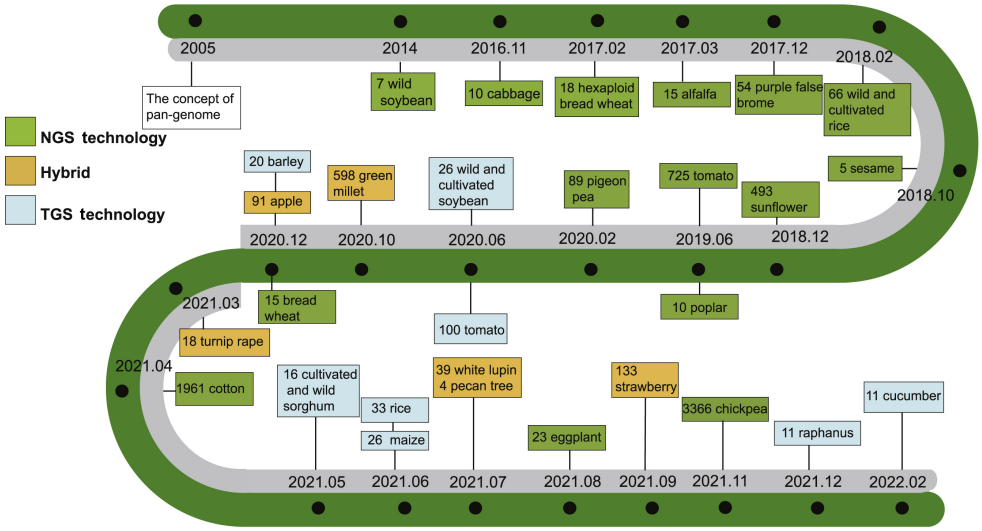

Figure 3: Chronologie et informations relatives aux pangénomes de plantes publiés.
Les différentes technologies de séquençage utilisées pour construire les pangénomes sont indiquées par des couleurs différentes. Les cercles noirs pleins indiquent les études de pan-génomique de plante. Les technologies sont indiquées à l'aide des rectangles colorés : vert clair, séquençage de nouvelle génération ; orange foncé, séquençage hybride ; bleu clair, séquençage à lecture longue. Figure issue de Li, Liu, et al. 2022.

Dans la revue, en plus de la définition usuelle centrée sur les gènes, nous avons proposé une définition plus complète qui englobait toutes les séquences du génome (qu'elle soit génique ou non) (Figure 2). Nous avons ensuite discuté de certaines étapes clé d'une analyse pangénomique. Cette dernière se compose de différentes étapes, de la construction du pangénome à la visualisation du pangénome (le graal) et à la caractérisation du pangénome (Figure 4). Avant de commencer la construction du pangénome proprement dite, il y a une phase importante d'échantillonnage et de séquençage qui dépendra de la question scientifique à laquelle on souhaite répondre. Il est également important de noter que plusieurs facteurs peuvent avoir un impact sur la construction et la caractérisation d'un pangénome, à commencer par la taille de l'échantillon. Des questions récurrentes se posent donc au début d'une telle analyse :

- Combien de génomes doivent être séquencés pour maximiser la diversité au sein d'un groupe?
- Est-ce qu'il y aura toujours le même ensemble de séquences partagées par tous les individus si un génome nouvellement séquencé est ajouté?
- Combien de nouveaux gènes "dispensable" seront-ils découverts avec le séquençage d'individus supplémentaires?

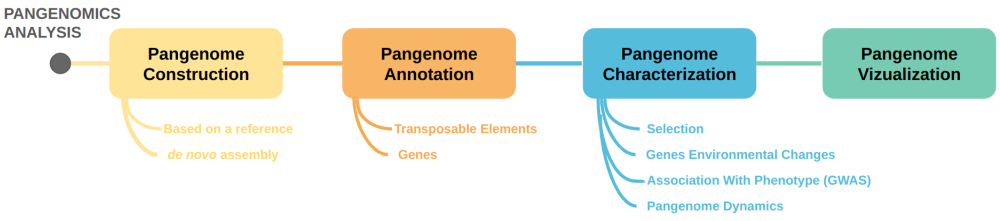

Figure 4: Vue d'ensemble d'une analyse pangénomique.

Ainsi, la taille du pangénome et du "core genome" peuvent être sous-estimées si un nombre insuffisant d'échantillons est utilisé ou si les échantillons choisis ne représentent qu'une petite fraction de la diversité réelle. Deux analyses pangénomiques sur le riz ont illustré l'impact du nombre d'échantillons. En 2014, une étude portant sur trois accessions de riz asiatique a révélé un pangénome de 40 362 gènes, dont

8% étaient variables (Schatz et al. 2014). Quatre ans plus tard, dans une étude basée sur plus de 3 000 accessions de riz (Wang, Mauleon, et al. 2018), Wang, Mauleon, et al. ont identifié un pangénome beaucoup plus important (48 098 gènes) avec un pourcentage plus élevé de gènes variables (41%). Au-delà du choix des échantillons, les propriétés génétiques du modèle étudié peuvent également influencer les résultats d'une étude pangénomique tels que la taille du génome (ex: contenu en TE), le mode de reproduction ou le niveau de ploïdie (ex: allogamie ou autogamie). En conclusion, la comparaison d'analyses pangénomiques, même réalisées sur des espèces proches, peut s'avérer complexe en raison de nombreux facteurs, y compris ceux liés aux méthodes utilisées pour construire le pangénome. En 2019, deux approches étaient utilisées pour assembler les pangénomes et elles se basaient essentiellement sur le reséquençage de génome à partir de "reads" courts : les méthodes "assemble-then-map" et "map-then-assemble". La première consiste en un assemblage du génome *de novo* suivi d'une comparaison des génomes (Gordon et al. 2017; Hu et al. 2017; Li, Zhou, Ma, et al. 2014; Schatz et al. 2014; Zhao et al. 2018), tandis que la seconde est basée sur l'alignement des "reads" suivi de l'assemblage *de novo* des reads non alignés (Golicz, Bayer, et al. 2016; Montenegro et al. 2017; Yao et al. 2015).

Dans cette revue, nous avons également synthétisé les connaissances acquises ces dernières années pour déchiffrer à la fois la structure des pangénomes et leur dynamique, ainsi que la façon dont la structure des pangénomes est façonnée par des mécanismes (ex: introgression, transfert horizontal) et des processus évolutifs (ex: mutations, sélection, dérive génétique). La taille du pangénome dépend de la dynamique du génome du groupe considéré. A cet égard, les bactéries ont un pangénome relativement plus grand que les plantes en raison de leur niveau plus élevé de flux de gènes. La capacité du pangénome à croître ou à rester stable, ainsi que de passer du "dispensable genome" ou "core genome" (et vice-versa) est étroitement liée à l'équilibre entre les événements de gain et de perte et sans doute aussi même si cela est plus discuté à la capacité d'adaptation à divers environnements (Figure 5). La pangénomique peut ainsi contribuer à mieux comprendre les mécanismes évolutifs qui permettent aux organismes de s'adapter à de nouveaux environnements et d'étudier les processus d'adaptation et de sélection au sein d'un groupe (ex: population, espèce, genre).

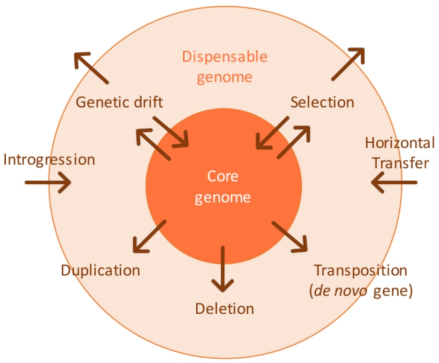

Figure 5: Vue d'ensemble de la dynamique de la structure du pangénome façonnée par différents événements et forces.
De nouvelles séquences sont ajoutées au génome "dispensable" par le biais de mutations, de duplications, de délétions et de transpositions, tandis que le contenu du génome "core" peut diminuer par le biais de deletions et de transpositions. Le transfert horizontal et l'introgression ont également un impact sur le compartiment "dispensable" du génome (gain de séquences). En outre, la sélection positive et négative ainsi que la dérive génétique ont un impact sur les génomes "core" et "dispensable" (gain et perte de séquences) ainsi que sur le pangénome (perte de séquences). Figure adaptée de Tranchant-Dubreuil, Rouard, and Sabot 2019.

Le concept de pangénome, combiné aux technologies de séquençage de troisième génération et aux méthodes renseignant l'agencement des chromosomes (telles que Hi-C, carte optique) offre de nouvelles possibilités d'aborder les questions biologiques sous un nouvel angle. Cependant, la pangénomique sera

confrontée à de nouveaux défis en matière d'analyse, de stockage et de visualisation de la masse de données générée par cette approche nouvellement développée. Cette étude bibliographique a été valorisée dans un article de revue publié à *Annual Plant Reviews* en 2019 (Tranchant-Dubreuil, Rouard, and Sabot 2019) et nous a permis de développer une approche et un outil pour construire le premier pangénome du riz africain.

Développement de l'outil FrangiPANe pour contruire le pangénome du riz africain

Pour étudier la diversité génétique d'une population via une approche pangénomique, la première étape consiste à détecter l'ensemble des variations, c'est-à-dire construire le pangénome, afin de définir ce qui est "core" ou "dispensable" dans une population dans une deuxième étape. Cette étape primordiale sera d'autant plus difficile si l'on veut détecter toutes les variations (géniques ou non), sur un grand nombre d'individus. Si, de plus, cette analyse est réalisée sur des Eucaryotes, qui ont des génomes larges et complexes (ex. contenu élevé en séquences répétées, polyploïdie), la construction du pangénome est un processus encore plus complexe et long. Nous avons développé une approche et un outil "tout-en-un" qui simplifie cette tâche complexe de construction d'un pangénome à partir de multiples séquençages individuels du génome basés sur des lectures courtes. Très peu d'outils sont disponibles pour effectuer toutes les étapes en même temps, étant soit développés pour les bactéries (Ding, Baumdicker, and Neher 2018; Laing et al. 2010; Page et al. 2015), soit basés sur l'approche *de novo* 'assemble-then-map' (Hu et al. 2017). L'outil FrangiPANe a été mis en œuvre pour appliquer facilement l'approche "map-then-assemble" et pour créer une pangénome pour n'importe quel organisme eucaryote à partir de données de lectures courtes (Figure 6). FrangiPANe simplifie la construction d'un pangénome en fournissant à la fois l'ensemble du processus à partir de ses propres données, *i.e.* de l'alignement des "reads" sur le génome à l'ancrage des contigs assemblés sur ce même génome, ainsi que l'ensemble des logiciels bioinformatiques nécessaires, au sein d'une seule machine virtuelle. En outre, une interface unique permet d'effectuer chaque étape de l'analyse et d'analyser progressivement les données générées de différentes manières, telles que des tableaux ou des graphiques, via un jupyterbook unique.

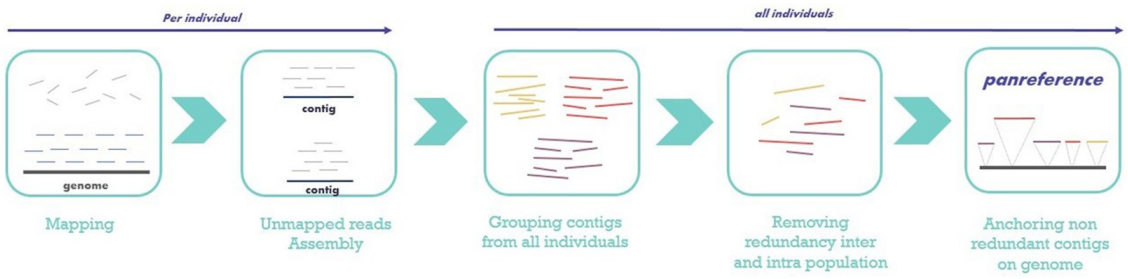

Figure 6: Résumé de l'approche "Map-then-assemble" mise en oeuvre dans FrangiPANe.
Les lectures courtes sont alignées sur le génome de référence, indépendamment pour chaque échantillon, et les lectures non alignées sont assemblées. Ensuite, l'ensemble des contigs sont regroupés pour éliminer la redondance et les contigs non redondants sont ancrés sur le génome. Figure issue de Tranchant-Dubreuil et al. 2023.

Notre méthode et FrangiPANe ont été validées en utilisant les données de séquençage "short reads" du génome de 248 accessions de riz africain cultivé et sauvage (Cubry et al. 2018; Monat, Pera, et al. 2016), comme preuve de concept pour construire le premier pangénome du riz africain. Ces échantillons se composaient de 164 plantes domestiquées et de 84 plantes sauvages, représentatives de la diversité génétique au sein des deux espèces de riz africain *O. glaberrima* et *O. barthii* (Orjuela et al. 2014). L'ensemble du matériel et des méthodes ainsi que tous les résultats sont détaillés dans un article publié dans *NAR Genomics and bioinformatics* (Tranchant-Dubreuil et al. 2023), les principaux résultats sont les suivants :

- Un taux d'alignement des données de reséquençage de type "short reads" contre le génome de référence CG14 (*Oryza glaberrima*) élevé : 96% et 97.8% pour *O. barthii* et *O glaberrima*

respectivement;

- Après assemblage des "reads" non alignés, 8 Mb de séquences ont été obtenus par individu en moyenne, soit un total de 1,65 Gb et 1 306 706 séquences ;
- Après élimination de la redondance, nous avons identifié 513,5 Mb de séquences nouvelles (484 394 contigs) avec une taille de séquence en moyenne de 1 060 pb variant de 301 pb à 83 704 pb;
- 31,5% des contigs non redondants (152 411) ont été placés à une position unique sur le génome de référence (145 Mb ; Figure 7) tandis que 8% (39 630) des contigs ont été placés à des positions multiples (31 Mb) et, enfin, 60,3% des contigs n'ont pas été ancrés.

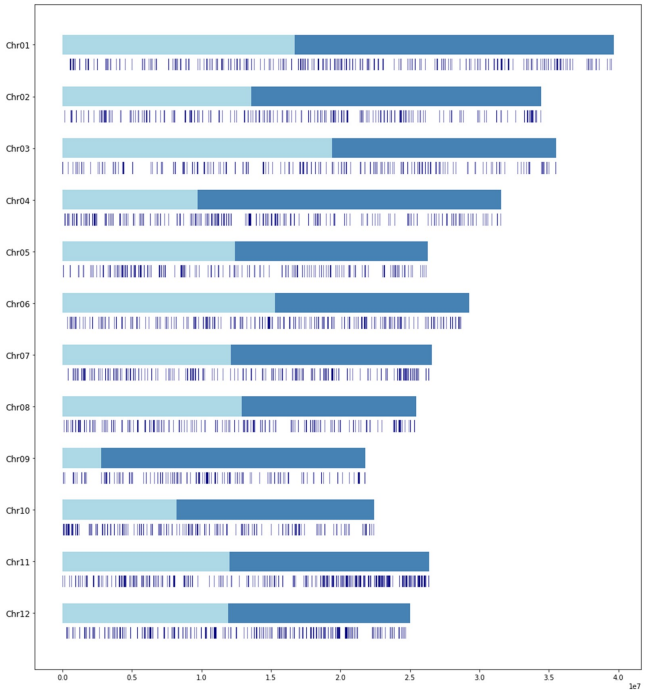

Figure 7: Position des contigs sur les 12 chromosomes de CG14.
(152,411 séquences, 31.5% de l'ensemble des contigs). Figure issue de Tranchant-Dubreuil et al. 2023.

L'approche a été validée avec les données de séquençage "long read" et "short read" du cultivar TOG581. 5 318 contigs (7,9 Mb) ont été ainsi assemblés à partir des données "short read" et un taux de 97,7 % de ces contigs ont été retrouvés, sur le génome assemblé à partir des données "long read", aux positions observées sur le génome de référence.

Nous avons ainsi identifié un total de 515 Mb de nouvelles séquences. Cette part de nouvelles séquences se situe dans la fourchette de valeurs observées chez le riz asiatique, variant entre 268 Mb (Wang, Mauleon, et al. 2018) et 1,3 Gbp (Qin et al. 2021). Si nous comparons notre résultat avec la valeur la plus faible, basée sur une approche similaire réalisée sur 3 010 génomes de riz asiatiques, nous avons trouvé deux fois plus de séquences mais dans notre étude, les espèces sauvages ont été intégrées.

Les séquences assemblées sont enrichies en éléments transposables, avec un taux total d'éléments transposables de 52,1% dans la panréférence incluant le génome et les nouvelles séquences. Ce résultat est similaire au taux de 52,7% observé avec une approche "long reads" mené sur le riz asiatique (Qin et al. 2021). Il serait intéressant d'annoter plus finement ces TEs et de détecter les copies complètes ainsi que leur site d'insertion. Nous avons juste utilisé ici un transfert de l'annotation du génome de riz asiatique, *Oryza sativa japonica* dans ce premier article. L'annotation est une des étapes critiques de la construction d'un pangénome, comme pour un génome de référence. En effet, la majorité des analyses ultérieures seront basées sur cette annotation, à partir de laquelle les gènes seront ensuite classés en "core" et "dispensable", ou des variations structurales seront associées aux gènes, par exemple. Une annotation

de novo a ensuite été effectuée dans le cadre des analyses présentées dans la partie suivante qui s'est concentrée sur l'impact de la domestication sur l'architecture et la dynamique des pangénomes inter- et intra-espèces et, plus généralement, sur l'histoire évolutive du riz africain.

La domestication a remodèlé le pangénome chez le riz africain

Les technologies de séquençage à haut débit ont ouvert la voie à la comparaison des génomes des plantes cultivés avec ceux de leurs parents sauvages. De nombreuses analyses de la diversité ont montré que la domestication a façonné les génomes en réduisant la diversité, principalement la diversité allélique (Huang et al. 2012; Hufford, Xu, et al. 2012; Lin, Zhu, et al. 2014; Qi, Liu, et al. 2013). Cependant, les conséquences de la domestication sur le pangénome n'ont pas encore été examinées, bien qu'il s'agisse de questions émergentes qui ont été récemment abordées dans les dernières revues de pangénomique végétale (Khan et al. 2019; Petereit et al. 2022).

Après avoir développé FrangiPANe et construit le pangénome de riz africain comme preuve de concept (Tranchant-Dubreuil et al. 2023), nous avons ensuite exploré la diversité du pangénome au cours de la domestication du riz africain, en nous concentrant sur la façon dont la domestication a impacté le pangénome du riz africain et comment la sélection a agi sur son organisation et sa dynamique.

L'ensemble des résultats sont décrits dans un troisième article en cours de rédaction, et les principaux résultats sont les suivants:

- 22,765 nouveaux gènes annoté sur les 513,5Mb de nouvelles séquences assemblées, soit un total de 63 318 gènes dans le pangénome;
- le pangénome du riz africain se compose de 64.2% de gènes "core" (présent dans au moins 95% des accessions) et 35.8% de gènes "dispensable" (Figure 8a);
- Nous avons observé une taille du pangénome du riz sauvage et cultivé différente avec au total 60 110 et 57 497 gènes dans *O. barthii* et *O. glaberrima* respectivement. Partageant le même nombre de gènes principaux (environ 39 000 gènes), la différence vient d'un nombre plus important de gènes "dispensable" dans les riz sauvages, c'est-à-dire 2 613 gènes supplémentaires. Cela a conduit à un ratio de gènes "dispensable" de 34,8% et 30,8% dans les accessions de riz sauvages et cultivés (Figure 8b);
- Parmi le génome dispensable, 3 523 et 910 gènes étaient spécifiques à *O. barthii* et *O. glaberrima* respectivement. Les gènes spécifiques aux riz sauvages étaient significativement enrichis dans la liaison des polysaccharides et des hydrates de carbone, ainsi que dans la réponse de défense au stress biotique, tandis que les gènes spécifiques aux riz cultivés étaient enrichis dans des fonctions moléculaires telles que la liaison à la calmoduline ;
- Des gènes ont été identifiés comme "core" dans le pangénome d'une espèce et "dispensable" dans le pangénome de l'autre espèce :
  - 508 gènes "core" dans *O. barthii* ont été identifiés comme "dispensables" dans le pangénome de *O. glaberrima* ;
  - 1 093 gènes "dispensable" dans le pangénome *O. barthii* se sont révélés être des gènes "core" dans le pangénome *O. glaberrima*.

  Parmi ces gènes qui changent de compartiment "core"/"dispensable" selon l'espèce, certains sont significativement enrichis dans des voies telles que le transport de composés azotés, les NAD(P)H, les quinones ou les Peptidyl-prolyl cis/trans isomérases.
- En utilisant la couverture en lecture des 484 394 séquences pour réaliser une analyse en composantes principales (ACP), nous avons montré que notre analyse ACP basée sur les contigs récapitule parfaitement la structuration des individus réalisée précédemment avec les SNP (Cubry et al. 2018). Les accessions sauvages ont été clairement séparées des accessions domestiquées et ont pu être divisées en 3 groupes (Figure 9).
- Sur la base de notre analyse pour la détection des valeurs extrêmes contribuant à la différenciation entre les riz sauvages et cultivés, nous avons identifié 7 579 contigs comme étant putativement

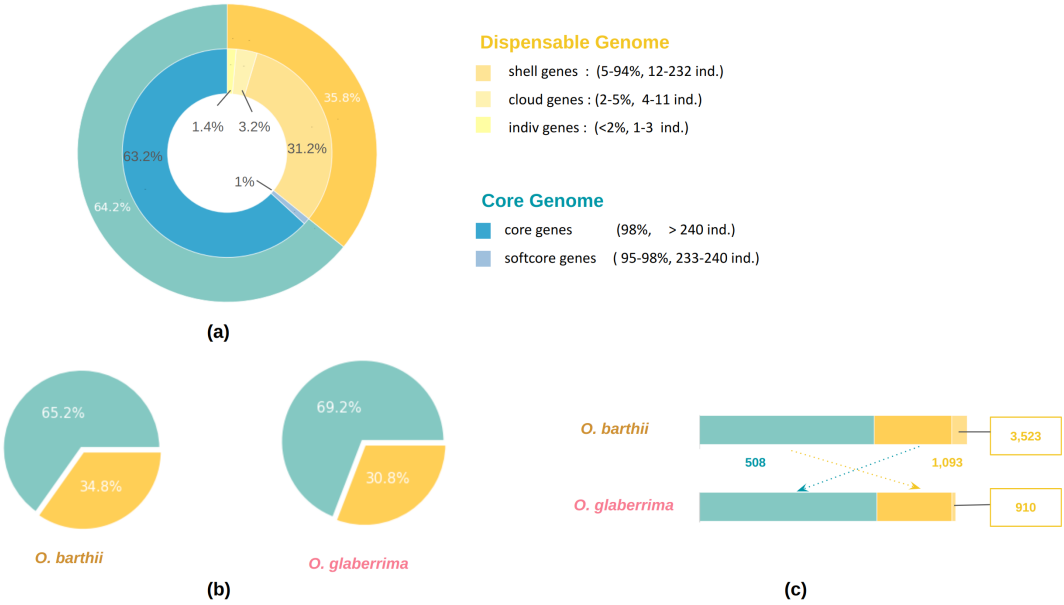

Figure 8: Pangénome du riz africain.
(a) Nombre de gènes dans les 247 accessions de riz domestiqués et sauvages. Le diagramme circulaire extérieur affiche le nombre de gènes "core" (jaune doré) et "dispensable" (bleu). L'anneau intérieur montre la proportion de gènes "core" divisé en gènes "core" et "soft" (couleurs bleues), et de gènes "dispensable" divisés en gènes "shell", "cloud" et individuels (couleurs jaunes). (b) Nombre de gènes "core" et "dispensable" dans les accessions *O. barthii* (à gauche) et *O. glaberrima* (à droite). (c) Dynamique des pangénomes du riz africain sauvage et cultivé. Chaque pangénome est représenté par une barre horizontale affichant les gènes "core" (en jaune) et les gènes "dispensable" (en bleu). Le pangénome de *O.barthii* (en haut) est plus grand que celui de *O. glaberrima* (en bas). Le dernier compartiment, à la fin de chaque barre en jaune clair, indique le nombre de gènes spécifiques à chaque espèce, soit 3 523 et 910 gènes présents uniquement dans *O. barthii* et *O. glaberrima* respectivement. La transition des gènes entre les compartiments "core" et "dispensable" des pangénomes des deux espèces est représentée par les flèches en pointillés : 508 gènes sont "core" dans *O. barthii* et "dispensable" dans *O. glaberrima*, tandis que 1 093 gènes "dispensable" dans *O. barthii* sont "core" dans *O. glaberrima*.

sélectionnés au cours de la domestication, dont 683 correspondent potentiellement à de nouveaux gènes (approche PCAdapt).

- Une signature de sélection a été trouvée pour le gène *PROG1*, un gène déjà connu pour avoir été sélectionné au cours de la domestication du riz africain et asiatique, associé à une délétion majeure au cours de la domestication du riz africain. Dans notre analyse, *PROG1* faisait partie des gènes spécifiques des riz sauvages, comme attendu.
- En plus d'une variation significative du nombre de gènes, nous avons également observé que pour les variations structurales correspondant à nos 152 411 contigs placés sur le génome de référence, ces contigs sont placés dans un 30% des gènes du génomes et ce ratio atteint 70% des gènes si nous prenons en compte les gènes ainsi que la séquence flanquante de 5kbp.

Les premiers résultats de notre dernière étude montrent le potentiel des approches pangénomiques pour explorer la diversité au sein d'une céréale, le riz africain, et la manière dont cette diversité a été remodelée au cours de sa domestication. Nous commençons à avoir un premier aperçu du rôle que ces variations structurales ont joué dans la composition en gènes et l'adaptation, et de la manière dont les forces évolutives ont façonné l'organisation et la dynamique du (pan)génome.

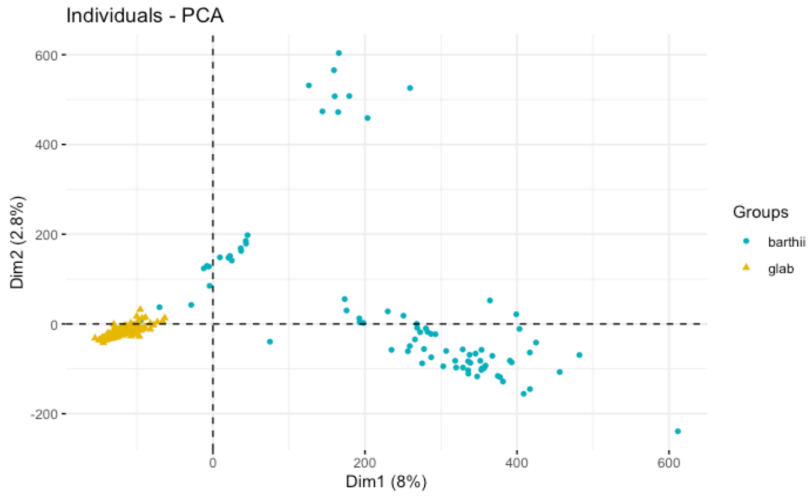

Figure 9: Analyse en composantes Principales des 484,394 nouvelles séquences absentes du génome de référence *Oryza glaberrima*.

Pour conclure

Bien que très prometteuse, la pangénomique a encore de nombreux défis à relever pour réussir la transition d'une approche émergente à une approche plus classique intégrée aux études de génomique et de génétique. Afin d'exploiter les données de séquençage "longs reads" (et aussi "short reads"), une nouvelle approche basée sur les graphes a émergé récemment pour construire un pangénome sous la forme d'un graphe (Figure 10), qui va intégrer toutes les variations au sein d'une population (Garrison et al. 2018; Li, Feng, and Chu 2020; Sirén et al. 2021).

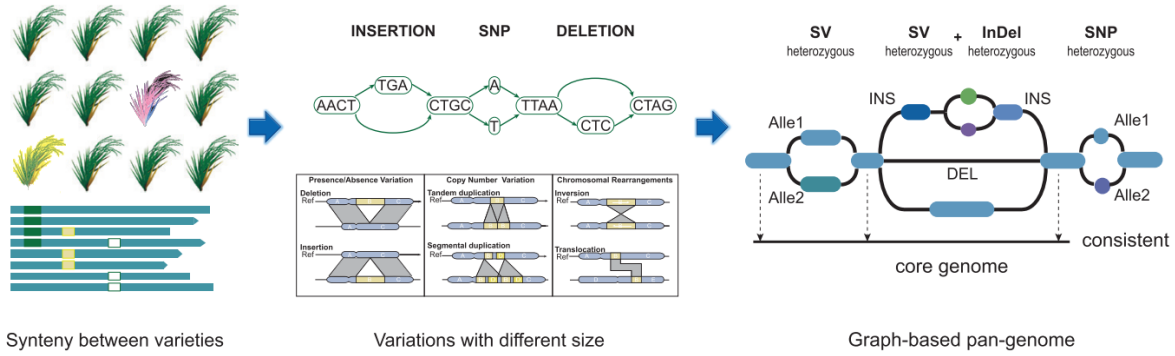

Figure 10: Ensemble des processus pour construire un graph de pangénome.
(à gauche) Sélection des variétés représentatives de la diversité et comparaison de leur génome séquencé. (au milieu) Détection des variations structurales. (à droite) Construction du pangénome basé sur un graphe. Figure issue de Wang, Qian, et al. 2023.

Combinés aux technologies de séquençage à haut débit de troisième génération, les graphes de pangénome sont très prometteurs, même s'ils n'en sont qu'à leurs balbutiements. Ils devront être testés sur de nombreux génomes et à plus grande échelle, pour que ces analyses soient standardisées et puissent être lancées dans n'importe quel laboratoire de recherche. D'autres défis seront par exemple liés à l'annotation de ces (pan)génomes, qu'il s'agisse d'éléments transposables ou de gènes, et plus largement à l'intégration, dans les analyses pangénomiques, de toutes les autres informations biologiques disponibles (ex: RNAseq,

méthylome, phénotype, épigénome).

Plus de 20 ans après le séquençage du premier génome de plante, la pangénomique ouvre une nouvelle ère, prometteuse avec de nombreux défis à relever. Elle peut aussi nous amener à repenser nos connaissances des processus évolutifs tels que la domestication en comprenant, par exemple, pourquoi des gènes ont été potentiellement perdus, quelles étaient leur fonction et quel est l'impact de cette variabilité sur la régulation des gènes. Plus largement, elle peut être l'occasion de pousser les scientifiques à sortir des sentiers battus pour approfondir et repenser l'ensemble des connaissances acquises ces dernières années.

Remerciements

Traitez les gens comme s'ils étaient ce qu'ils devraient être, et vous les aiderez ainsi à devenir ce qu'ils peuvent être – Johann Wolfgang Goethe

C'est sûrement une des pages les plus difficiles à écrire... comment trouver les mots "justes" pour exprimer toute la gratitude que je ressens et remercier toutes les personnes qui m'ont permis de croire que je pouvais faire une thèse et celles qui ont fait partie de près ou de loin de cette aventure... ce n'est pas une mince affaire...

Tout d'abord, je voudrais remercier mes deux boss qui m'ont guidée et accompagnée tout au long de ce projet doctoral un peu particulier, réalisé à temps partiel, en parallèle de mes activités d'ingénieure. Comme autorisé par l'école doctorale, j'aurai pu réaliser cette thèse en six ans tellement j'ai bien vécu cette aventure grâce à vous deux! Vous y avez échappé de peu! Merci à vous deux d'avoir fait de cette ~~folie~~ aventure professionnelle, une belle aventure humaine ~~au moins pour moi~~... Vous m'avez aidée à maintenir ce fragile équilibre entre mes activités d'ingénieure et de doctorante et surtout à garder le cap... Je n'aurai pu y arriver sans vous. J'ai compté, plus de 150 heures de réunions hebdomadaires, pendant lesquelles nous avons discuté et où vous m'avez ~~torturée~~ questionnée afin de développer ~~cet obscur~~ ce fameux questionnement scientifique! Je ne sais pas si j'ai atteint tous les objectifs mais ce dont je suis sûre est que j'ai immensément appris durant ces 4 années.

Merci à Yves Vigouroux d'avoir accepté de co-encadrer cette thèse alors qu'on se connaissait peu, pour son soutien, sa disponibilité, sa pédagogie pour m'initier ~~au côté obscur,~~ aux statistiques et à la génétique des pop, et sa rigueur scientifique sans faille dans l'analyse et la discussion des résultats. Et merci pour ton humour !

Merci à François Sabot... Il y a des rencontres professionnelles (et humaines) qui sont décisives dans sa carrière... Merci de m'avoir soutenue depuis fort, fort longtemps, d'avoir cru en moi et pour toutes ces idées et projets à foison... comme celle de faire cette thèse sur la pangénomique des riz africains ! Tu m'as appris à ~~dire non~~ optimiser mon temps face à tous nos projets d'équipe passionnants... Toujours enthousiaste et disponible pour des discussions scientifiques (mais pas que), la fin de cette thèse me fait encore plus réaliser le plaisir de travailler avec toi depuis toutes ces années ! comme quoi le travail peut être sérieux, productif ~~j'espère~~ tout en étant fun et cool...

En résumé, même si cette aventure se termine, je compte bien continuer à ~~vous embêter~~ travailler avec vous dans le cadre de nouvelles aventures !

Merci à Alain Ghesquière d'avoir accepté, quand il était DU de l'UMR DIADE, que je réalise ce doctorat... de m'avoir fait confiance, ~~en 2002~~ il y a quelques années, et permis de commencer mon premier contrat à l'IRD... qui sera suivi d'un concours en 2004 que j'aurai... bref... je mesure tout le chemin parcouru grâce à toi... et je te remercie d'avoir accepté de faire partie de ce jury de thèse...

Merci à Serge Hamon, Alexandre de Kochko de leur appui et de m'avoir encouragée à réaliser un doctorat tout en étant ingénieure... Il m'aura fallu quelques années pour passer le cap mais l'idée était lancée ! Merci à Laurence Albar pour tous ses encouragements, son soutien et de m'avoir aidée à ne pas trop dévier du cap quand elle était la boss de l'équipe Rice... Merci à Jérôme Duminil de m'avoir accueillie dans l'équipe Dynadiv !

Je voudrais aussi remercier les membres de mes 3 comités de suivi de thèse : Hervé Etienne, Maud Tenaillon, Clémentine Vitte et Vincent Ranwez. Pareil que pour mes boss, j'aurai pu faire 6 années de thèse tellement j'ai apprécié nos discussions et échanges...

Un grand merci chaleureux à toute la team I-trop...I-trop powwweeeer!!!!!!! Aurore, Valérie, Alexis, Bruno, Ndomassi et Julie!!! C'est un plaisir de travailler à vos côtés ! Promis, Julie, je sors de ma "grotte" après la soutenance et "il faut qu'on cause" !!!

Un grand merci à tous les membres de l'équipe Dynadiv de m'avoir accueillie dans un contexte particulier entre ce projet doctoral et la covid... sans aucune corrélation... Je vais bientôt ne plus me cacher dans mon bureau... promis ! Un spécial merci à Marie Couderc qui, en plus d'avoir bossé comme une dingue sur le séquençage de génomes de riz, a aussi réalisé de nombreuses analyses bioinformatiques dans le cadre de ce projet ! Promis, on va les regarder de plus près ces données ONT!!! Un merci aussi à Philippe Cubry avec qui j'ai partagé les ~~nombreuses~~ galères lors des analyses SNPs... Et oui faut arrêter GATK ! et oui va falloir qu'on le dépose ce projet !

Merci aux étudiants en thèse et postdoc pour tous ces chouettes petits moments de partage... Marine, Thomas, Fabrice, Tram, Stella... et au petit minion violet ~~Kevin~~ Eloi, pour tous ces moments ~~de pétage de plombs et de fous rires~~ qui ont permis de décompresser...

Merci à Ezéchiel, Romaric, Fidèle, pour leur soutien de loin mais de près aussi...

Un grand merci à tous mes collègues de l'UMR DIADE et de la plateforme South Green pour leur soutien durant ces années de thèse... Merci à Maître Python (il se reconnaîtra ;))... A Mathieu Rouard pour sa bienveillance et en autre la rédaction de la revue pangénomique... ça a été un plaisir et il va vraiment falloir qu'on collabore sur un projet un jour... Merci à Cécile Triay de m'avoir supportée en binôme agile git et en formation python... et aussi pour sa gentillesse et sa bonne humeur...

Merci à mes amis et ma famille qui ont suivi ce projet de loin, qui ont vraiment essayé de comprendre ce que je cherchais toutes ces années... progressivement, vous avez arrêté de me demander sur quoi portait mon projet de thèse, j'ose espérer que j'ai été assez claire à un moment. Petit mot spécial pour Orianne, ma ~~seule~~ nièce préférée ! Un grand merci à Ienke pour tous ces encouragements, les cours d'anglais, les séances de coaching, les repas, les sorties bateau, la Corse, sa bonne humeur...

Merci à mes parents, pour leur immense amour, pour cette enfance merveilleuse sous les "tropiques" et ailleurs... ils m'ont tellement appris... notamment que tout était possible dans la vie avec le travail et le respect des autres! Merci à mon frère pour m'avoir ~~supportée~~ soutenue !

Et enfin, je terminerai par mes deux hommes... merci à Ilan d'avoir supporté une maman qui ~~s'amuse~~ travaille à la maison, sur un projet qui aura duré une partie de ton collège et de ton lycée... Ouf, j'aurai fini cette thèse avant que tu passes le BAC!

Et un immense merci à toi, Arnaud... On partage notre vie depuis un petit bout de temps... et comme on aime bien les challenges et pour ne pas s'endormir dans la routine, on a décidé ces 4 dernières années de vivre plein de choses en parallèle... une luxation du genou dans notre maison à étage, une pandémie, deux déménagements, les "choses" de la vie, les travaux de l'ancienne maison, des concours, les travaux de la nouvelle maison et accessoirement cette aventure... Je n'aurai jamais pu y arriver sans toi, ton soutien permanent, tes ~~raleries~~, ton 200% à la maison, ta bonne humeur et ton énergie... merci d'être dans ma vie...

Contents

List of Figures

List of Tables

Introduction

1- Domestication

What is not started will never get finished – Johann Wolfgang von Goethe

The domestication of plants and animals began 12,000 to 11,00 years ago (Fuller et al. 2014; Larson et al. 2014), a very recent and short period of time in the 200,000 years of human life. This process has revolutionized our lives, shaped our societies and natural landscapes. It provides most of our current food and was a prerequisite for the development of our agricultural system and, more broadly, of our cities and civilization.

Advances in archaeology and genetics offer insight into the domestication process addressing many questions about domestication such as :

- When, where and how often did domestication take place?
- Which species or wild population did the cultivated plant come from?
- Which phenotypic changes occurred during domestication and at which rate?

1.1 What is domestication?

Defining domestication is not as straightforward as one might think (Purugganan 2022), and many definitions have been proposed, more or less focused on human, such as :

- A first anthropocentric definition emitted by Doebley, Gaut, and Smith (2006):

  > *"Plant domestication is the genetic modification of a wild species to create a new form of plant altered to meet human needs."*

- A non-human-centered one used by Fuller et al. (2014):

  > *"It is a process of speciation and/or species transformation that occurs when one species (the domesticator) begins to control the reproduction and dispersal of another species (the domesticated) in order to meet the needs of the former, most notably (but not exclusively) for food."*

- A broader biological definition of domestication proposed by Purugganan (2022):

  > *"It is a coevolutionary process that arises from a mutualism, in which one species (the domesticator) constructs an environment where it actively manages both the survival and reproduction of another species (the domesticate) in order to provide the former with resources and/or services."*

The two latter definitions could be used to describe agriculture developed independently in other species than human. Examples of domestication in the animal kingdom have been studied, notably in at least three orders of insect : the ambrosia beetles, the fungus-growing termites and the fungus-growing ants (Mueller et al. 2005). The cultivation of fungi by attine ants (*Myrmicinae subfamily*) is one of the best studied insect-associated domestications. Ants plant and cultivate fungus: as any farmer, they manage their growing conditions by regulating temperature and humidity, while taking care to protect them from other herbivores, pests and diseases, and fertilizing their fungal gardens. Finally, they harvest the cultivated mushrooms for food (Schultz et al. 2005).

Overall, domestication enhances the survival and fitness of the domesticator, but this system of co-evolution cannot be considered one-sided, as it also increases the fitness of the domesticated relative

to its wild parent by increasing the size of the population over generations, as well as its geographical dispersal outside its area of origin (Mueller et al. 2005; Schultz et al. 2005). So, domestication can be considered as a beneficial process for both the domesticator and the domesticated.

In conclusion, domestication can be considered rather a plural than a univocal concept. It can be sometimes difficult to determine the limits of the domestication, i.e. whether a species is truly domesticated, or in the process of being domesticated, or simply the result of a commensal relationship or beneficial association if the species is considered to be able to survive without human assistance. Furthermore, it can be considered that there are different forms and degrees of domestication, ranging from fewly domesticated plants to semi-domesticated crops with several agronomic adaptations such as flax or olive to fully domesticated crops that are completely dependent on humans, such as maize or barley (Stetter et al. 2017).

1.2 Plant domestication

Of the more than 250,000 species of angiosperms worldwide, only about 1% have been partially or fully domesticated (Dirzo and Raven 2003), spread over a third of the 500 families of flowering plants. Two families group together the largest number of cultivated species, i.e. 30%, the grass family (Poaceae) and the legumes (Fabacaeae), with 379 and 337 domesticated species respectively (Dirzo and Raven 2003). Of all crop plants, only 103 provide 90% of the world's food energy intake, with rice, maize and wheat accounting for two-thirds. Other staple foods for humans are millet and sorghum, tubers such as potatoes, cassava, yams, and taro (Dirzo and Raven 2003). Besides being a source of food, more than 15 plants are cultivated as sources of fiber, and thousands more as ornamental plants or sources of medicine (Dirzo and Raven 2003).

Domestication began 12,000 to 11,000 years ago, at the end of the Pleistocene, the most recent glacial period, during the transition to the Holocene, the current interglacial period (Fuller et al. 2014) (Figure 1.1, page 26).

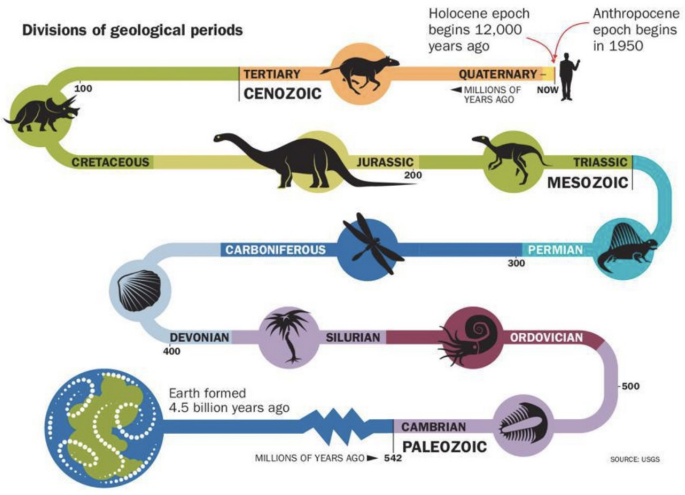

Figure 1.1: The Earth timeline.
The 4.6 billion years are divided into Eons, Eras, Periods, Epochs and Ages. Earth's current epoch, the Holocene, began at the end of the last ice age (the Pleistocene epoch), about 12,000 years ago. This era is part of the Quaternary Period, which is part of the Cenozoic Era. The figure shows only the eras of the present (fourth) Eon, the Phanerozoic, which began about 540 million years ago. Figure from U.S. Geological Survey.

The reasons for this shift from hunting and gathering to agriculture are still debated: is it due to the increase of the human population, to climate change, or to new ways of life? In any case, this transition

to agriculture is associated with fundamental changes in human life such as the birth of settlements (prerequisite for the development of cities). Domestication occurred separately on different continents and in different cultural traditions at the same time (Figure 1.2, page 27).

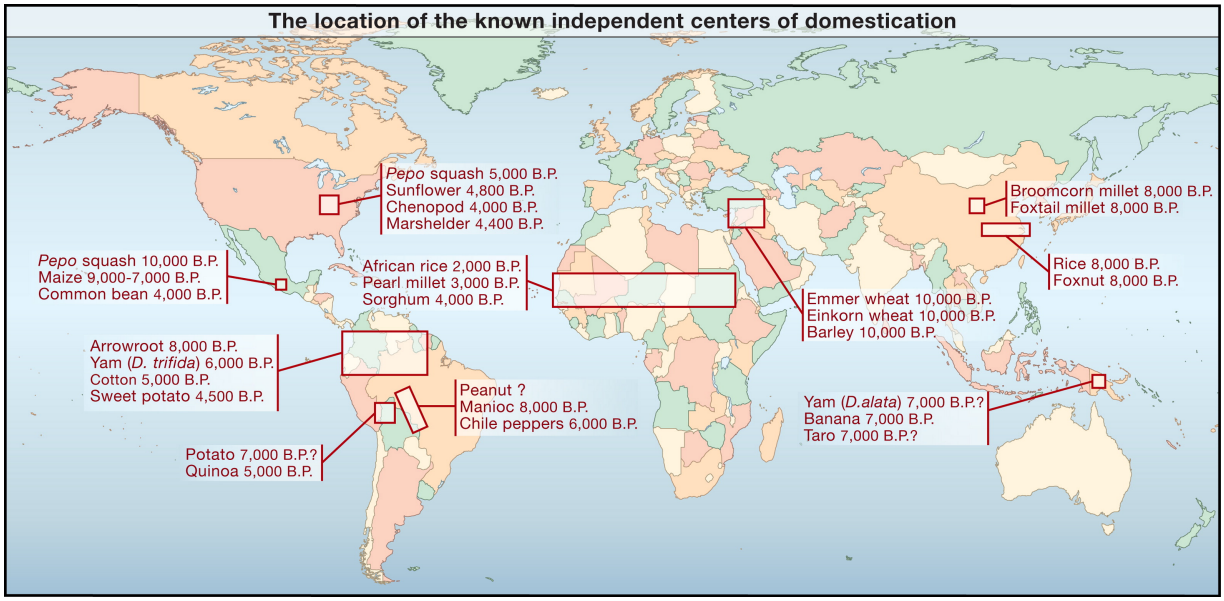

Figure 1.2: The independent centers of domestication.
For each region, principal crop plants and estimates of when they were brought under domestication based on available archaeological evidence are shown. Figure from Doebley, Gaut, and Smith 2006.

Before addressing the question of the duration of the domestication process, we will see what can differentiate a domesticated plant from its wild ancestor at the phenotypic and physiological level, and what the domestication syndrome is.

1.3 Domestication syndrome

In his book "*The Variation of Animals and Plants Under Domestication*", Charles Darwin described behavioral, morphological, and physiological traits shared by domesticated animals, but not by their wild ancestors, such as docility, floppy ears, modified tails, original coat colors and patterns, smaller brain size and smaller tooth (Darwin 1868; Wilkins, Wrangham, and Fitch 2014). These shared traits became known as the domestication syndrome (Wilkins, Wrangham, and Fitch 2014).

The syndrome has been extended to plants to similarly define a series of traits common to domesticated plants that distinguish them from their ancestors (Hammer 1984). These physiological and phenotypic changes were targets of the human selection (Doebley, Gaut, and Smith 2006) (Figure 1.3, page 28), and, depending on the species, may include:

- More robust plants overall
- Robust growth of the central stem relative to the lateral stems due to increased apical dominance
- Bigger fruits or grains but fewer per plant
- Bigger seeds
- A reduced capacity to disperse seeds by retaining them more on the plant (noshattering)
- Reduction in seed dormancy, more controlled germination
- Changes in photoperiod sensitivity, and synchronized flowering
- A decrease in bitter substances in edible fruits

,

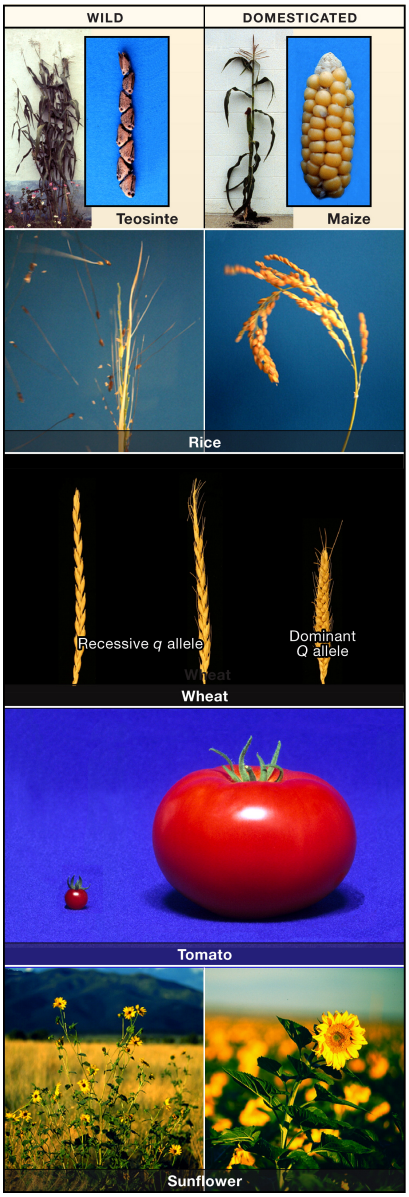

Figure 1.3: Phenotypes of some crops and their progenitors.
(Top row) A plant of the maize progenitor, teosinte, with multiple stalks and long branches (left), a plant of cultivated maize with its single stalk (right). A teosinte ear with its grain (not visible) enclosed in the triangular casing that comprises the ear (inset left), a maize ear bearing its grain naked on the surface of the ear. (Second row) Wild rice with a panicle that shatters (left), cultivated rice with a solid panicle of grain (right). (Third row) Cultivated wheat with the dominant allele of the Q gene and a condensed and tough spike (right), Cultivated wheat with the recessive allele q (center) and wild wheat (left) with the recessive allele and slender, fragile spikes. (Fourth row) The massive fruit of cultivated tomato (right), the minuscule fruit of its progenitor (left). (Fifth row) A wild sunflower plant with many small heads borne on multiple slender stalks (left), a cultivated sunflower plant with a single large head borne on a thick stalk (right). Figure from Doebley, Gaut, and Smith 2006.

Initially, several studies highlighted the domestication as a relatively rapid process with the fixation of the domestication traits over a few hundred years as the result of human selection (Abbo, Lev-Yadun, and Gopher 2010; Hillman and Davies 2008). Several archaeological studies have subsequently shown that domestication traits turns out to be much longer and more complex than originally thought, resulting from a continuum of relationships between humans and plants over a long period (Figure 1.4, page 30).

Domestication has been described as a 3-step process (Weiss, Kislev, and Hartmann 2006):

1. The annual gathering of grains and fruits of wild species available in the natural environment;
2. The cultivation of individuals, which are sown in fields. Initially, only wild plants are cultivated and then the proportion of cultivated individuals increases due to both conscious and unconscious selections according to the selected traits;
3. The domestication with the cultivation of crops that present interesting phenotypes and that are selected and improved at each new generation.

Thus, the non-shattering phenotype has been fixed over several millennia (from 1,000 to 4,000 years) in several cereals species such as wheat, oat, rice or barley (Fuller et al. 2014; Tanno and Willcox 2006; Tanno and Willcox 2012; Weiss, Kislev, and Hartmann 2006). Several explanations can therefore be put forward to explain why the domestication was a protracted process spread over several millennia according to the species (Purugganan 2022):

- A recurrent gene flow between wild and plant under human selection (Allaby 2010);
- A more or less strong human selection resulting from a conscious but also unconscious selection (Darwin 1859; Darwin 1868);
- Polygenic traits influenced by multiple genes with little effect and which take longer to become fixed (Stetter et al. 2017).

1.4 Tracing domestication in genome

Numerous genetic studies have examined the impact of human selection on the genetic diversity of crop species by comparing the genomes of crop plants and of their wild ancestors (Berger et al. 2012; Eyre-Walker et al. 1998; Hyten et al. 2006). They provided a better understanding of how domestication has "tinkered" with a crop both by filtering out some alleles from the standing allelic variations of the ancestors and by selecting new mutations responsible for interesting physiological or phenotypic traits.

By using a limited number of initial wild individuals, only a fraction of the diversity was captured and retained (founder effect). This diversity is also expected to be further reduced if humans use a limited number of seeds for each new generation. The reduction of this diversity within the genome is called the domestication genetic bottleneck (Doebley, Gaut, and Smith 2006), more or less severe depending on different factors such as the population size, the length of time between the beginning of domestication and its full establishment, and gene flows.

Moreover, this loss of diversity can be different along the genome with, for instance:

- A strong decrease in diversity within genes conferring an advantageous trait corresponding to an increase in the frequency of the favored allele and a decrease in the other alleles;
- A higher level of diversity for the so-called neutral genes depending on the impact of genetic drift and population size (Doebley, Gaut, and Smith 2006, Figure 1.5, page 31).

Genes controlling domestication traits or varietal differences have been identified in different species (Table 1.1, page 32). It has been observed that traits associated with domestication or post domestication selection result mainly from mutations in regulatory genes such as transcription factors and enzymes (Meyer and Purugganan 2013). In addition, several mutations have been found in a large number of these genes, offering different alternative targets for selection. Loss-of-function and altered gene expression are the two main causes of change, mainly due to nonsense mutations but also cis-regulatory and missense mutations (Meyer and Purugganan 2013). Two traits can be used as examples to illustrate how domestication and selection have acted on the same genes or gene networks but in different ways in different species, here the African and Asian cultivated rice (Cubry et al. 2018):

- The transition from prostrate to erect growth is controlled by the *PROG1* gene in both two rice species. In Asian rice, this gene has one mutation causing a loss of function of this gene whereas it is absent in the cultivated African rice (but present in the wild relative species);

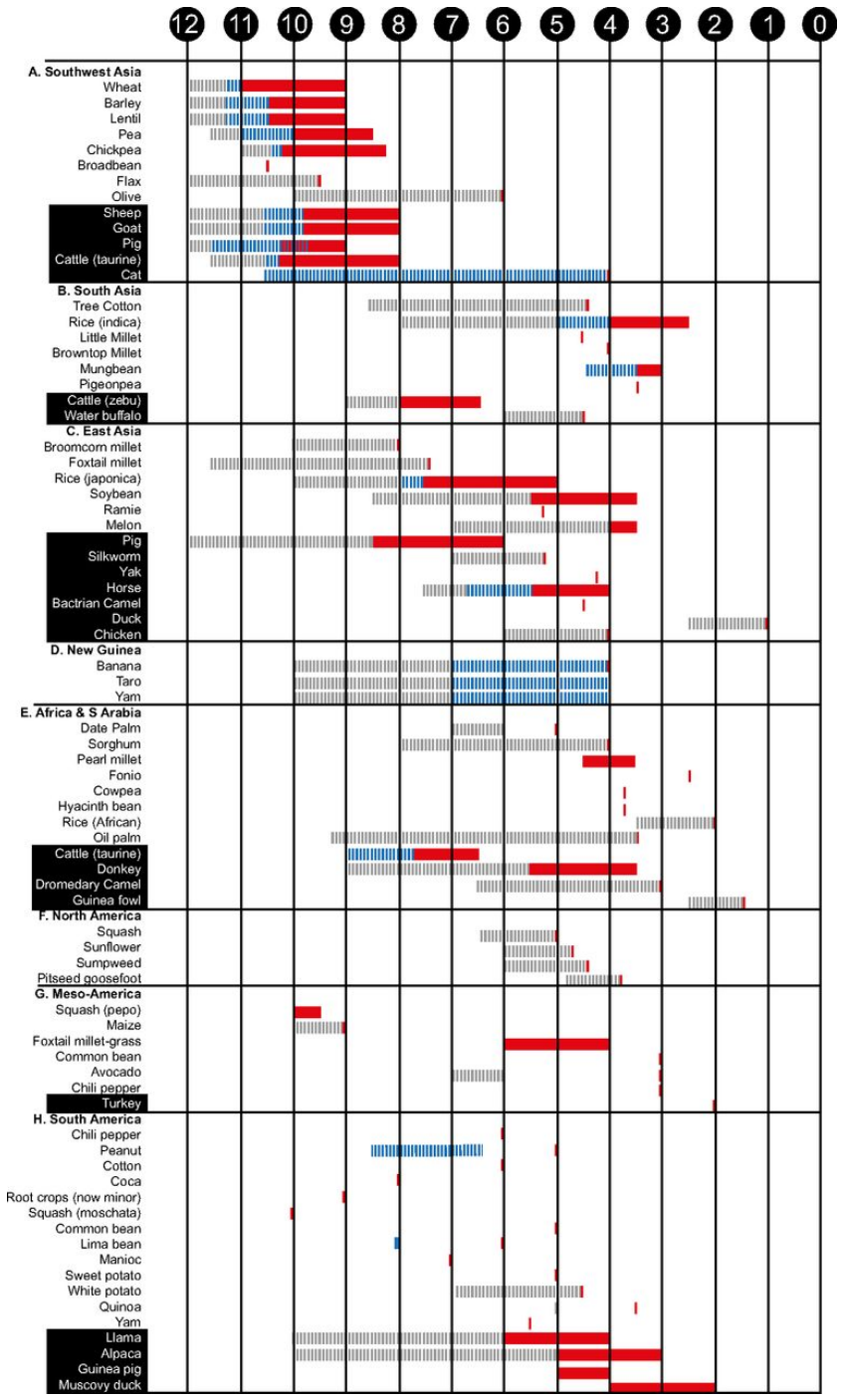

Figure 1.4: A chronological chart of key plants and animals domesticated.
This figure lists the regions where, and the time frames over which, key plants and animals were domesticated. The numbers in the black circles represent thousands of years before present. Gray dashed lines represent documented exploitation before domestication or posited as necessary lead-time to domestication. Blue dashed lines represent either the management of plants or animals or predomestication cultivation of plants. Red bars frame the period over which morphological changes associated with domestication are first documented and a short, solid red bar represents the latest time by which domestication occurred. Figure from Larson et al. 2014.

- The dehiscence trait is controlled by different genes on which selection has acted in a similar gene regulatory network in the two Rice species, but in different ways (Figure 1.6, page 31).

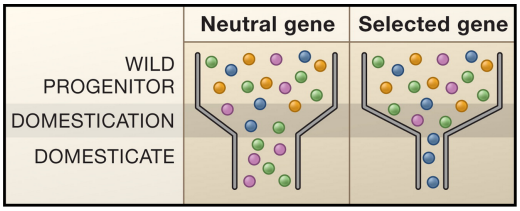

Figure 1.5: The effects of the domestication bottleneck on genetic diversity.
(Left) Population bottlenecks are a common important demographic event during domestication. Genetic diversity is represented by shaded balls; the bottleneck reduces diversity in neutral genes, as shown by the loss of the orange and blue variants. (Right) Selection decreases diversity beyond that caused by the bottleneck, as shown by the loss of all but one genetic variant in the domesticated species. Note, however, that an exceptionally strong domestication bottleneck could leave little variation in neutral genes. In that case, it may be very difficult to distinguish selected from neutral loci. Figure from Doebley, Gaut, and Smith 2006.

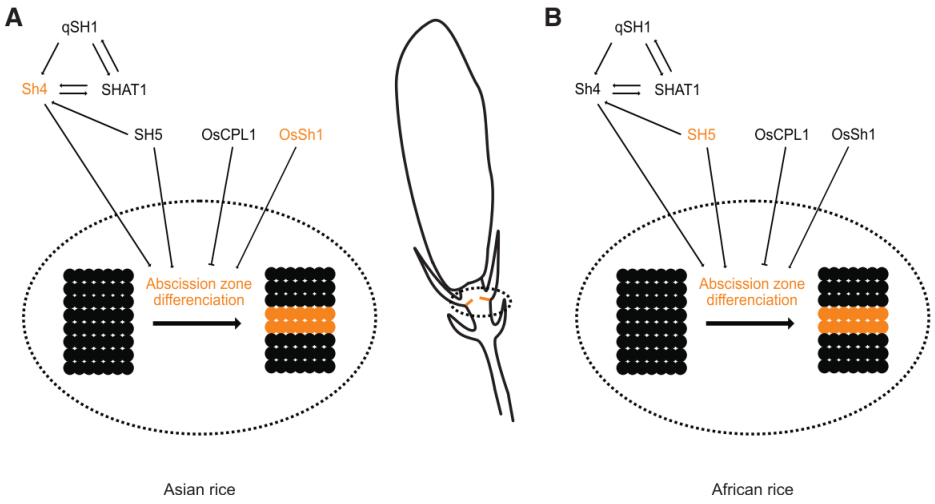

Figure 1.6: Regulatory gene network in Rice involved in abscission zone differentiation.
The currently accepted model of abscission describes a first step that corresponds to tissue differentiation defining the abscission zone (AZ). Genes showing evidence of selection in Asian rice (A) or in African rice (B) are colored in orange. O. sativa (A) Sh4 carries a single non-synonymous substitution leading to partial function of the AZ, which is fixed in all cultivated Asian rice. OsSh1, a YABBY transcription factor underlies a minor QTL in rice but was a target of selection during rice domestication. qSH1 carries a causative mutation located 12 kb upstream of the BEL1-type homeobox gene, thus decreasing its expression level and interfering with the development of AZ in temperate japonica varieties only. For three other known shattering genes, SHAT1, OsCPL1, and SH5, no evidence of selection during the domestication process of Asian rice is known. In African Rice (B), evidence of selection during the domestication process was found only on SH5. In addition, OsSh1 is absent from the genome of some *O. glaberrima* individuals, whereas it is present in all individuals of the wild relatives. Figure from Cubry et al. 2018.

Crop	Gene	Molecular Function	Causative Change	Ref.
GENES IDENTIFIED AS CONTROLLING DOMESTICATION TRAITS				
Plant and inflorescence structure				
Maize	*Teosinte branched1 tb1*	Transcriptional regulator (TCP)	Regulatory change	Wang, Stec, et al. 1999
Wheat	*Q*	Transcriptional regulator (AP2)	Regulatory/ amino change	Simons et al. 2006
Abscission layer formation, shattering				
Rice	*qSH1*	Transcriptional regulator (homeodomain)	Regulatory change	Konishi et al. 2006
Rice	*sh4*	Transcriptional regulator (Myb3)	Regulatory/ amino change	Li, Zhou, and Sang 2006
Wheat	*Q*	Transcriptional regulator (AP2)	Regulatory/ amino change	Simons et al. 2006
Fruit weight				
Tomato	*Fruit weight 2.2 fw2.2*	Cell signaling	Regulatory change	Frary et al. 2000
Seed casing				
Maize	*Teosinte glume architecture*, *tga1*	Transcriptional regulator (SBP)	Amino acid change	Wang, Nussbaum-Wagler, et al. 2005
Seed color				
Rice	*Rc*	Transcriptional regulator (bHLH)	Disrupted coding sequence	Sweeney et al. 2006
GENES IDENTIFIED AS CONTROLLLING VARIETAL DIFFERENCES				
Grain number				
Rice	*grain number 1*, *gn1*	Cytokinin oxydase/ dehydrogenase	Regulatory/early stop codon	Ashikari et al. 2005
Fruit shape				
Tomato	*ovate*	unknown	Early stop codon	Liu, Van Eck, et al. 2002
Sticky grains				
Rice	*waxy*	Starch synthase	Intron splicing defect	Olsen and Purugganan 2002; Wang, Zheng, et al. 1995

Table 1.1: List of some genes of interest in crop domestication and improvement. Table adapted from Doebley, Gaut, and Smith 2006.

2- Genetic diversity shaped by evolutionary processes

It is not the strongest of the species that survives, or the most intelligent; it is the one most capable of change – Charles Darwin

In order to explain the great phenotypic diversity between individuals of the same species, the theory of evolution proposed by Charles Darwin is based on two main ideas: numerous heritable variations between individuals of the same species and the action of natural selection on these variations (and hence on individuals). In a given population, each individual is a unique combination of traits, and individuals with variations that currently confer an advantage in a given environment will increase their chances of survival and reproduction within the population (resulting in passing the advantageous variations to the next generation) and thus trait spending (Darwin 1859).

Since then, the theory of evolution has continued to be enriched and refined as knowledge has advanced combined with technological progress. And it has generated over time a large number of new questions such as:

- How to measure diversity within a population or, in other words, how to estimate the rate of change within a species?
- What are the mechanisms generating these changes in individuals?
- What is the impact of factors such as population size on this variability? Is selection the only force acting on diversity?

First of all, what type of variations can be found in DNA and what are the mechanisms behind these variations?

2.1 A large range of variations in DNA

Genetic diversity can be observed at the DNA level through a broad continuum of mutations from single base variation (SNPs), small insertions-deletions (indels, $<$ 50bp) to larger structural variations (SVs). These latter includes a multitude of types such as duplication, inversion, translocation, insertion/deletion, transposition or Copy Number Variation (CNV) and Present Absence Variation (PAV) (Gabur et al. 2018). Structural variations can involve genomic regions of several kilobases or megabases, even at the chromosomal scale shaping the individual genome they impacted.

The mechanisms by which structural variations arise are various. They can be generated by errors in cellular mechanisms such as DNA replication or recombination, or even by misrepair of DNA following DNA strand breaks (Table 14.1 in Appendix, page 147, Gabur et al. 2018; Saxena, Edwards, and Varshney 2014). Changes in ploidy in plants can also cause structural variations. Studies of the evolutionary history of many angiosperms have shown ancient or recent polyploidization and/or whole-genome duplication (Alix et al. 2017; Jiao, Wickett, et al. 2011; Van De Peer, Maere, and Meyer 2009). In addition, the mobilization of transposable elements can also contribute significantly to structural variations.

Transposable element, a main actor shaping the structure and the size of plant genome

A large part of the variation in plant genome corresponds to repetitive DNA mainly due to the activity of transposable elements (TEs) (Grzebelus 2018). TEs, also called mobile elements or mobile DNA, are self mobilized DNA sequences that are able to amplify and move into the genome using host cell

machinery to express their own genes. They can be classified into 2 classes according to their mechanism of amplification and proliferation:

- Class I or retrotransposons using a RNA intermediate to transpose through a copy and paste mechanism and leading to the original/donor TE remained to its initial position and a new copy integrated in a new site;
- Class II or DNA transposons using a DNA intermediate to transpose through a copy and cut mechanism.

Every TE class is divided into orders, then into superfamilies, according to the classification proposed by Wicker et al. 2007 (Figure 14.3 in Appendix, page 148). In plants, the two superfamilies *Copia* and *Gypsy* superfamilies belonging to the LTR retrotransposons order, encompass the majority of all plant TEs. Within the class 2, most plant TEs are members of the TIR order, divided into five superfamilies including *hAT*, *Mutator*, *CACTA*, *PIF/Harbinger* and *Tc1-mariner* (Grzebelus 2018).

With a more than 2,400-fold genome size variation in seed plants (angiosperms and gymnosperms), the major difference is mainly due to the TE content (Novák et al. 2020). This latter ranges, in plant genomes, from 3% in the tiny 82-megabase genome of carnivorous plant *Utricularia gibba* (Ibarra-Laclette et al. 2013) to more than 85% in large plant genomes such as wheat (Appels et al. 2018). Due to their transposition mechanisms, class 1 elements are highly abundant in genomes and associated with large plant genome such as in the 17 Gb wheat genome, in which LTR retrotransposons account for over 65% of the TE content (Appels et al. 2018) (Figure 2.1, page 35).

From junk and selfish DNA to regulatory elements

TEs were called *"controlling elements"* by Barbara McClintok, who first described them 76 years ago in maize (McClintock 1947). For a long time defined only as junk or selfish DNA (Ohno 1972, Orgel and Crick 1980), TEs are now also known as regulatory elements that can impact on the expression of their host's genes, and are even considered as domesticated and exapted elements to serve the host (Feschotte 2008, Jangam, Feschotte, and Betrán 2017). Thus, they can inactivate a gene by inserting themselves into it, such as the insertion of the LRT retrotransposon *Gret-1* into the *VvMybA-1* promoter responsible for a variation in the color of grape berry (Kobayashi, Goto-Yamamoto, and Hirochika 2004). They can also modify gene expression following their insertion directly into or near gene regulatory regions (Feschotte 2008, Figure 2.2, page 36). For example, the LTR retrotransposon *hopscotch*, which is inserted into a regulatory region of the maize domestication gene *teosinte branched1* (*tb1*), acts as an enhancer to induce overexpression of this gene and consequently a decrease in the number of branches (Studer et al. 2011). In blood oranges, the insertion of the LTR retrotransposon *Rider*, upstream of the *Ruby* gene provides it with an alternative promoter, involved in cold-dependent red fruit colouration (Butelli et al. 2012). These variations will serve as raw material for evolutionary forces, such as selection.

2.2 Evolutionary processes

A population can be seen as a gene pool whose composition can vary over time as a result of evolutionary processes. Mathematical models have been proposed to investigate the relationship between allele frequencies in populations of organisms and evolutionary change such as that of G.H. Hardy and W. Weinberg. The Hardy-Weinberg Theorem models the distribution of genotype frequencies in populations that are not subject to any evolutionary forces. This theorem states that allele frequencies in a population remain constant from one generation to next over time only if the following assumptions are satisfied: no mutation, no natural selection, random mating and infinite population size (Andrews 2010).

Mutations can be important drivers of genetic diversity, on which other forces can act. In addition to mutations, other evolutionary mechanisms will impact on allele frequencies in a population, leading to violations of the Hardy-Weinberg equilibrium:

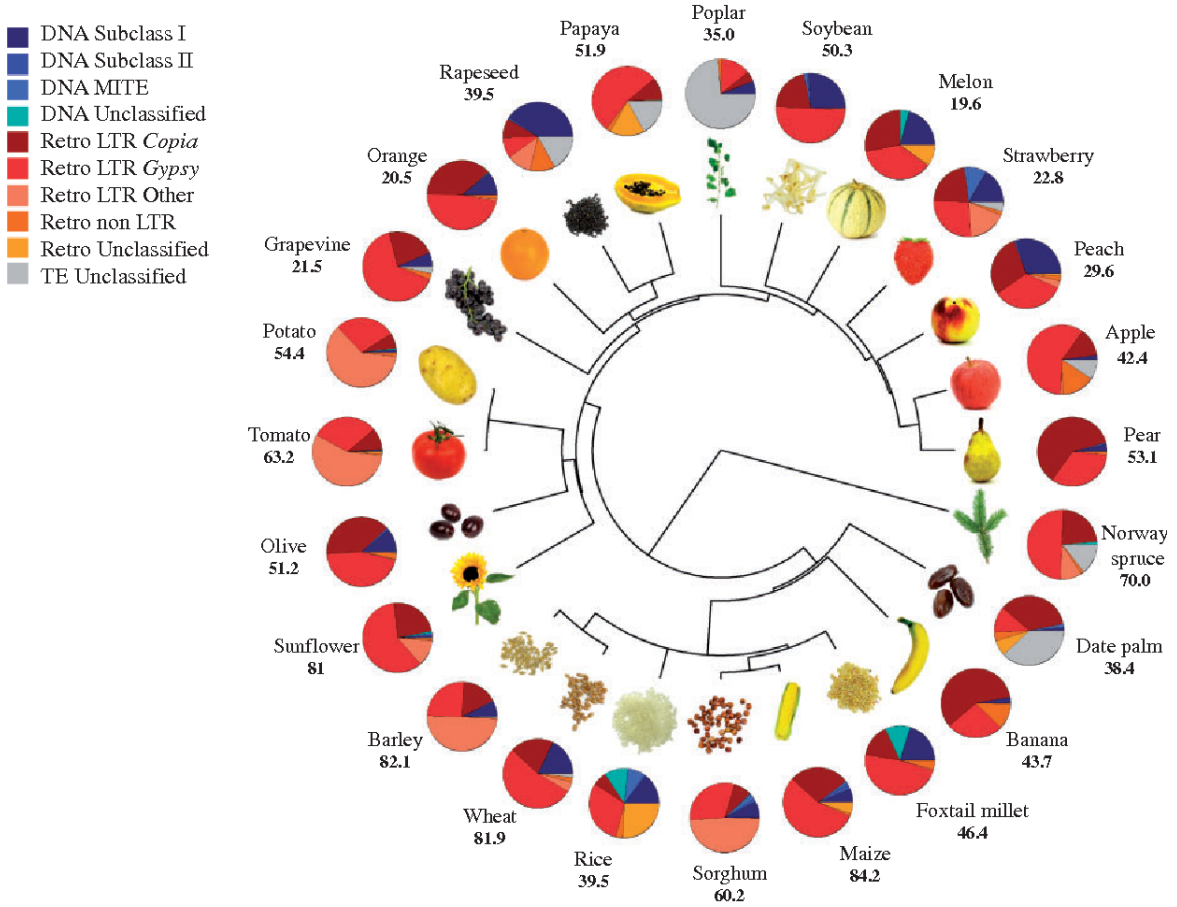

Figure 2.1: Relative proportion of different TE types within 24 sequenced crop genomes.
Phylogenetic relationships among species are shown (divergence time derived from http://www.timetree.org/ and [13]). Numbers in bold indicate for each crop the percentage of genomic sequence that has been annotated as TE-derived DNA. Pie charts illustrate the proportions of variousTE classes (DNA or RNA/retro elements), subclasses, orders. Relative contribution of TE types estimated as percentage of theTE-derived genomic fraction in different plant genomes. Figure from Vitte et al. 2014.

- **Non-random mating of individuals** according to their genotype or reproduction mode (autogamy, allogamy, asexual reproduction and so on);
- **Gene flow or migration** which corresponds to genes movement into or out a population as a result of the movement of individuals (or their gametes, such as pollen dispersal by a plant);
- **Genetic drift** describing random events (e.g.: random gametes mating) resulting in variation in allele or genotype frequencies within a population, independent of other evolutionary processes. It can lead to the loss of some alleles and the fixation of others, especially when the population size is reduced due to natural disaster (bottleneck effect) or the separation of a small group from the main population (founder effect) for example. In other words, the smaller the population, the greater the effect of drift;
- **Natural selection** as the combination of gene alleles that make an organism more or less fit and thus able to survive and reproduce in a given environment.

Three type of natural selection are defined : (i) **purifying (or negative) selection** which eliminates deleterious variations, (ii) **positive selection** which favors alleles that spread throughout a population, and (iii) **balancing selection** which maintains two or more alleles at a given locus (Figure 2.3, page 37).

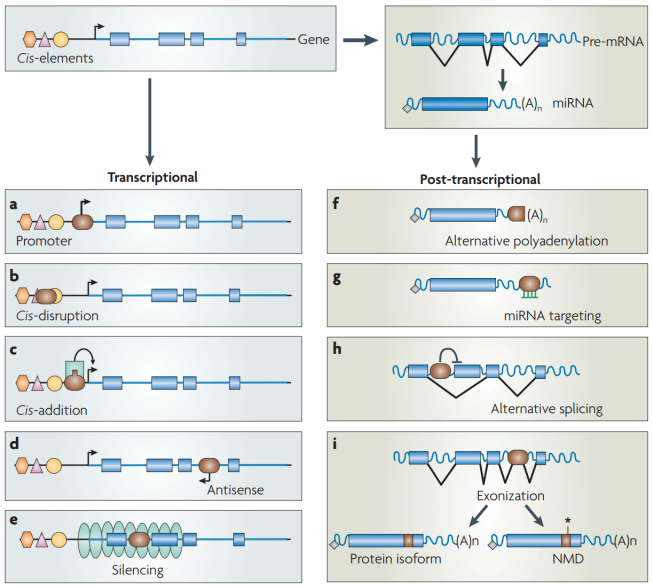

Figure 2.2: Impacts of transposable elements on gene expression.
Transposable elements can influence gene expression in many ways. At the transcriptional level, a TE (in brown) that has been inserted upstream of a gene can insert promoter sequences and introduce an alternative transcription start site (a), disrupt existing cis-regulatory element or elements (b), or introduce a new cis element such as a transcription factor binding site (c). In addition, a TE that is inserted within an intron can drive antisense transcription and potentially interfere with sense transcription (d). Finally, a TE can serve as a nucleation centre for the formation of heterochromatin (green ovals), potentially silencing the transcription of an adjacent gene or genes (e). At the post-transcriptional level, a TE that has been inserted in the 3' UTR of a gene can introduce an alternative polyadenylation site (f), a binding site for a microRNA (g) or for an RNA-binding protein (not shown). A TE that has been inserted within an intron can interfere with the normal splicing pattern of a pre-mRNA (h), provoking various forms of alternative splicing (for example, intron retention and exon skipping). A TE that is inserted within an intron and contains cryptic splice sites can be incorporated (exonized) as an alternative exon (i). This can result in the translation of a new protein isoform, or in the destabilization or degradation of the mRNA by the nonsense-mediated decay (NMD) pathway, especially if the exonized TE introduces a premature stop codon (represented by an asterisk). Figure from Feschotte 2008.

2.3 Association of structural variations with plant phenotypes

We have seen that there is a wide range of variations generated by different mechanisms, on which evolutionary mechanisms act, and some studies have been cited as example of the association of structural variations (due to TEs), with plant phenotypic traits. Overall, the number of studies has been growing steadily over the last 10 years, mainly in human at the beginning because of their association with numerous diseases such as various autoimmune disorders (Mamtani et al. 2010), HIV infection (Gonzalez et al. 2005), or Parkinson's or Alzeihmer's disease (Rovelet-Lecrux et al. no date; Singleton et al. 2003). These studies have been progressively extended to plants and have improved our understanding of structural variations and its impacts on genetic diversity (Saxena, Edwards, and Varshney 2014; Springer et al. 2009).

Thanks to advances in sequencing technologies, an increasing number of studies have reported examples of the impact of structural variations on phenotypic traits in plants associated with biotic and abiotic stress, flowering time, breeding traits as grain size or plant height (Table 2.1, page 38). In one of the first studies, Winzer et al. identified a 221-kb cluster of 10 genes, in opium (*Papaver somniferum*), associated with the synthesis of the anticancer alkaloid noscapine, which was absent in non-noscapine producing lines (Winzer et al. 2012). Many examples of CNVs have been described, such as the one of the *Ppd-B1* and *Vrn-A1* genes contributing to differences of flowering time in wheat, including photoperiod-

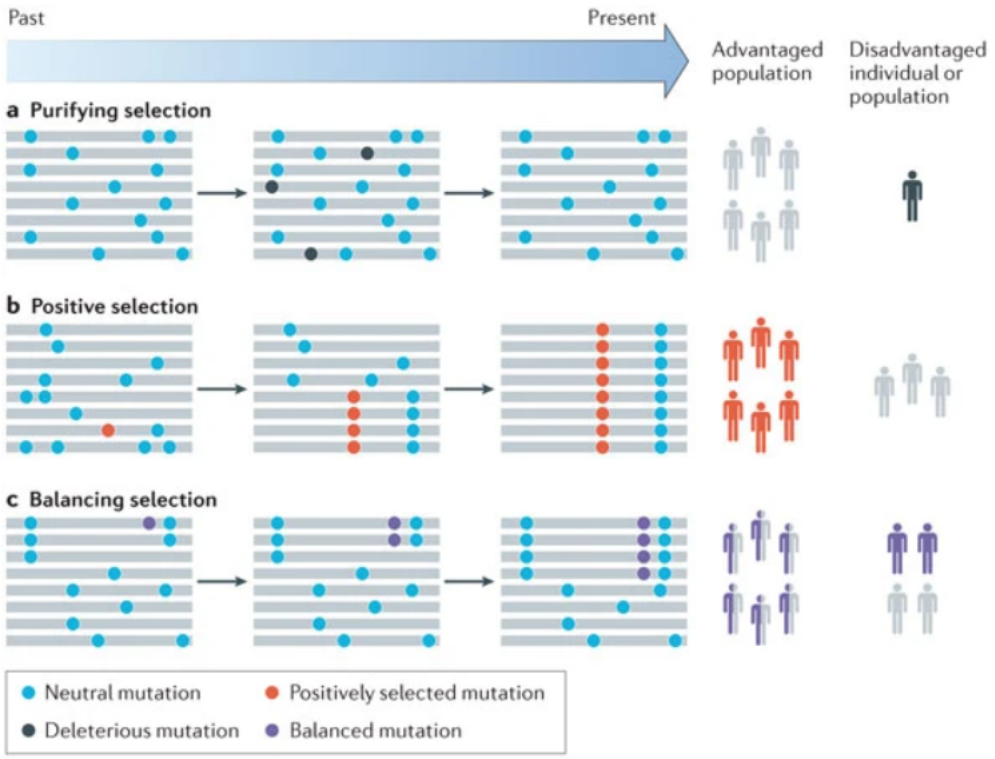

Figure 2.3: Types of selection.
The evolutionary fate of different types of mutations is represented in a sample of eight chromosomes. Blue circles indicate neutral polymorphisms. a | Purifying selection removes deleterious alleles (indicated by black circles) from the population. The pace at which deleterious mutations are purged from the population depends on their effect on host survival, which can range from lethal (immediately removed from the population) to mildly deleterious (tolerated but kept at low population frequencies). b | Positive selection increases the frequency of an advantageous mutation (indicated by a red circle) in the population. Advantageous mutations can be fixed (completed selective sweep) or polymorphic (ongoing selective sweep; not shown) in the population. c | Balancing selection maintains polymorphism in the population as a result of heterozygote advantage and frequency-dependent advantage (not shown). In the illustrated example, a mutation (indicated by a purple circle) confers a selective advantage at the heterozygous state, so individuals who are heterozygous at this particular position have a greater fitness than homozygous individuals. Figure from Quintana-Murci and Clark 2013.

sensitivity or vernalization requirement (Díaz et al. 2012). Another study based on the same cereal also highlighted the impact of a CNV on plant yield, targeting the *Taxkx4* gene associated with the leaf chlorophyll content after anthesis as well as grain weight (Chang et al. 2015). Finally, in maize, Maron et al. found a recent tandem triplication of the MATE1 gene that was only observed in three lines that were both aluminium-tolerant and originated from regions with highly acidic soils. Totally absent in teosinte, the authors suggested that this 30 kbp CNV appeared recently, after domestication, and was potentially being selected, conferring local adaptation to a specific environment (Maron et al. 2013). Pangenomics will provide a new way to explore genome variability on a large scale.

Species	SVs Type	Traits associated	Reference
Barley, *Hordeum vulgare*	CNV	Boron toxicity tolerance	Sutton et al. 2007
	CNV	Disease resistance	Muñoz-Amatriaín et al. 2013
Maize, *Zea mays*	PAV, CNV	Domestication	Springer et al. 2009
	CNV	Disease response, heterosis	Beló et al. 2010
	CNV	–	Swanson-Wagner et al. 2010
	CNV	Breeding selection	Jiao, Zhao, et al. 2012
	CNV	Aluminium tolerance	Maron et al. 2013
Rice, *Oryza sativa*	PAV, CNV	Grain size, disease resistance	Xu, Liu, et al. 2012
	CNV	Disease resistance	Yang, Li, et al. 2013; Yu, Wang, et al. 2013
	InDel	Root system architecture	Uga et al. 2013
Soybean, *Glycine max*	PAV, CNV	Stress responses	Haun et al. 2010; McHale et al. 2012
	CNV	Disease resistance	Lee, Kumar, et al. 2015
Sorghum, *Sorghum bicolor*	PAV, CNV	Disease resistance	Mace et al. 2014; Zheng et al. 2011
Wheat, *Triticum aestivum*	CNV	Vernalization, flowering time	Díaz et al. 2012; Würschum, Boeven, et al. 2015
	CNV	Plant height	Li, Xiao, et al. 2012
	PAV	Heading date	Nishida et al. 2013
	CNV	Frost tolerance	Sieber et al. 2016
	CNV	Winter hardiness	Würschum, Longin, et al. 2017
Oilseed rape, *Brassica napus*	PAV, CNV	Flowering time	Schiessl et al. 2017
	HE	Seed fibre	Stein, Coriton, et al. 2017
	PAV	Stay-green	Qian et al. 2016
	PAV	Disease resistance	Gabur et al. 2018

Table 2.1: List of structural variations with effects on agronomic traits in different crop species. Table from Gabur et al. 2018

3- From the genome to the pangenome

The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom – Isaac Asimov

3.1 Revolutionary improvements in sequencing technologies

... or how advances in sequencing have paved the way for pangenomics? As discussed in the previous section, genetic diversity consists of a wide range of variations from single-nucleotide polymorphisms (SNPs) to larger structural variations including PAVs, CNV and large scale chromosomal rearrangements. All that diversity constitutes the raw material for all evolutionary processes such as genetic drift or selection that will act on the frequency of each variation within a population, the evolution thus shaping the genome structure over time.

20 years after the publication of the first plant reference sequence, *Arabidopsis thaliana* (Initiative 2000), more than 1000 plant genomes from 788 species have been published thanks to advances in low-cost high-throughput sequencing technologies, and well-established assembly algorithms (Sun, Shang, et al. 2022, Marks et al. 2021). This made it possible to sequence large and complex plant genomes (Figure 3.1, page 40) such as the largest plant genome sequenced, *Pinus lambertiana* of 31 Gb (Stevens et al. 2016). Furthermore, in recent years, long-read genome sequencing has enabled to produce more and more high-quality assembled genomes (Figure 3.2, page 41). Since the publication of the first plant genomes, many studies on genetic diversity were performed, first on SNPs (Atwell et al. 2010; Lai et al. 2012; Xu, Liu, et al. 2012; Zhou, Jiang, et al. 2015) and then increasingly on structural variations (Muñoz-Amatriaín et al. 2013; Springer et al. 2009; Swanson-Wagner et al. 2010), as high-throughput sequencing technologies became more affordable in terms of cost and access, and as adequate computational methods were implemented. Thus, in the last 10 years, more and more studies based on genome (re)sequencing have described a wide range of impacts on gene content variation such as deletions in the *sh1* or the *GmCHX1* genes involved in loss of seed dispersal in rice or soybean salt tolerance respectively (Lin, Li, et al. 2012; Qi, Li, et al. 2014), or duplications of the *GL7* or *SUN* genes contributing to variations in rice grain size or tomato fruit shape respectively (Wang, Xiong, et al. 2015; Xiao et al. 2008). One extreme effect of a structural variation on a gene is the presence or absence of that gene. It has become increasingly clear that a single reference genome is insufficient to capture all the genetic diversity present within a species. Gradually, we have moved from the concept of the reference genome to that of the pangenome.

Pangenome concept was used, in 2005, by Tettelin et al., to refer all genes present in a bacterial species, including the core genome that contains genes present in all strains and the dispensable genome composed of genes present in a subset of strains (Tettelin et al. 2005). Morgante, De Paoli, and Radovic 2007 were the first to use the pangenome concept in plants, to describe all genomic segments, relying on the large amount of variations observed in genic and intergenic regions in plant genome, largely due to transposable elements (Figure 3.4, page 42). We will come later on these two definitions that have been proposed, focused on genes or on the whole genome, in the part II (Section 5.2.1, page 52) and we will discuss the possible impacts on the pangenomics analysis according to the definition used.

Over the past decade, pangenomics has facilitated access to the vast unexplored content of structural variations. For instance, Gordon et al. showed that up to 8Mb of sequences were present in any *Brachypodium distachyon* individual and absent from the reference genome (Gordon et al. 2017). In another study of the tomato pangenome, Gao et al. revealed that 4,873 genes absent from the reference

genome were present in other sequenced cultivated and wild accessions (Gao et al. 2019). Before discussing some of the avenues that can be explored with plant pangenomics, we will describe the main steps to build a pangenome.

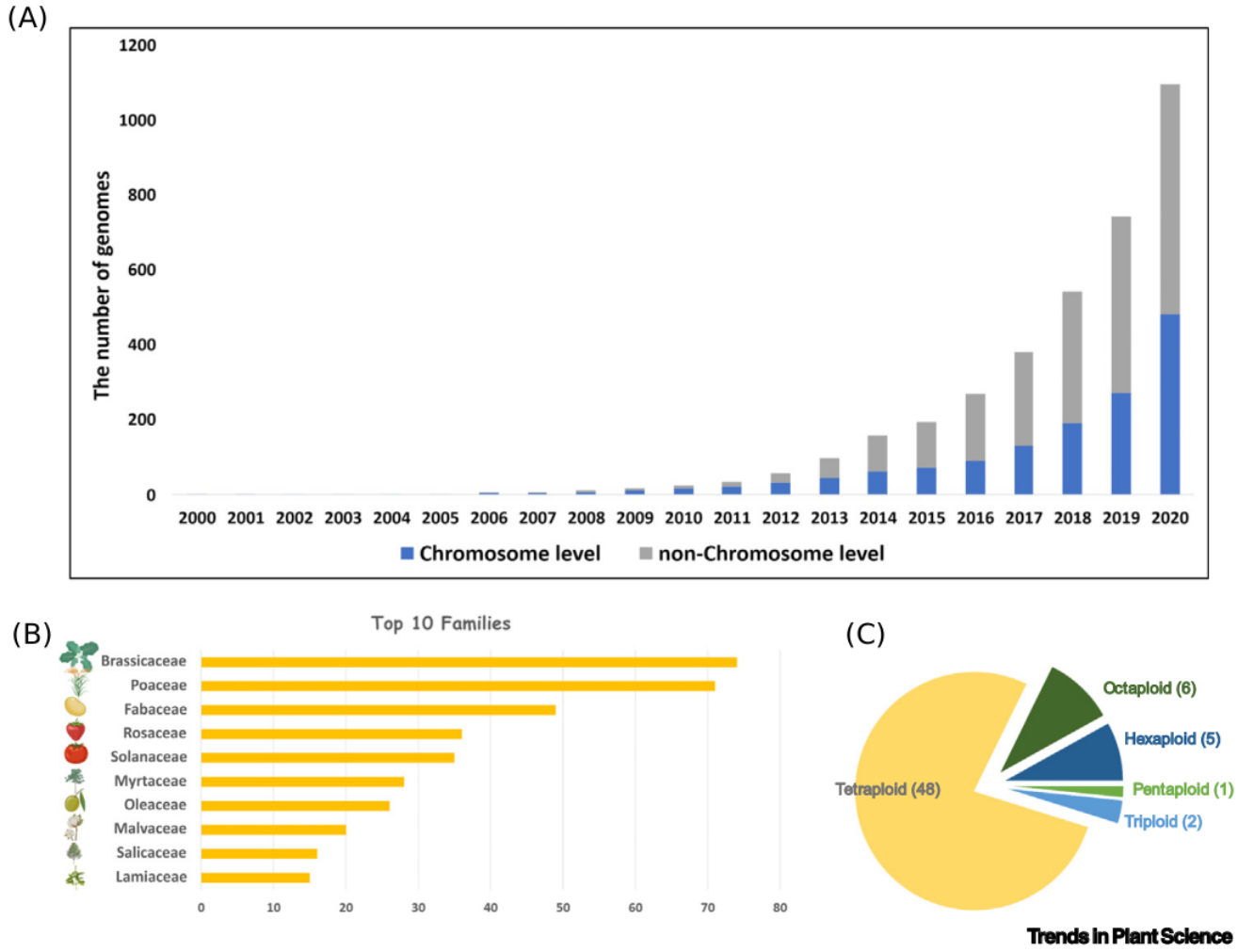

Figure 3.1: Published plant genomes.
(A) The number of plant genomes sequenced at the chromosome and non-chromosome levels since publication of the Arabidopsis thaliana genome in 2000. (B) The top 10 families with the most sequenced genomes in angiosperms. (C) Number of polyploid genomes sequenced in angiosperms. Figure from Sun, Shang, et al. 2022.

3.2 Pangenome construction

Before starting the construction of the pangenome, there is an important sampling and sequencing phase that will depend on the scientific question to be answered. At this stage of designing the experimental and methodological plan, it is also important to keep in mind all the factors that can impact the completeness of the pangenome construction. This last point will be discussed in chapter II (section 5.2.3, page 53). Overall, a pangenomics analysis consists of different steps from pangenome construction to pangenome visualization (the grail) and pangenome characterization (Figure 3.3, page 42). Let's start with the first (technical) step, but not the most trivial one.

We will only describe the methods used for plants, which can be divided into three main approaches:

- The ***assemble-then-map approach*** that can be used with both short and long read sequencing data;

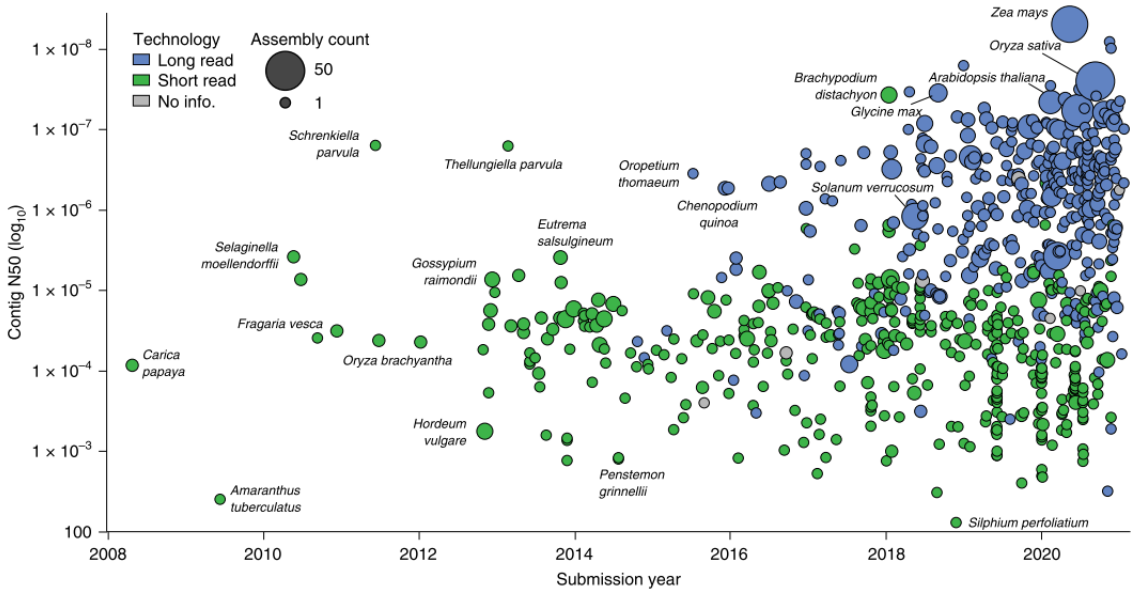

Figure 3.2: Changes in land plant genome assembly quality and availability over time.
Assembly contiguity by submission date for 798 land plant species with publicly available genome assemblies. Points are coloured by the type of sequencing technology used and scaled by the number of assemblies available for that species. There is an improvement in contiguity associated with the advent of long-read sequencing technology, and a noticeable increase in the number of genome assemblies generated annually. Figure from Marks et al. 2021.

- The ***map-then-assemble approach*** based primarily on short-read sequencing data;
- The latest one, ***the graph-based approach***, that is still in development, using both long and short reads.

The ***assembly-then-map approach*** starts with *de novo* assembly of multiple genomes followed by genomes comparison (*i.e* mapping of contigs against reference genome or pair-wise genome comparison or only genes comparison) (Gordon et al. 2017; Hu et al. 2017). This approach requires a high sequencing depth to produce a good quality sequenced genome, that can be relatively expensive or complicated with large genomes or large sampling, even more with short reads only.

The ***map-then-assemble approach*** consists in short reads mapping against a reference genome followed by *de novo* assembly of the unmapped reads. This approach were used, for example, in human (Sherman et al. 2019), in sunflower (Hübner et al. 2019) or in a slightly different form in *Brassica oleracea* (Golicz, Bayer, et al. 2016) or banana (Rijzaani et al. 2021) (Table 3.1, page 44). This approach is less consuming-time and less costly, as it requires lower sequencing depths. The fact that contigs are not positioned on a genome can make it difficult to differentiate similar sequences and to know whether they are alleles or gene copies.

The last method uses **graphs** to represent all the variations contained in different accessions of the same species. Building a graph and using it for diversity studies for example is still complex with several tools under development, formats specific to each tool but the fact of being able to integrate all the variations detected in each pangenomic analysis, within an unique structure enriched at each analysis, is very promising and attractive.

3.3 What ways can be explored with plant pangenomics?

As we have seen, a single reference genome does not capture the whole diversity within a species and pangenomics is proving to be a particularly interesting new approach for exploring the broad content of

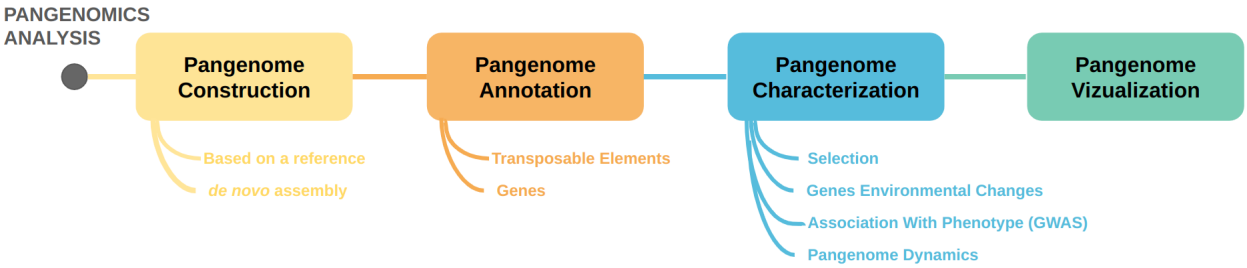

Figure 3.3: Overview of a pangenomic study

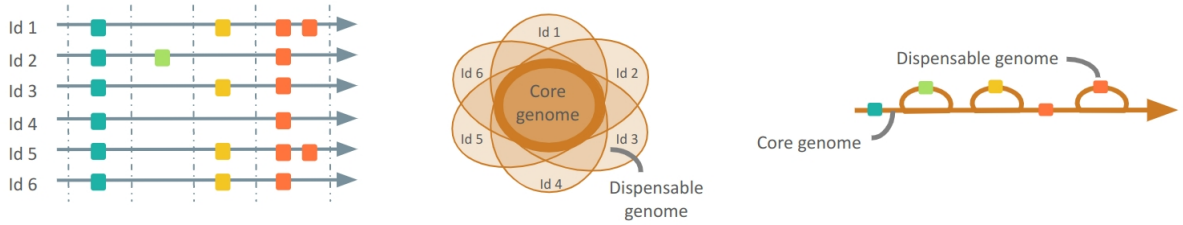

Figure 3.4: Three ways of representing pangenome.
(Left) Pangenome is displayed as an inventory of genomic items shared or not (including genes) within a group of related individuals. (Center) pangenome is represented as a Venn diagram in which the core genome is the common set of sequences shared by all individuals of the group; the remaining belongs to the dispensable genome. (Right) Pangenome is represented as an oriented graph, in which alternative paths replace both the structural and unique variants.

structural variations at the population, species or broader scale. Thus, several studies have highlighted that a large number of sequences, including a significant number of genes, are part of the variable genome. For example, in a study of 18 wheat cultivars, 12,150 new genes were identified as absent from the reference (Montenegro et al. 2017). Similarly, in rice, 10,872 genes were detected as missing from the reference in a study based on 66 accessions (Zhao et al. 2018). The table 3.1 (page 44) also provides an overview of the variable genome rate within about 25 species, ranging from 10% (Torkamaneh, Lemay, and Belzile 2021) to 80% (Yang, Liu, et al. 2022), as observed in pangenomic studies carried out over the past 10 years. In addition, a growing number of studies have revealed that SVs in the variable genome can be associated with variations in phenotypic traits such as flowering time (Gordon et al. 2017; Song et al. 2020) or, more specifically, variations between between cultivated and wild accessions such as seed coat color in soybean (Song et al. 2020) or fruit color in the strawberry (Qiao et al. 2021). In a study performed on a population of 493 cultivated and wild sunflowers, Hübner et al. 2019 also showed that about 10% of the cultivated sunflower genome came from introgressions of genomic regions from wild sunflower species. These latter examples illustrate that pangenomic studies, conducted jointly within cultivated species and their wild relatives, can help identify genes lost during the processes of domestication and, more broadly, of selection and adaptation throughout the history of a species.

My PhD project used the African rice as a study model, to investigate diversity within cultivated and closely related wild plants and the impact of evolutionary forces on this diversity on a relatively short time scale, that of domestication and human selection... and all this from a pangenomic perspective.

Species	#ind.	#genes	Core	Disp	Approach	wild sp.	References
Arabidopsis thaliana	19	37 789	70%	30%	assemble-then-map	no	Contreras-Moreira et al. 2017
Brachypodium distachyon	54	37 886	54%	46%	assemble-then-map	no	Gordon et al. 2017
Brassica oleacera	10	61,379	81%	19%	map-then-assemble*	yes	Golicz, Bayer, et al. 2016
Brassica napus	53	94,013	62%	38%	map-then-assemble*	no	Hurgobin et al. 2018
-	9	105,672	56%	44%	assemble-then-map	no	Song et al. 2020
Capsicum	355	51,757	56%	44%	map-then-assemble	no	Ou et al. 2018
Cucumis sativus L.	11	26,822	69%	31%	assemble-then-map	yes	Li, Wang, et al. 2022
Fragaria spp.	6	25,687	44%	56%	assemble-then-map	no	Qiao et al. 2021
Helianthus annuus L.	493	61,205	73%	27%	map-then-assemble	yes	Hübner et al. 2019
lupinus albus L.	39	46,890	68%	32%	assemble-then-map	yes	Hufnagel et al. 2021
Medicago	15	74,700	42%	58%	assemble-then-map	no	Zhou, Silverstein, et al. 2017
Populus	22	-	81%	19%	mapping only	no	Pinosio et al. 2016
Oryza sativa	3	40,362	92%	8%	assemble-then-map	no	Schatz et al. 2014
	66	42,580	62%	38%	assemble-then-map	yes	Zhao et al. 2018
	453	23,876	72%	38%	assemble-then-map	no	Wang, Mauleon, et al. 2018
	33	66,636	31%	69%	assemble-then-map**	no	Qin et al. 2021
	251	51,359	43%	57%	assemble-then-map**	yes	Shang et al. 2022
	111	75,305	51%	48%	assemble-then-map**	yes	Zhang, Xue, et al. 2022
	56	38,998	80%	20%	map-then-assemble	no	Woldegiorgis et al. 2022
Oryza glaberrima	163	39,106	86%	14%	map-the-assemble	yes	Monat, Tranchant-Dubreuil, et al. 2018
Oryza barthii	86	40,475	98%	2%	map-the-assemble	yes	Monat, Tranchant-Dubreuil, et al. 2018
Pisum	118	112,776	51%	49%	map-then-assemble	yes	Yang, Liu, et al. 2022
			19%	81%			
Raphanus	11	41,952	36%	%	assemble-then-map	no	Zhang, Liu, et al. 2021
Sesamum indicum	5	26,472	58%	42%	assemble-then-map	no	Yu, Golicz, et al. 2019
Solanum lycopersicum	586[a]	40,369	74%	26%	assemble-then-map	yes	Gao et al. 2019
Solanum melongena L.	26	35,148	92%	8%	assemble-then-map	yes	Barchi et al. 2021
Sorghum	13	44,079	36%	64%	assemble-then-map	yes	Tao et al. 2021
Soybean	7	59,080	80%	20%	assemble-then-map	yes	Li, Zhou, Ma, et al. 2014
	26	57,492	50%	50%	assemble-then-map[g]	yes	Liu, Du, et al. 2020

	204	54,531	93%	7%	assemble-then-map	no	Torkamaneh, Lemay, and Belzile 2021
Bread Wheat	18	128,656	64%	34%	map-then-assemble	no	Montenegro et al. 2017
Zea mays	26	103,033	31%	69%	assemble-then-map	no	Hufford, Seetharam, et al. 2021

Table 3.1: Overview of plant genome studies to date. For each study, this table displays general information such as the species, the number of samples as the method used to build the pangenome and whether wild species were used. This table also gives the total number of pangenes identified as the core and dispensable ratio.

4- PhD context and objectives

If we knew what it was we were doing, it would not be called research, would it ? – Albert Einstein

4.1 The African rice as a study model

Rice is the most widely consumed cereals in the word, belonging to the angiosperms (flowering plants) and more precisely to the large group of monocotyledons such as orchids (family *Phalaenopsis*), palms (family *Arecaceae*), banana (genus *musa*), rushes (family *Juncaceae*) and other grasses (family Poaceae). It is part of the relatively small genus *Oryza*, which comprises only 23 species with genomes of different ploidy levels (Table in appendix 14.4, page 149) and a wide range of habitats spread over the tropical and subtropical regions of the world such as forests, savanna, mountainsides, rivers or lakes (Figure 4.1 page 45, Stein, Yu, et al. 2018; Vaughan, Morishima, and Kadowaki 2003). Of these 23 species, only two are cultivated today, the African rice (*Oryza glaberrima*) and the Asian rice (*Oryza sativa*). These two species were domesticated independently, both in different regions, at different periods and from separate wild relatives (Vaughan, Morishima, and Kadowaki 2003). The cultivated species *O. glaberrima* was domesticated from the wild rice *O. barthii* in the inner delta of the Niger River in Mali 3,500 years ago (Cubry et al. 2018; Wang, Xiong, et al. 2015). Although Asia is the world's largest producer and exporter of rice with highly appreciated taste qualities, *O. glaberrima* has nevertheless developed interesting agronomic traits compared to its Asian cousin, to overcome biotic and abiotic stresses such as drought, salinity and flooding. The Asian and African cultivated rice are diploid (n=12), autogamous and self-pollinating plants.

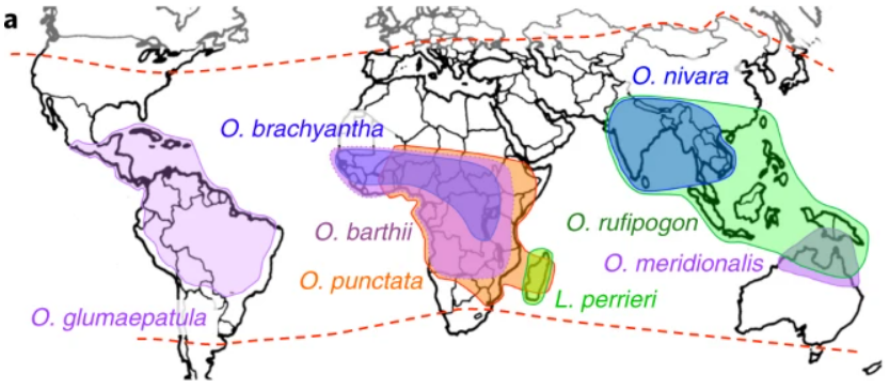

Figure 4.1: Geographic distribution of wild species of *Oryza* and *Leersia*.
Geographic ranges of wild *Oryza* species and the outgroup species *L. perrieri* sequenced in this study. Dashed red lines show the limits of rice cultivation. Mapped ranges are adapted from IRRI's Knowledge Bank. Figure from Stein, Yu, et al. 2018.

.

Oryza glaberrima was cultivated long before Europeans arrived on the continent. The cultivated African rice *Oryza glaberrima* was named and described as a new species of rice by Steudel in 1855, from samples collected by Edelestan Jardin in Portuguese Guinea between 1845 and 1848 (Portères 1955). African rice cultivation was gradually replaced by Asian rice introduced by the Portuguese as early as

the middle of the 16th century (Portères 1962). However, *Oryza glaberrima* continues to be cultivated, particularly because of its use in sacred rites (Linares 2002).

There are three main types of rice cultivation in West Africa (Figure 4.2, page 46 and Figure 4.3, page 46):

- **rainfed lowland rice**, including mangrove rice, which is predominant, for instance, in Guinea, in the coastal areas of its gulf
- **rainfed upland rice** in upland areas such as Guinea, southern Mali, western Benin and Nigeria.
- **irrigated rice cultivation** mainly in three large areas located in Senegal, Mali and Nigeria.

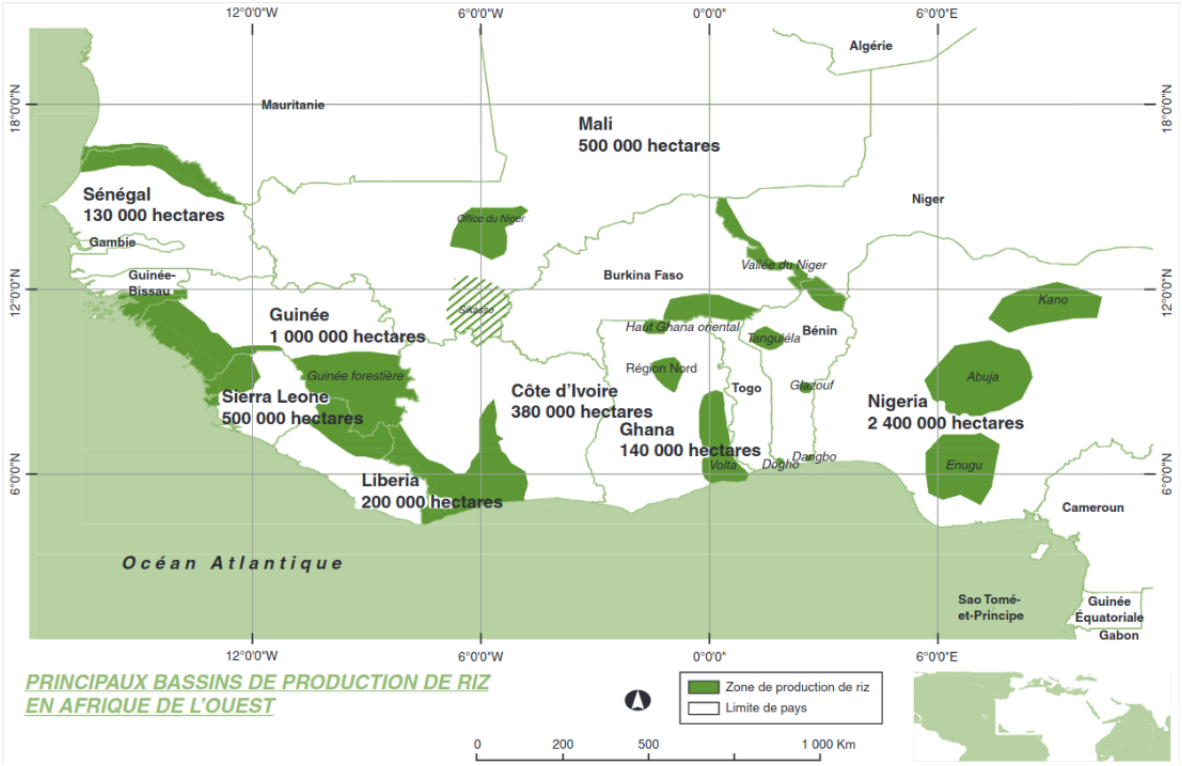

Figure 4.2: Main areas of rice production in West Africa.
Figure from Villar and Bauer 2013.

Burkina Faso, 2003
© IRD -Mathilde Savy

Guinea, 2012
© IRD / Univ. Montpellier

Mali, 1985
© IRD - Henri Guillaume

Senegal, 1950
© IRD - Roger Fauck

Figure 4.3: Different rice-growing landscapes in West Africa.
Source : IRD mutimedia (https://multimedia.ird.fr/).

Rice is the second most consumed cereal in Africa, after maize. This consumption is constantly increasing, particularly in West Africa, due to strong demographic growth, urbanisation and changes in lifestyle (Figure 4.4, page 47). As African rice production does not cover needs, there has been a sharp increase in rice exports, which represented 20% of the rice consumed in the early 1960s compared to 40%

in 2013 (Villar and Bauer 2013). Despite efforts to develop local rice production since the sharp rise in rice prices in 2008, this dependence has continued to the present day, with 20% of global rice exports going to West African markets.

It is in this context of food security and climate change that my PhD project was developed to characterize the diversity of African rice and better understand how this diversity can be of interest for crop improvement.

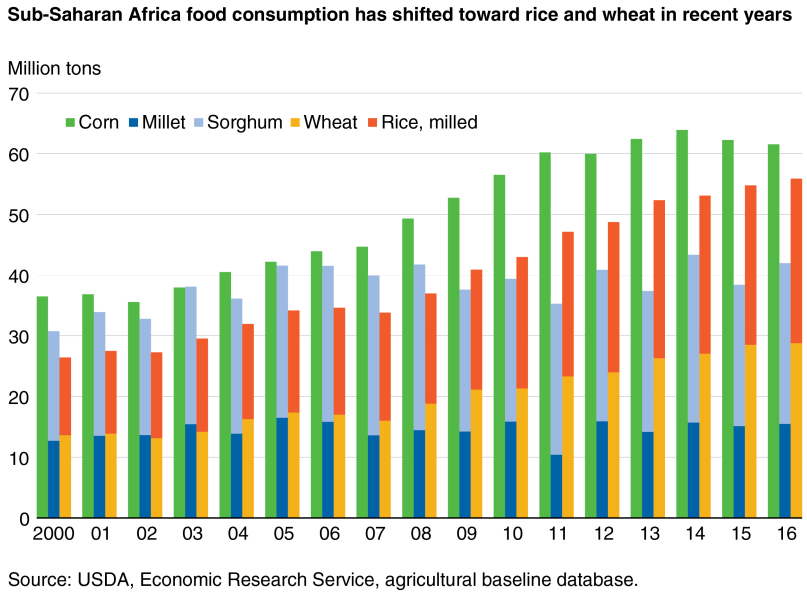

Figure 4.4: Sub-Saharan Africa food consumption.
In this region, the diet has shifted toward rice and wheat in recent years. Figure fromUSDA, Economic Research Service, agricultural baseline database (https://www.ers.usda.gov/amber-waves/2017/october/sub-saharan-africa-is-projected-to-be-the-leader-in-global-rice-imports).

4.2 Scientific Objectives

The main objectives of my PhD project were to fully explore the genetic diversity within cultivated and wild african rice, including structural variations, and to improve knowledge on the evolutionary mechanisms such as domestication and human selection on this diversity. We will focus on questions such as which roles do these structural variations play in gene composition and adaptation or how evolutionary forces have shaped (pan)genome organization and dynamics.

Firstly, a state of the art on pangenomics was carried out in order to define how this approach could be used to study genome variations in a population. Secondly, we set up a tool to build a plant pangenome from scratch using whole genome sequencing by short reads. Taking advantage of 247 african rice genomes resequencing, we applied our approach as a proof-of-concept to build the first african rice pangenome. Finally, after characterizing these pangenomes at inter- and intra-species levels, we investigated how domestication shaped the African rice pangenome.

II Pangenome: From the concept to large-scale studies across species

5- Context and main points

If you had to specialise in order to learn, you have to be open to understand – François Kourilsky

5.1 Why to write a review about Plant Pangenomes ?

By 2019, beginning of this work, the pangenome concept had become more widely used to study the gene content variation within bacterial species (Medini et al. 2005; Tettelin et al. 2005, Vernikos et al. 2015). With the spread of low-cost, high-throughput sequencing technologies, this concept was progressively extended to the higher organisms, although pangenomics analysis were (and is still today) challenging, due to their large genome size and complexity (repeat content or polyploidy). At that time, the number of pangenomic analyses on crops was growing (Contreras-Moreira et al. 2017; Golicz, Bayer, et al. 2016; Gordon et al. 2017; Li, Zhou, Ma, et al. 2014; Schatz et al. 2014; Wang, Mauleon, et al. 2018; Yao et al. 2015; Zhao et al. 2018, Figure 5.1) and even more, so we took into account large-scale analyses studying structural variations or gene content variation between genomes without mentioning the pangenome concept (Berger et al. 2012; Springer et al. 2009; Swanson-Wagner et al. 2010). However, there were very few article reviews that synthesized knowledge on this emerging research topic in plants and the paper by Golicz, Batley, and Edwards, published in 2016, was one of the few examples (Golicz, Batley, and Edwards 2016).

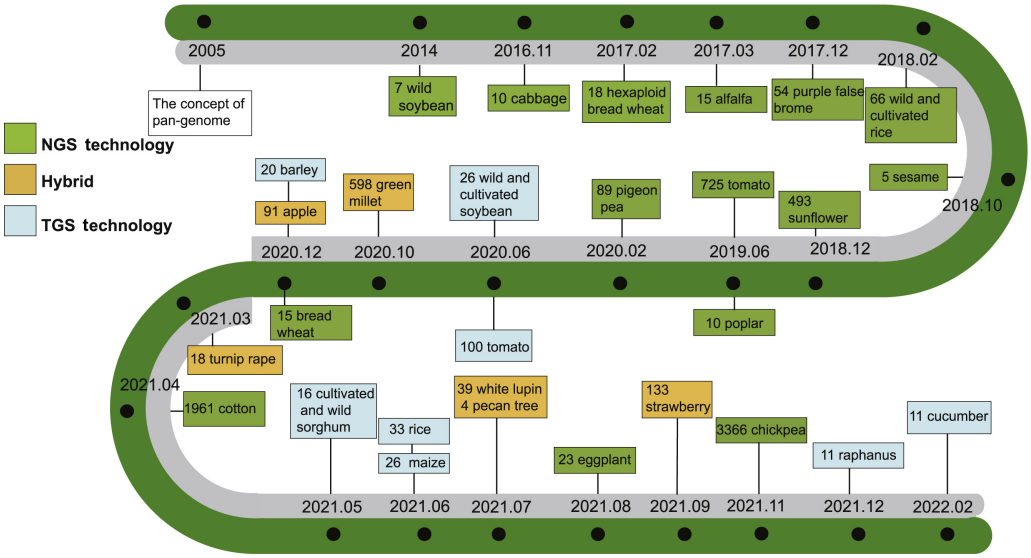

Figure 5.1: Timeline and basic information for the released plant pangenomes.
The different sequencing technologies used to construct the pangenomes are indicated using different colors. Solid black circles indicate past events in plant pan-genomics. The technologies are indicated using colored rectangular boxes: light green, next-generation sequencing; dark orange, hybrid sequencing; light blue, long-read sequencing. The sample size and species are indicated in the colored rectangular boxes. Figure from Li, Liu, et al. 2022.

Therefore, the proposal to write a review of current literature appeared very appropriate to perform the current state-of-the-art on this emerging concept, to identify challenges, limitations and research ideas

on this topic. We also thought that this review could be a useful resource for other researchers who were asking the same questions as us, such as:

- Why a paradigm shift from the reference genome to the pangenome ?
- What is a pangenome ? Does it only involve genes ?
- What do you know about pangenome especially in plants ?
- How to build a pangenome ?
- Which factors and forces impact the ability of the pangenome to increase in size ?

5.2 Back to 2019... What did we know about plant pan genomes ?

With the dramatic advances in high-throughput sequencing methodologies and their decreasing cost, a growing number of studies highlighted the limitation of using a single genome to assess genetic diversity and identify complex DNA polymorphisms including structural variations such as large insertion/deletion or Presence Absence Variations and Copy Number Variations (Berger et al. 2012; Springer et al. 2009; Swanson-Wagner et al. 2010).

Previous studies had also shown the absence of specific genes in a reference genome such as the genes conferring tolerance to submergence (Submergence 1, Sub1) (Xu, Xu, et al. 2006) or deep water (SNORKEL1, SNORKEL2) (Hattori et al. 2009) in the Asian rice. However, pangenomic studies highlighted that this could involve a large number of genes. For example, among the earliest papers on crop pangenomics, two studies in wheat (Montenegro et al. 2017) and rice (Yao et al. 2015) showed that 12,150 genes and 8,000 genes, respectively, were missing from the reference genome. In 2017, Gordon et al. identified genes absent from the *Brachypodium distachyon* genome that were involved in environmental responses such as the flowering time variation (Gordon et al. 2017). Specifically, they found a gene encoding an NF-Y subunit Transcription factor present in all delayed or extremely delayed flowering lines but absent from rapid or intermediate flowering lines (including the reference genome).

Thus, over the past few decades, it became increasingly clear that a single genome reference was insufficient to capture all the sequences present in a species and that the concept of a reference genome had to be rethought. This was one of the starting points for the shift from the reference genome paradigm to the pangenome.

5.2.1 What is a pangenome ?

The pangenome concept was first proposed by Tettelin et al. (Medini et al. 2005; Tettelin et al. 2005) to refer to the entire set of genes present in all individuals of a species, which includes (i) a **core genome** containing genes present in all individuals and (ii) a **dispensable genome** shared only in a subset of individuals. The terms core and dispensable are commonly used today but given current knowledge, the term dispensable does not seem as appropriate as when it was proposed in 2005 by Tettelin et al and the term variable is also used (Golicz, Bayer, et al. 2016; Hufnagel et al. 2021; Montenegro et al. 2017; Rijzaani et al. 2021) or shell (Gao et al. 2019; Gordon et al. 2017).

In addition to this exclusively gene-based definition, we proposed a more comprehensive definition encompassing all DNA sequences, both genes and non-genic sequences that we called structure-based definition (Figure 3.4, page 42). Depending on the definition used, the classification of genes in the core genome or in the variable genome could be different. If we consider highly similar sequences such as recent paralogs, in a function-based definition, paralogs will be considered as a unique sequence. The presence or absence of genes will be thus calculated without taking into account its location in the genome, and the function will be then classified as core. Another example could be genes involved in a genomic recombination in a subset of a population, and, depending the structure- or function-based definition used, the gene could be considered as variable or core. In addition, the structure-based definition also takes into account transposable elements, whether inserted in a gene or outside of it. Indeed, more and more papers have shown that TEs can influence gene expression in many ways following their insertion in a gene or

upstream of a gene (such as the promoter sequence) (Figure 2.2, page 36, Feschotte 2008). However, dealing with multiple copies of multiple TEs families can rapidly become a complex and challenging task especially in a pangenome.

Finally, in 2019, before the concept of pangenome graph was widespread, we proposed the term **panreference** to describe the set of sequences in a pangenome, including a reference genome and all sequences assembled (and absent from this reference), with the additional information of the position of these sequences (all or part) on the reference.

5.2.2 To be Core or Dispensable ?

Core genes are likely to be involved primarily in essential and basic functions such as DNA replication, maintenance of cellular homeostasis or cellular process as glycolysis. This explains why they tend to be conserved. In contrast, dispensable genes contribute to the diversity of a species, allowing it to adapt to various environmental conditions such as biotic and abiotic stress (including defense and response genes), as well as developmental genes such as genes controlling flowering time. Thus, many studies have shown the trend of faster evolution of dispensable genes compared to core genes:

- (i) A higher SNP density in dispensable genes (Soybean (Li, Zhou, Ma, et al. 2014), *Brachypodium* (Gordon et al. 2017), Rice (Wang, Mauleon, et al. 2018));
- (ii) The ratio non-synonymous to synonymous mutations higher in the variable genes (Soybean (Li, Zhou, Ma, et al. 2014), *Brachypodium* (Gordon et al. 2017)).

We will discuss these contrasting features in more detail in section 5.2.5 (page 54).

When a pangenome analysis is initiated, one of the first questions raised is "how many genomes should be sequenced to capture the full diversity of a given group ?".

5.2.3 Points to keep in mind before pangenome construction

It's worth noting that several factors can impact the completeness of a pangenome construction, starting with the sample size. So, recurrent questions have to be arisen:

- How many genomes should be sequenced to maximise diversity within a group ?
- Will there still be the same set of sequences shared by all individuals even if a newly sequenced genome is added ?
- Will new dispensable genes still be discovered with the sequencing of additional individuals ?

In order to validate whether the final size of the pangenome has been reached, Tettelin et al. proposed to represent the total number of sequences found after each new individual sequenced (Tettelin et al. 2005, Figure 5.2, page 53). He also introduced the concept of an open or closed pangenome. A pangenome is closed when only a few new sequences are added when new genomes are included in the analysis. In contrast, the pangenome is defined as open if it is always unlimited regardless of the number of new genomes sequenced.

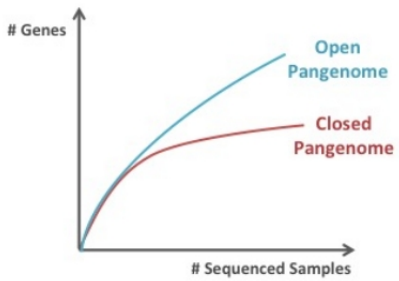

Figure 5.2: Open and closed pangenomes.
Figure From Tranchant-Dubreuil, Rouard, and Sabot 2019.

Thus, the size of the pangenome and its core compartment may be underestimated if too few samples are used or if the samples chosen represent only a small fraction of the true diversity. Two pangenomics analyses in Rice clearly illustrated the impact of sample number. In 2014, a study of three asian rice accessions reported a pangenome of 40,362 genes, of which 8% were dispensable (Schatz et al. 2014). Four years later, in a study based on over 3,000 rice accessions (Wang, Mauleon, et al. 2018), Wang, Mauleon, et al. identified a much larger pan-genome (48,098 genes) with a higher percentage of dispensable genes (41%).

In addition to sample size, sample selection can have a significant impact on pangenome size: using closely related samples would underestimate the pangenome size and could lead to the incorrect conclusion of a closed pangenome. Conversely, integrating wild accessions would generate a larger pangenome with a higher dispensable genome than using only cultivated crops (Shang et al. 2022).

Beyond the choice of samples, genetic properties of the selected samples, including genome size (e.g., TEs content), mode of reproduction or ploidy level (e.g., allogamy or autogamy), or living conditions, can also influence the results of a pangenome study.

To conclude, comparison of pangenomic analyses, even when performed on closely related species, can be complex because of the many factors that can impact the pangenome completeness, including technical limitations related to the methods used to build pangenome.

5.2.4 Methods to build a pangenome

In 2019, two approaches were mainly used to assemble pangenomes based primarily on short reads resequencing: the **"assemble-then-map"** and the **"map-then-assemble" methods**. The former consists of *de novo* genome assembly followed by genome comparison (Gordon et al. 2017; Hu et al. 2017; Li, Zhou, Ma, et al. 2014; Schatz et al. 2014; Zhao et al. 2018) while the latter is based on the mapping of reads followed by the *de novo* assembly of unmapped reads (Golicz, Bayer, et al. 2016; Montenegro et al. 2017; Yao et al. 2015). A full description of the approaches used to build a pan-genome from both short- and long-read sequencing technologies is provided in part I (section 3.2, page 40), along with the advantages and limitations of each approach.

5.2.5 Dynamics of pangenome compartments

Pangenome encompasses all the diversity present in a group of individuals. The core genome contains genes involved in essential pathways, often described as the minimal genome necessary for a cell to live. The variable genome encompasses the diversity within a group of individuals, including genes that allow adaptation to different environments. Pangenomic analyses have revealed dynamic evolution of genomes within a species or between related species, including a broad spectrum of structural variations. These result from a variety of mechanisms including TE activity, recombination, introgression, described in the part I (section 2, page 33). These mutations provide raw material for evolutionary forces such as genetic drift or selection, thus contributing to the diversity of the pangenome. The size of the pangenome will depend on the genome dynamics of the group under consideration; for example, in this regard, bacteria have a relatively larger pangenome than plants due to their higher level of gene flow. The ability of the pangenome to grow or remain stable, as well as to switch from core to dispensable and back again, is strongly connected to the balance between gain and loss events and the ability to adapt to various environments (Figure 5.3, page 55).

All the issues discussed in this section were reviewed in *Annual Plant Reviews* in 2019, which is presented hereafter in part 6 (page 57).

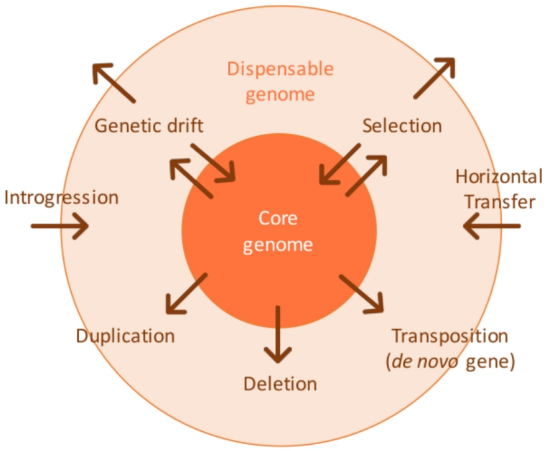

Figure 5.3: Dynamic overview of the pangenome structure shaped by different events and forces.
New sequences are added to the dispensable genome through mutations, duplications, deletions and transpositions, while the core genome content may decrease by deletion and transposition. Horizontal transfer and introgression also impact on the dispensable genome compartment (sequence gain). Moreover, positive and purifying selection as well as genetic drift impact on the core and dispensable genomes (sequence gain and loss) as well as on the pangenome (sequence loss). Figure adapted From Tranchant-Dubreuil, Rouard, and Sabot 2019.

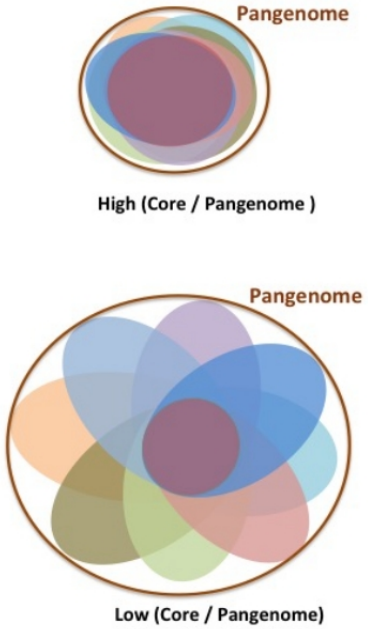

Figure 5.4: Core/Pangenome ratio illustration.
Figure From Tranchant-Dubreuil, Rouard, and Sabot 2019.

6- Review "Plant pangenome: impact on phenotypes and evolution"

Life is like riding a bicycle. To keep your balance, you must keep moving – Albert Einstein

Annual Plant Reviews (2019) **2**, 1–25
doi: 10.1002/9781119312994.apr0664

http://onlinelibrary.wiley.com

PLANT PANGENOME: IMPACTS ON PHENOTYPES AND EVOLUTION

Christine Tranchant-Dubreuil[1,3], Mathieu Rouard[2,3] and Francois Sabot[1,3]

[1] *DIADE, University of Montpellier, IRD, Montpellier, France*
[2] *Bioversity International, Montpellier, France*
[3] *South Green Bioinformatics Platform, Bioversity, CIRAD, INRA, IRD, Montpellier, France*

Abstract: With the emergence of low-cost high-throughput sequencing technologies, numerous studies have shown that a single genome is not enough to identify all the genes present in a species. Recently, the pangenome concept has become widely used to investigate genome composition of a collection of individuals. The pangenome consists in the core genome, which encompasses all the sequences shared by all the individuals, and the dispensable genome, composed of sequences present in only some individuals. Pangenomic analyses open new ways to investigate and compare multiple genomes of closely related individuals at once, and more broadly new opportunities for optimising breeding and studying evolution. This emerging concept combined with the power of the third-generation sequencing technologies gives unprecedented opportunities to uncover new genes, to fully explore genetic diversity and to advance knowledge about the evolutionary forces that shape genome organisation and dynamics.

Keywords: pangenome, gene diversity, adaptation, evolution, population genomics, structural variation

1 Introduction

Revolutionary advances in high-throughput sequencing technology during the last two decades have offered new ways to study genome diversity and evolution. Limited initially to a few reference genomes,

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

current capabilities allow sequencing and analysis of multiple genomes of closely related species. Indeed, for years genomic studies typically used a reference-centric approach, which relied mainly on the expensive and low-throughput Sanger sequencing, limiting large-scale population studies to a few loci, or to markers such as simple sequence repeats (SSRs) (Zhang and Hewitt, 2003; Schmid et al., 2005). Since the advent of next-generation sequencing (NGS), a transition has occurred from a single-genome/species to multiple-genomes/species analysis. The data deluge produced by these NGS data revealed that individuals from the same species do not systematically share the same genetic content (Redon et al., 2006).

1.1 Genetic Diversity and Structural Variations

Many genetic diversity studies have focused on single-nucleotide polymorphisms (SNPs) as the main source of genetic variation (Cubry et al., 2018; Li et al., 2014; Lin et al., 2014; Qi et al., 2013). However, larger structural variations (SVs), including copy number variation (CNV) and presence/absence variation (PAV), have been shown to play a major role on genetic variability and are thought to contribute to phenotypic variations (Redon et al., 2006; Saxena et al., 2014; Springer et al., 2009). Moreover, even if there are variations within genes, such as SNPs or small insertions or deletions (InDels), several studies showed that all the genes from a given species are not obtained using a single genome (Hurgobin and Edwards, 2017; Monat et al., 2017, 2018). In plants, evidence first from maize (Morgante et al., 2005, 2007) showed that only half of the genomic structure is conserved between two individuals. Similarly, a study of 18 wheat cultivars revealed the absence of 12 150 genes from the reference genome (Montenegro et al., 2017). Previous studies performed on rice showed that genes absent from the Asian rice *Oryza sativa japonica* subspecies are present in other rice varieties (Schatz et al., 2014) and confer tolerance to submergence (*Submergence 1, Sub1*) (Xu et al., 2006), deep water (*SNORKEL1, SNORKEL2*) (Hattori et al., 2009), or low-phosphorus soils (*Phosphorus-Starvation Tolerance, Pstol1*) (Gamuyao et al., 2012). In the same species, Yao et al. (Yao et al., 2015) highlighted that 41.6% of trait-associated SNPs (from GBS markers) were not present in the reference genome sequence. In the wild *Brachypodium distachyon*, the flowering time divergence is directly linked to SV and pangenomic variations (Gordon et al., 2017).

1.2 Origin of the Pangenome Concept

Studies on bacteria benefited earlier by the NGS potential due to their small genome size and large populations, and gave rise to the Pangenome concept, first introduced by Tettelin et al. in 2005 (Medini et al., 2005; Tettelin et al.,

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

2005), to refer to the full genomic content of a species. The pangenome was first defined to consist of the core genome shared among all individuals and the dispensable genome, shared only between some individuals. In plants as in bacteria, the dispensable genome turns out not to be so 'dispensable' (Marroni et al., 2014) and encompasses a large portion of structural variants that affect a large number of genes. The dispensable genome may contribute to phenotypic trait diversity (Saxena et al., 2014) such as biotic resistance, organ size, or flowering time (Gordon et al., 2017) and may play a role in adaptation to various environments. The pangenome view of the genome opens new ways to investigate diversity, adaptation and evolution with strong impacts on the species concept itself.

2 What Is a Pangenome?

Since the pangenome concept was first proposed (Medini et al., 2005; Tettelin et al., 2005), definitions and objectives fluctuated between various interpretations including (i) the total number of nonredundant genes that are present in a given dataset (Guimarães et al., 2015), (ii) the full gene repertoire of a species (Plissonneau et al., 2018), (iii) the result of genomic comparison of different organisms of the same species or genus (Alcaraz et al., 2010; Snipen et al., 2009), (iv) the similarity-based representation of the total set of genes, which are present in a group of closely related species or strains of a single species (Rasko et al., 2008) or (v) the sum of the genes of all living organisms, viruses, and different mobile genetic elements (Tetz, 2005).

2.1 Different Ways to Define a Pangenome

These multiple definitions highlight the flexibility of the pangenome concept and the levels of granularity possible in relation to taxonomy (genus, species, subspecies) or composition of the core genome (e.g. single copy genes versus CNV, gene, and nongenic). Here, we propose to define the pangenome in two different ways: a function based and a structure based. Whatever definition is used, the considered group can be a species, a subspecies, a genera, or even a family. Thus, the limits of a specific pangenome will change if the referential group changes.

2.1.1 Function-based Definition

The function-based definition states that the pangenome is the sum of all genes within a given set of individuals; it can be extended at the gene family level, as in Guimarães et al. (2015). This is similar to the definition used in bacteria and relies on the identification of gene clusters (genes with close

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

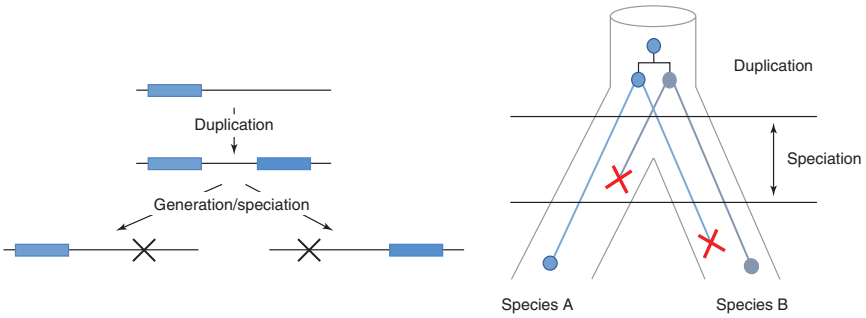

Figure 1 After a recent duplication, through generation or speciation, one individual will conserve the dark blue paralog while the second individual will conserve the light blue paralog.

phylogenetic relationships, that may be scattered all along the genome). Highly similar sequences (e.g. recent paralogs) may be considered as the same sequence. Once all gene clusters per individual are identified, the presence/absence for each gene is scored, wherever the location. In this article, if at least one member of a gene family is present in each individual (whatever is the sequence itself), the function belongs to the core genome. Such a definition is gene-centric and does not take into account transposable elements (TEs) or noncanonical genes (e.g. tRNA, miRNA). Most of the current pangenome analyses use this definition.

2.1.2 Structure-based Definition

The structure-based definition states that the pangenome is defined as the complete set of nonredundant sequences approximately 100 base pairs (bp) in length or more (except for few SNP and InDels, see below) within a given group of individuals. The advantage of this definition is that it allows both genes and nongenic sequences to be taken into account. However, this definition may be difficult to apply when dealing with copies of TEs (see below). Sequences of 100 bp can be identified and annotated with few ambiguities (e.g. the size of a small TE, miRNA locus, or tRNA gene). The presence or absence of a sequence here is purely position based. Thus, in the case of a recent duplication followed by an alternative deletion (i.e. individual A conserves A copy and individual B the B copy; see Figure 1), none of the copies are in the core genome. In the same way, genomic recombination in a portion of the population can change the location of a given region, and thus will not be included in the core genome. Transposition of a Class I (Copy-and-Paste) or of a Class II TE (Cut-and-Paste) (Wicker et al., 2007) will also change the core genome content. Indeed, more and more studies show that the position of these

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

events (recombination and transposition) will impact the expression of adjacent genes (Elgin and Reuter, 2013). Thus, the location of a given sequence may modify its impact on the phenotype, in addition to selection and evolution.

2.2 The Different Compartments of the Pangenome

2.2.1 The Core Genome

The core genome is the common set of sequences shared by all individuals of the group and is generally described as the minimal genome sequence required for a cell to live. Indeed, the core genome has been shown to include the main essential gene functions: (i) maintenance of the basic functions of the organism which include DNA replication, translation, and maintenance of cellular homeostasis (Tettelin et al., 2005), and (ii) essential cellular processes (e.g. glycolysis) (Gordon et al., 2017).

However, some authors (Collins and Higgs, 2012) proposed that the core genome consists of two sub-compartments, one essential and the other 'persistent'. The persistent core genome sub-compartment includes genes or sequences that were perhaps necessary at one time in the life history of an organism but have lost their necessity and have not yet been removed by the genetic drift.

2.2.2 The Dispensable Genome

An unexpectedly large number of sequences, including a surprising number of genes, belong to the dispensable genome (Monat et al., 2017). Thus, PAVs were identified within 38% of genes in the *Brassica napus* pangenome (Hurgobin et al., 2018). Similarly, Zhao et al. (2018) identified 10 872 novel genes (absent from the reference genome) using 66 rice accessions. Among those, several genes detected in previous studies as absent in the reference genome were reported, such as *SUBMERGENCE1A* (*Sub1A*), *SNORKEL1*, and *SNORKEL2* (Hattori et al., 2009; Xu et al., 2006), controlling submergence tolerance, and *PHOSPHORUS-STARVATION TOLERANCE 1* (*Pstol1*), implied in the tolerance to phosphorus-deficient soil (Gamuyao et al., 2012), respectively.

Dispensable genes in bacteria are thought to contribute to diversity and adaptation (Tettelin et al., 2005). In plants, the dispensable genome seems to be enriched in abiotic and biotic stress-related genes, including defence and response, and developmental genes such as those that control flowering time (Golicz et al., 2016b; Gordon et al., 2017; Schatz et al., 2014; Xue et al., 2008; Zhao et al., 2018). Disease resistance-related genes are some of the most prevalent types in the dispensable genome (McHale et al., 2012; Xu et al., 2011). In rice, Schatz et al. showed that 5–12% of the dispensable genes within three divergent genomes contain the NB-ARC domain (nucleotide-binding

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

domain of plant R-genes), versus only 0.35% within the core genome. In *Arabidopsis*, the largest part of the dispensable genome assembly (absent from the Columbia reference) belongs to nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes (Cao et al., 2011). Other gene families that are also enriched in the dispensable genome include auxin- or flowering-related genes, or genes that encode enzymes involved in secondary metabolism (e.g. glucosinolates) (Golicz et al., 2016b). Finally, more 'accessory' functions are linked to the dispensable genome sequences such as telomere maintenance or negative regulation (Gordon et al., 2017). However, those 'accessory' functions can drive major differentiation within a species, as in the wild *B. distachyon* with which flowering time is the main population splitter (Gordon et al., 2017). In this last study, almost 77.6% of the protein-coding genes from the core genome has similarities with known *InterPro* domains, a much higher proportion than that in the dispensable genome set (35.8%). This observation led some authors to suspect that a portion of the PAV genes in the dispensable genome set may be just annotation artefacts or pseudogenes (Contreras-Moreira et al., 2017).

2.3 Individual Specific Genome

The individual-specific compartment contains sequences uniquely detected for one individual and therefore potentially responsible for specific features of the individual. Although this compartment may contain sequences with real biological functions (for highly divergent sequences or neogenes) (Li et al., 2009), many of them are probably artefacts, misannotations, or contaminations. It may also be the result of sampling bias; additional individual data may transfer those sequences to the dispensable genome. Consequently, this compartment might either be merged with the dispensable genome (Li et al., 2014) or discarded for subsequent analyses as in *Brachypodium* analyses (Gordon et al., 2017).

2.4 To Be or Not to Be Core

The core genome is generally considered as the set of sequences common to all individuals of the considered group. However, even if in theory this is a valid definition, due to various limitations (e.g. sampling, sequencing, and technical issues linked to GC-content), some sequences may not be detected in some individuals even though they are present. Thus, we propose that a sequence belongs to the core genome when 90–95% of the individuals harbour it, as published previously (Gordon et al., 2017; Zhao et al., 2018). All sequences not included in the core genome are by default placed in the dispensable genome (Figure 2). Other authors proposed a less strict definition than core and dispensable genomes, using more sophisticated

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

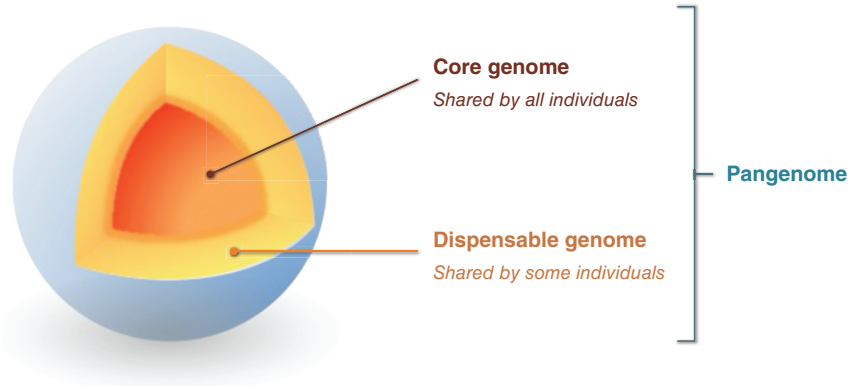

Figure 2 The pangenome is seen as a sphere that contains all the genome of a collection of individuals. The core genome gathers all the common sequences shared by all individuals while the dispensable genome consists of sequences shared by only some individuals.

statistical approaches to define persistent, shell, and cloud levels in the pangenome (Collins and Higgs, 2012). Some sequences may be unique to a single individual, while some may be shared only by less than 90–95% of the group. While individual-specific sequences are most of the time artefacts or contamination, they could indeed be new genes (see below).

From a functional point of view of the pangenome, a gene family will remain in the core if any member of this family is able to perform the function and is present in each individual. As the classification in a given family will depend on the threshold used in its computation, using a functional definition may be complex and may also depend on the clustering method used (e.g. OrthoMCL (Li et al., 2003), MCL (Enright et al., 2002), Mutual Best Hit (Tatusov et al., 1997), GET_HOMOLOGUES-EST (Contreras-Moreira et al., 2017)).

2.5 How Many Genomes to Capture the Whole Genome Content of a Given Group?

For each pangenome analysis, recurrent questions arise: (i) How many genomes should be sequenced to maximise the diversity within a group? (ii) Will there always be the same set of sequences shared by all members of a group even when newly sequenced individuals are added? (iii) Will new specific sequences still be discovered with additional individual sequencing?

In order to validate whether the definitive pangenome size has been reached, Tettelin et al. (2005, 2008) proposed to represent the evolution of

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

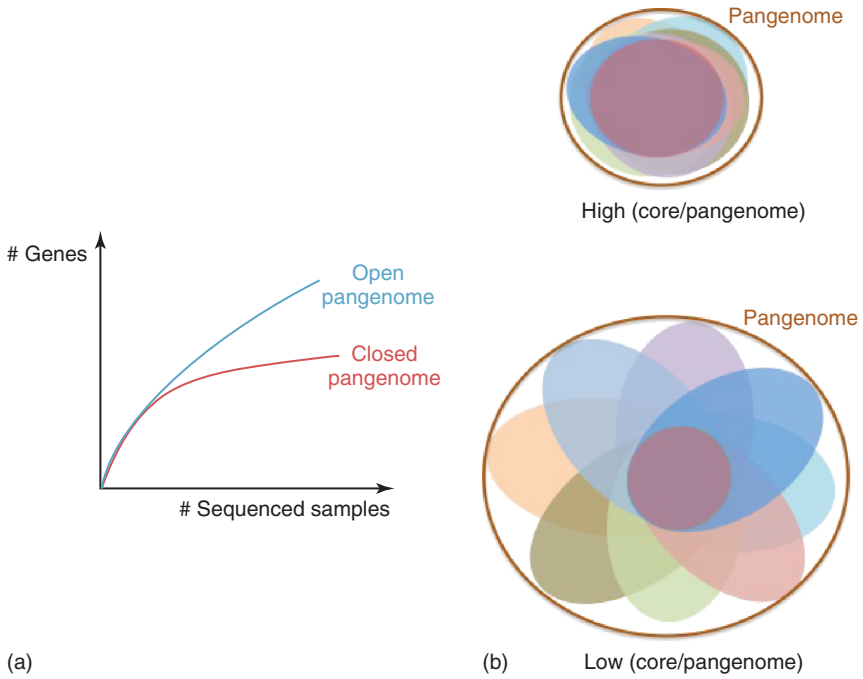

Figure 3 (a) Open and closed pangenomes. (b) C/P ratio illustration.

the total number of sequences found after the addition of each individual sequenced. If the number of sequences levels off to a plateau with each newly added genome, the pangenome is closed. Otherwise, the pangenome is still unlimited and defined as open (Figure 3a).

Tettelin et al showed that the pangenome of the bacteria *S. agalactidae* is very large and open with numerous new unique genes identified even after hundreds of genomes were sequenced (Tettelin et al., 2008). Within plants, the pangenome was shown closed for several models such as soybean, *Brassica oleracea*, maize, or *Medicago* with a small number of samples (Golicz et al., 2016b; Hirsch et al., 2014; Y-h et al., 2014; Zhou et al., 2017). Indeed, the size of the pangenome will depend on the genome dynamics of the considered group; thus, in this regard, bacteria have a relatively larger pangenome than plants because of their higher level of gene flow.

Such observations are generally performed based on gene sequences only (i.e. standard protein-coding genes *a fortiori*), and not on nonprotein genes, neogenes, and TEs. In addition, the sample choice is critically important: an under-representation of the diversity within a given group may indicate that

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

the pangenome is closed, while if the population of individuals sampled is increased, the pangenome may be found to be open. In this case, the largest possible population of individuals should be targeted for sampling (Montenegro et al., 2017).

The core/pangenome ratio seems to be related to an organism's capacity of adaptation (Caputo et al., 2015), with values under 85% showing a huge adaptability (Figure 3B). In plants, the core genome represents from 40% to 80% of the total pangenome, depending on the organism and group's structure (Table 1), indicating a large potential for plants.

3 Methods for Pangenome Assembly

Whichever pangenome definition is used, the first step is to obtain sequences per individual. Up to now, three main approaches have been used in plants to assemble pangenomes.

3.1 Assemble-then-map: Complete *de novo* Assembly Approach

With bacteria, pangenome studies used to complete *de novo* assemblies of small genomes (and their subsequent annotations) (Tettelin et al., 2005, 2008). With plants, most studies (Gan et al., 2011; Gordon et al., 2017; Y-h et al., 2014; Sakai et al., 2014; Schatz et al., 2014; Zhao et al., 2018) also used a similar approach (Figure 4a). With this method, sequences from each individual are assembled separately, then mapped all-versus-all sequences and also to a reference in order to reduce redundancy and to identify shared and nonshared sequences. This approach is time-consuming, requires costly computing and sometimes leads to errors when short read sequencing (e.g. Illumina) is used for large genomes. Indeed, repeated sequences are difficult to resolve using short reads sequences and such assemblies generally result in fragmentation of contigs, leading to a loss of collinearity of fragments. However, the recent and rapid development of long-read sequencing technologies such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences will allow better assemblies and longer contigs, which should resolve the main problem with this approach.

3.2 Metagenomic-like Approach

Yao et al. (2015) combined low-coverage data of around 1500 rice genomes to perform a pangenome assembly using a metagenomic-like approach (Figure 4b). They assembled the whole sequence data together then reassigned the different contigs to individuals through mapping of the single

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

Table 1 Current overview of current plant pangenome studies.

Organism	Sample number	Method	Total pangenes	Core[a]	Dispensable[a]	References
Arabidopsis thaliana	19	Assemble-then-map	37 789	69.7	30.3	Contreras-Moreira et al. (2017)
Brachypodium distachyon	54	Assemble-then-map	37 886	54	46	Gordon et al. (2017)
Brassica oleacera	10[b]	Map-then-assemble	61 379	81.3	18.7	Golicz et al. (2016b)
Capsicum	355	Map-then-assemble	51 757	55.7	44.3	Ou et al. (2018)
Medicago	15	Assemble-then-map	74 700	41.9	58.1	Zhou et al. (2017)
Poplar	22	Mapping only	–	80.7	19.3	Pinosio et al. (2016)
Asian rice	66[b]	Assemble-then-map	42 580	61.9	38.1	Zhao et al. (2018)
Asian rice	453/3000	Assemble-then-map	46 115[c]/47 288[d]	52.9/61.3	47.1/28.7	Wang et al. (2018)
African rice	120 cultivated/ 74 wild	Map-the-assemble	35 198/36 252	86.5/98.6	13.5/1.4	Monat et al. (2018)
Soybean	7[b]	Assemble-then-map	59 080	80.1	19.9	Li et al. (2014)
Bread wheat	18	Map-then-assemble	128 656	64.3	33.7	Montenegro et al. (2017)

[a]Percentage of total pangenes.
[b]Wild and cultivated.
[c]*Indica* subspecies.
[d]*Japonica* subspecies.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

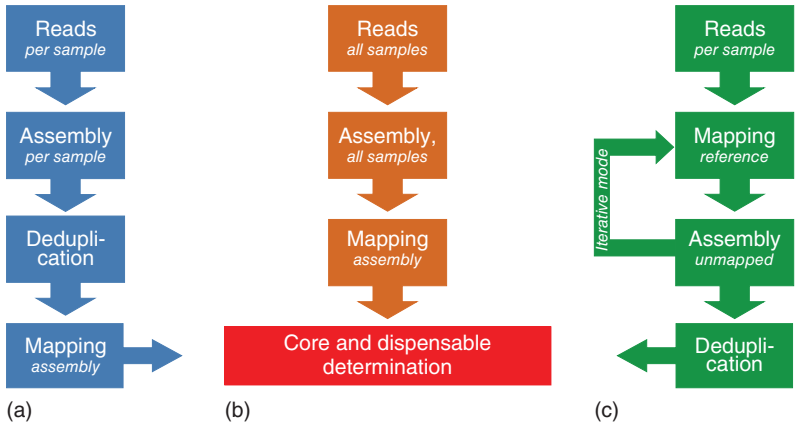

Figure 4 The three approaches for pangenome sequence data assembly: (a) assemble-then-map; (b) metagenomic-like; (c) map-then-assemble. For assemble-then-map and map-then-assemble methods, reduction of redundancy is performed (deduplication step) to identify the common sequences of different individuals.

individual data on their metagenomic assembly. While this allows working with low coverage data from a large number of samples, such an approach may result in chimerical assembly of artefactual sequences.

3.3 Map-then-assemble: Reference-based Approach

The map-then-assemble approach allows to perform individual assemblies after shared sequences are identified (Figure 4c). The idea here is to map all the sequences upon a reference sequence, then to reassemble per individual the unmapped data (Cao et al., 2011; Laine et al., 2019; Monat et al., 2018). An alternative way is to reassemble through an iterative mode (Golicz et al., 2016b), where samples are mapped successively on a panreference, which is updated each turn by the newly assembled sequences. In such a way, shared repeated and complex regions are resolved immediately. Assemblies per individuals are then grouped and deduplicated to avoid redundancy. This approach is less time-consuming than the *de novo* assembly previously described; however, it may impair the detection of recent duplicated sequences. Reads from the two copies that are the result of a recent duplication may map on the single target. In addition, this approach is generally performed using short reads and which may lead to short contigs as described in the sections above for the metagenomic-like or *de novo* assembly approaches.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

3.4 Creating a Panreference

Whatever approach is selected to identify the core and dispensable sequences, the dispensable genome is generally anchored into a reference sequence in order to create a panreference needed for subsequent analyses. Anchoring to a reference sequence may be performed by gene synteny (Gordon et al., 2017) using the collinearity of nearby core genes to identify the position of the dispensable sequences. This method allows to anchor the data precisely, but only if the dispensable sequence is located close to core genes (annotation-based). Another approach is the use of linkage disequilibrium (LD) between genetic markers (e.g. SNPs) on dispensable sequences and on core markers (Yao et al., 2015). It can be faster than gene synteny methods and allows working with nongenic data, but the anchoring is not precise and generally depends on the local LD value. Similarity-based approaches can also be used to identify where the border of dispensable sequences are located within the reference sequence. While this approach can be precise at the single-base scale, in the case of repeated sequences, the similarity can occur with multiple regions and the exact location between all these similar regions may be difficult to distinguished.

3.5 Annotation of the Core and Dispensable Genome

As for any classical single-sequence genome annotation, sequences can be annotated to provide functions. Dispensable and core gene sequences are generally clustered using methods coming from comparative genomics: pair-wise BLASTP or MCL tools (e.g. OrthoMCL (Li et al., 2003), OrthoFinder (Emms and Kelly, 2015)), and clustering with GET_HOMOLOGUES-EST approaches (Contreras-Moreira et al., 2017). The stringency of the clustering level used here will heavily influence the results. Different studies used different thresholds, and these thresholds for clustering will mainly depend on the genetic diversity of the considered group. For instance, a highly recently diverged group will be analysed using a high threshold (up to 95% of similarity), while an older diverged group will use a more relaxed threshold (80% e.g.). For nongenic elements, such as TEs or miRNAs, the annotation will also be performed as with classical genomic annotation, using state-of-the-art tools (Ewing, 2015; Rishishwar et al., 2016).

With bacteria, the pangenomic analyses generally rely on gene annotation and gene family clustering. With plants, no specific trend (gene family or structure or synteny) has been clearly adopted by the community, and authors tend to combine several approaches within the same study (Gordon et al., 2017).

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

4 Dynamics of Pangenome Compartments

The ability of the pangenome size to increase or to be stable, as well as switching from core to dispensable and reversely, is strongly connected to the balance between gain and loss events and the ability to adapt to diverse environments (Guimarães et al., 2015). Different factors and forces can impact on the pangenome structure, including gene birth and death, horizontal transfers (HTs) and TE activity (Figure 5).

4.1 Gene Birth-and-death Processes

Gene creation and elimination can occur as a result of different processes, including errors during recombination that eliminate genes, TEs that mediate gene duplication, duplication events that result in gene gains, followed by diversification and neofunctionalization (Gordon et al., 2017). There is evidence that most of the dispensable genes may arise from these gene birth-and-death mechanisms. Unique protein-coding genes may emerged from (i) noncoding DNA (*de novo* genes), (ii) an older coding sequence by a combination of mechanisms such as duplication followed by rapid divergence, horizontal gene transfer (HGT), or ancient gene lost with important sequence variation followed by neofunction, or exapted transposon (domesticated by the host genome to provide a new biological function) (Arendsee et al., 2014; Schlötterer, 2015). Dispensable genes identified in several studies tend to display common features similar to young genes: short gene, weak Interpro homology, low expression, rapid evolution and turnover (Golicz et al., 2016a, b; Schatz et al., 2014; Stein et al., 2018). Several studies showed a regulatory role of these genes in response to numerous varying environmental conditions, biotic (Xiao and Wenfei, 2009) or abiotic stresses (Li et al., 2009), and potentially also in death gene processes (Zhou et al., 2017).

4.2 Transposable Elements, Umpires, and Players

Ubiquitous in all eukaryotic genomes, TEs represent a major part of many plant genomes. TEs are endogenous genomic elements able to duplicate themselves and to insert elsewhere in a host genome. They use different strategies to move, including RNA (retrotransposons, Class I) or DNA intermediates (DNA transposons, Class II) (Wicker et al., 2007). The Copy-and-Paste amplification strategy used by retrotransposons allows them to accumulate in the genome at a high-copy numbers, with the result that some TE families can represent a predominant part of a genome. For instance, 85% of the maize (*Zea mays*) genome consists of TEs, mainly

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

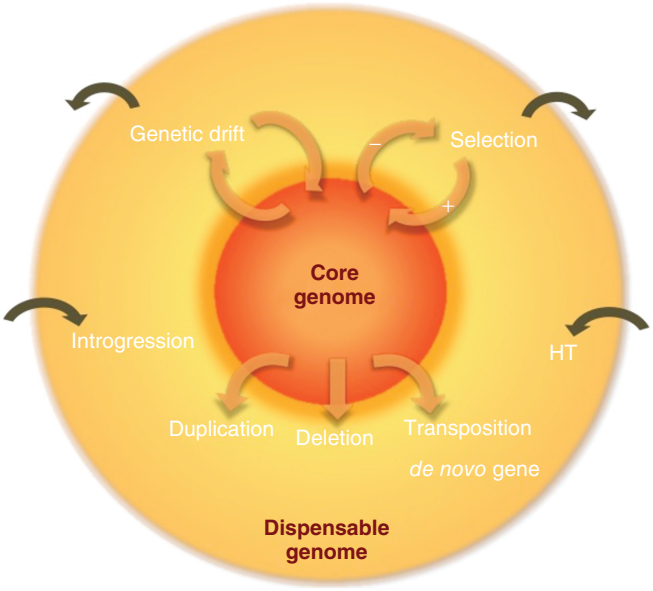

Figure 5 Dynamic overview of the pangenome structure shaped by different events and forces. New sequences are added to the dispensable genome through mutations, duplications, deletions, and transpositions, while the core genome content may decrease by deletion and transposition. Horizontal transfer and introgression also impact on the dispensable genome compartment (sequence gain). Moreover, positive and purifying selection as well as genetic drift impact on the core and dispensable genomes (sequence gain and loss) as well as on the pangenome (sequence loss).

LTR Retrotransposons (Schnable et al., 2009). TE movements are at the origin of numerous genomic variations within species (Morgante et al., 2007; Piegu et al., 2006). For instance, in maize, the activity of TEs, especially Helitron-like elements, was able to modify up to 50% of the genome structure in a vast majority of the collinear BACs analysed (Morgante et al., 2005, 2007). Owing to their ability to change location within the host genome (Wicker et al., 2007), they are the first candidates for dispensable genome creation (Morgante et al., 2007; Schatz et al., 2014), as every new insertion will belong to this compartment. Finally, their ability to spread through a population at a high rate allows them to invade even the core genome of species, such as the *P* element in different *Drosophila* species (Kofler et al., 2015).

Beyond their own intrinsic activity, the presence of TEs at a given position may alter the pangenomic structure. Golicz et al. (2016b) observed a higher

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

TE density surrounding variable genes in *Brassica oleacerea*. In the same way, Gordon et al. (2017) showed that noncore genes are more linked to TE activity than core genes. TEs can alter the expression of surrounding genes (Butelli et al., 2012; Hirsch and Springer, 2017) but also the global genome structure by serving as anchor for illegitimate recombination (Chantret et al., 2005).

4.3 Horizontal Transfers

It has been shown within bacteria the importance of HGT (Soucy et al., 2015) and its impact on the pangenome (Koonin, 2016). HGT has been shown in eukaryotes (Keeling and Palmer, 2008; Zhang et al., 2012), sometimes with a high success and essential functions, and can be selected to become a core gene. HTs from nongenic elements such as TEs (Brookfield, 2005; El Baidouri et al., 2014) may also impact the dispensable genome and could invade the host genome in a very short period of time (Kofler et al., 2015), and on a large array of species (Roulin et al., 2007, 2009).

5 Challenges, Perspectives, and the Way Forward

5.1 Links and Impacts on Phenotype

Dispensable genes are thought to be responsible for considerable phenotypic variation that could be suitable for breeding improved crop varieties and evolutionary studies of adaptive traits (Gordon et al., 2017). SVs (including CNV and PAV) can significantly have an impact on phenotypic variation in plants. For instance, Lu et al. (2015) investigated the contribution of PAVs to phenotypic variance using GWAS on four traits in maize and SNPs located in nonreference sequences were found enriched in the significant GWAS hits compared to reference-based SNPs, indicating their possible role in such variation. In the same way, in a study with rice, more than 40% of the agronomical traits were linked to nonreference sequences, thus dispensable (Yao et al., 2015). In addition, in wild African rice *O. barthii*, the PAV of the PROSTATE GROWTH 1 (*PROG1*) gene directly impacts global plant phenotype: when absent, plants are erect, while when present plants are not erect (Cubry et al., 2018). The absent state seems to have been selected in the cultivated relative *Oryza glaberrima*. Many other examples exist in rice and other crops that show numerous phenotypes of interest are not linked to SNPs but to PAVs.

5.2 Adaptation, Selection, and Speciation

The pangenome concept offers new perspectives to increase our knowledge about evolutionary mechanisms that allow organisms to adapt quickly to

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

new environments. Indeed, more and more studies show adaptive phenotypic changes in plants for various traits due to CNVs (e.g. flowering time (Díaz et al., 2012; Würschum et al., 2015), pest and diseases resistance (Cook et al., 2012; Hardigan et al., 2016), herbicide resistance (Patterson et al., 2017), plant height (Li et al., 2012)). The ability to acquire new genes and to generate gene allelic diversity has various potential effects including neutral, adaptive, or not on fitness (McInerney et al., 2017; Vos and Eyre-Walker, 2017). Pangenome analysis offers new ways to investigate the adaptation processes and to understand their impacts on the core and dispensable genomes. It would be particularly interesting to focus on different periods of divergence within a given group, such as speciation, when effective population size is small, genetic drift effect is important, and events such as the reproductive isolation is occurring. It was shown for some species that speciation will impact drastically the pangenomic structure (Golicz et al., 2016a, b; Monat et al., 2018). This will lead to additional questions and possibilities, such as what is a species in perspective of the pangenome? Is having the same core genome enough to be from the same species, or is it too restrictive or relaxed?

5.3 Graphical Visualisation

Graphical tools have been developed to handle and display bacterial pangenome datasets such as PanX (Ding et al., 2018), pan-Tetris (Hennig et al., 2015), PanViz (Pedersen et al., 2017), seq-seq-pan (Jandrasits et al., 2018), and PanACEAE (Clarke et al., 2018). However, fewer have been proposed for plant genomes including Rpan (Sun et al., 2017) and Brachypan (Gordon et al., 2017). Generally, publications on the topic have revisited Venn diagram or flower plot like representations (Collingro et al., 2011; Kant et al., 2014; Nourdin-Galindo et al., 2017; Paul et al., 2015) (Figure 6a) to illustrate PAV, but this representation has limitations with increased sample size, leading to the possibility of alternate visualisation tools, such as Upset (Lex et al., 2014). Various approaches are emerging using graph-based structures (Garrison et al., 2017; Paten et al., 2017) (genome and variation graphes; Figure 6b), for example using the de Bruijn Graph Algorithm. The main challenge remains to design scalable solutions for large panels of samples able to support PAV-based functional analyses that allow to zoom into chromosome segments to visualise individual SVs and SNPs for structural-based analyses. However, such comprehensive systems are still in their infancy and are not yet operational. Whatever solutions will be developed as reference tools, it would be recommended to build them upon existing systems with powerful capacity to explore genomes and variants, or at least to enable interoperability between them.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

5.4 Expected Contribution of Recent Sequencing Approaches

Conformation capture methods such as Hi-C, mate-pairs libraries, or 10x synthetic long-reads are second generation technology-based approaches that can be used in the near future to resolve panreference assemblies. Besides, third-generation sequencing technologies such as Pacific BioSciences SMRT or ONT offer single-strand long-reads sequences (up to 2 Mb, the current ONT record so far). These methods, and especially the low-cost Minion from ONT, will change the paradigm of one high-quality reference sequence and many draft sequence samples. Indeed, a golden-standard quality sequence can be performed for less than 1000 USD for genomes of around 150 Mb (Miller et al., 2018), and 1300 USD for a rice-sized genome (400 Mb; F. Sabot, unpublished data). Thus, the assemble-then-map approach (Figure 4) may become the standard for future pangenomic approaches. Indeed, the capacity of long-read assemblies to overcome the repeat sequence paradox and to solve the scaffolding difficulties will make this technology the best tool for pangenome analysis. Advantages of a portable solution such as Minion will allow rapid sequence capacities in any lab with any sample of interest to identify PAVs or CNVs of interest.

5.5 Conclusions and Future Perspectives

The pangenome concept combined with high-throughput third-generation sequencing will probably allow access to large gene repertoires of the wild relatives of cultivated plants, particularly interesting for crucial agronomic traits such as drought and salinity tolerance. Genome portals will have to evolve from a reference genome-centric view to adopt a pangenome reference view or to manage multiple reference assemblies with a granular level of display with standardised genome assembly and gene models nomenclature. Using dispensable genome data will allow identifying the genetic basis of phenotypes of interest in dedicated lines.

5.5.1 Summary Points

1 Pangenome view of a genome opens new challenging ways to explore genetic diversity, adaptation, and evolutionary mechanisms within a group.
2 Pangenome analyses give access to a surprisingly large reservoir of genes/sequences never identified when working on a single reference genome.
3 Increased knowledge of dispensable genes may be of high importance for breeding applications.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

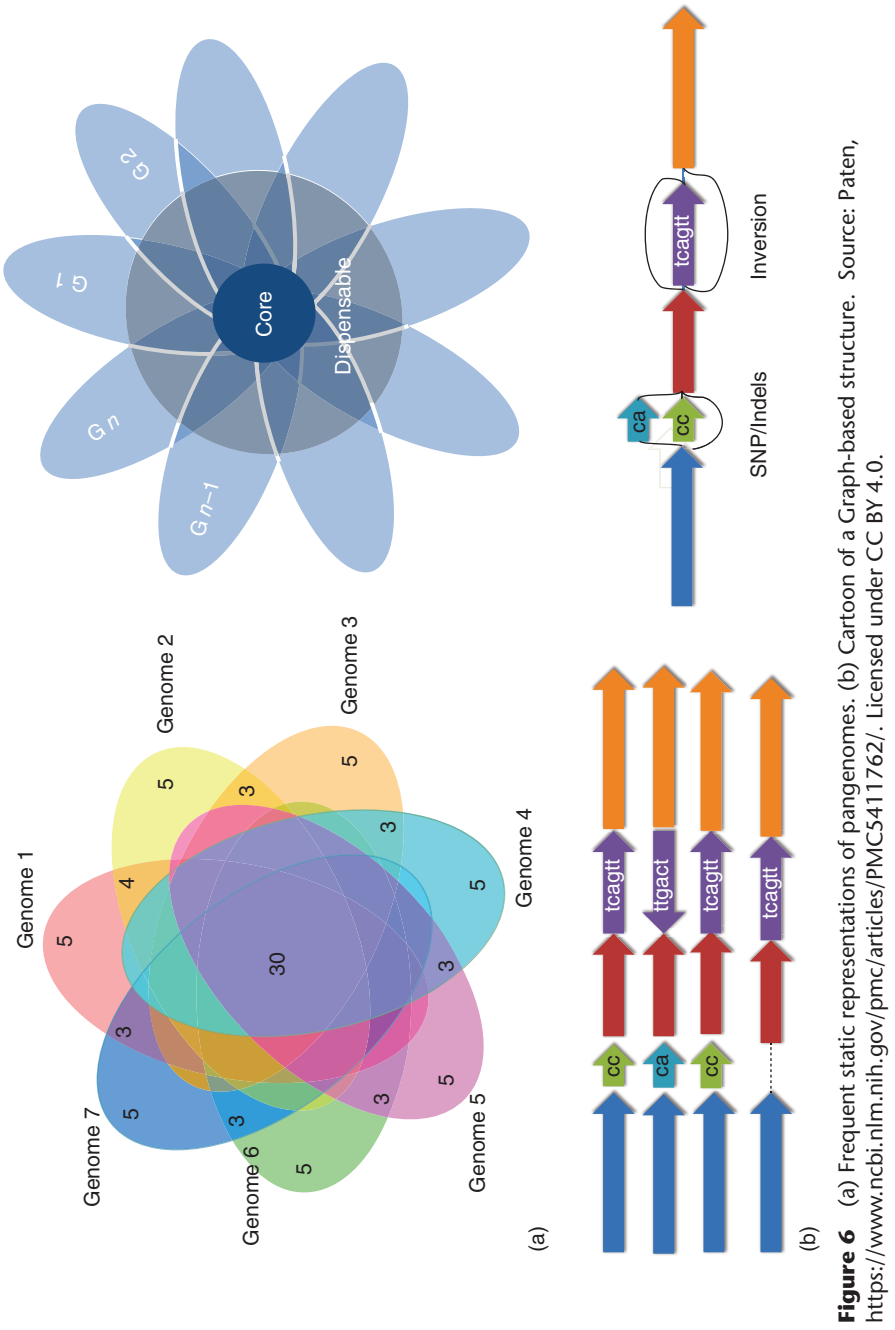

Figure 6 (a) Frequent static representations of pangenomes. (b) Cartoon of a Graph-based structure. Source: Paten, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5411762/. Licensed under CC BY 4.0.

Acknowledgements

Authors want to thank their respective host institutions for funding as well as the CGIAR Research Programs (CRP) on Roots, Tubers, and Bananas (RTB) and RICE.

Related Articles

Comparative Genomics
Genomics, Adaptation, and the Evolution of Plant Form
Comparative Evolutionary Genomics of Land Plants
Evolution and Taxonomy of the Grasses (Poaceae): A Model Family for the Study of Species-Rich Groups

Glossary

Core genome	Genes or sequences present in all individuals (or almost) in a given group.
Dispensable genome	Genes or sequences not present in all individuals in a given group.
Genetic drift	One of the main evolutionary mechanisms leading to the change in gene frequencies over the time due to chance or random sampling effect.
Negative or purifying selection	Evolutionary process that remove deleterious allele leading to the decrease of its frequency in a population.
Pangenome	The ensemble of the core genome and of the dispensable genome. It represents the whole set of sequences available within a given group.
Positive or diversifying selection	Evolutionary process driving to the increase in prevalence of a new advantageous allele in a population.

References

Alcaraz, L.D., Moreno-Hagelsieb, G., Eguiarte, L.E. et al. (2010). Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics. *BMC Genomics* **11**: 332.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

Arendsee, Z.W., Li, L., and Wurtele, E.S. (2014). Coming of age: orphan genes in plants. *Trends in Plant Science* **19** (**11**): 698–708.

Brookfield, J.F.Y. (2005). The ecology of the genome - mobile DNA elements and their hosts. *Nature Reviews. Genetics* **6** (**2**): 128–136.

Butelli, E., Licciardello, C., Zhang, Y. et al. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *The Plant Cell* **24** (**3**): 1242–1255.

Cao, J., Schneeberger, K., Ossowski, S. et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics* **43**: 956.

Caputo, A., Merhej, V., Georgiades, K. et al. (2015). Pan-genomic analysis to redefine species and subspecies based on quantum discontinuous variation: the Klebsiella paradigm. *Biology Direct* **10** (**1**): 55.

Chantret, N., Salse, J., Sabot, F. et al. (2005). Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *The Plant Cell* **17** (**4**): 1033–1045.

Clarke, T.H., Brinkac, L.M., Inman, J.M. et al. (2018). PanACEA: a bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes. *BMC Bioinformatics* **19** (**1**): 246.

Collingro, A., Tischler, P., Weinmaier, T. et al. (2011). Unity in variety—the pan-genome of the Chlamydiae. *Molecular Biology and Evolution* **28** (**12**): 3253–3270.

Collins, R.E. and Higgs, P.G. (2012). Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Molecular Biology and Evolution* **29** (**11**): 3413–3425.

Contreras-Moreira, B., Cantalapiedra, C.P., Garcia-Pereira, M.J. et al. (2017). Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. *Frontiers in Plant Science* **8**: 184.

Cook, D.E., Lee, T.G., Guo, X. et al. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science* **338** (**6111**): 1206–1209.

Cubry, P., Tranchant-Dubreuil, C., Thuillet, A.-C. et al. (2018). The rise and fall of African rice cultivation revealed by analysis of 246 new genomes. *Current Biology* .

Díaz, A., Zikhali, M., Turner, A.S. et al. (2012). Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS One* **7** (**3**): e33234.

Ding, W., Baumdicker, F., and Neher, R.A. (2018). panX: pan-genome analysis and exploration. *Nucleic Acids Research* **46** (**1**): e5.

El Baidouri, M., Carpentier, M.-C., Cooke, R. et al. (2014). Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Research* **24** (**5**): 831–838.

Elgin, S.C. and Reuter, G. (2013). Position-effect variegation, heterochromatin formation, and gene silencing in *Drosophila*. *Cold Spring Harbor Perspectives in Biology* **5** (**8**): a017780.

Emms, D.M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16** (**1**): 157.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30** (**7**): 1575–1584.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

Ewing, A.D. (2015). Transposable element detection from whole genome sequence data. *Mobile DNA* **6** (**1**): 24.

Gamuyao, R., Chin, J.H., Pariasca-Tanaka, J. et al. (2012). The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature* **488** (**7412**): 535–539.

Gan, X., Stegle, O., Behr, J. et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**: 419.

Garrison, E., Sirén, J., Novak, A.M. et al. (2017). Sequence variation aware genome references and read mapping with the variation graph toolkit. *bioRxiv* 234856.

Golicz, A.A., Batley, J., and Edwards, D. (2016a). Towards plant pangenomics. *Plant Biotechnology Journal* **14** (**4**): 1099–1105.

Golicz, A.A., Bayer, P.E., Barker, G.C. et al. (2016b). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications* **7**: 13390.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P. et al. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications* **8** (**1**): 2184.

Guimarães, L.C., Florczak-Wyspianska, J., de Jesus, L.B. et al. (2015). Inside the pan-genome - methods and software overview. *Current Genomics* **16** (**4**): 245–252.

Hardigan, M.A., Crisovan, E., Hamiltion, J.P. et al. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *The Plant Cell* **28** (**2**): 388–405. doi: 10.1105/tpc.15.00538.

Hattori, Y., Nagai, K., Furukawa, S. et al. (2009). The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. *Nature* **460** (**7258**): 1026–1030.

Hennig, A., Bernhardt, J., and Nieselt, K. (2015). Pan-tetris: an interactive visualisation for pan-genomes. *BMC Bioinformatics* **16** (**Suppl 11**): S3.

Hirsch, C.D. and Springer, N.M. (2017). Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1860** (**1**): 157–165.

Hirsch, C.N., Foerster, J.M., Johnson, J.M. et al. (2014). Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell* **26** (**1**): 121–135.

Hurgobin, B. and Edwards, D. (2017). SNP discovery using a pangenome: has the single reference approach become obsolete? *Biology* **6** (**1**): E21.

Hurgobin, B., Golicz, A.A., Bayer, P.E. et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal* **16** (**7**): 1265–1274.

Jandrasits, C., Dabrowski, P.W., Fuchs, S., and Renard, B.Y. (2018). Seq-seq-pan: building a computational pan-genome data structure on whole genome alignment. *BMC Genomics* **19** (**1**): 47.

Kant, R., Rintahaka, J., Yu, X. et al. (2014). A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*. *PLoS One* **9** (**7**): e102762.

Keeling, P.J. and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* **9** (**8**): 605–618.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

Kofler, R., Hill, T., Nolte, V. et al. (2015). The recent invasion of natural *Drosophila simulans* populations by the p-element. *Proceedings of the National Academy of Sciences of the United States of America* **112** (**21**): 6659–6663.

Koonin, E.V. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research* **5**. pii: F1000 Faculty Rev-1805.

Laine, V., Gossmann, T.I., van Oers, K. et al. (2019). Exploring the unmapped DNA and RNA reads in a songbird genome. *BMC Genomics* **20**: 19.

Lex, A., Gehlenborg, N., Strobelt, H. et al. (2014). UpSet: visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics* **20** (**12**): 1983–1992.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13** (**9**): 2178–2189.

Li, L., Foster, C.M., Gan, Q. et al. (2009). Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *The Plant Journal* **58** (**3**): 485–498.

Li, Y., Xiao, J., Wu, J. et al. (2012). A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. *New Phytologist* **196** (**1**): 282–291.

Li, J.-Y., Wang, J., and Zeigler, R.S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience* **3** (**1**): 8.

Lin, T., Zhu, G., Zhang, J. et al. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics* **46**: 1220.

Lu, F., Romay, M.C., Glaubitz, J.C. et al. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications* **6**: 6914.

Marroni, F., Pinosio, S., and Morgante, M. (2014). Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology* **18**: 31–36.

McHale, L.K., Haun, W.J., Xu, W.W. et al. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology* **159** (**4**): 1295–1308.

McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have pangenomes. *Nature Microbiology* **2**: 17040.

Medini, D., Donati, C., Tettelin, H. et al. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development* **15** (**6**): 589–594.

Miller, D.E., Staber, C., Zeitlinger, J., and Hawley, R.S. (2018). High-quality genome assemblies of 15 *Drosophila* species generated using nanopore sequencing. *G3 (Bethesda)*. **8** (**10**): 3131–3141.

Monat, C., Pera, B., Ndjiondjop, M.-N. et al. (2017). De novo assemblies of three *Oryza glaberrima* accessions provide first insights about pan-genome of African rices. *Genome Biology and Evolution* **9** (**1**): 1–6.

Monat, C., Tranchant-Dubreuil, C., Engelen, S. et al. (2018). Comparison of two African rice species through a new pan-genomic approach on massive data. *bioRxiv* . doi: 10.1101/245431.

Montenegro, J.D., Golicz, A.A., Bayer, P.E. et al. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal* **90** (**5**): 1007–1013.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

Morgante, M., Brunner, S., Pea, G. et al. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics* **37** (**9**): 997–1002.

Morgante, M., Paoli, E.D., and Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology* **10** (**2**): 149–155.

Nourdin-Galindo, G., Sánchez, P., Molina, C.F. et al. (2017). Comparative pan-genome analysis of *Piscirickettsia salmonis* reveals genomic divergences within genogroups. *Frontiers in Cellular and Infection Microbiology* **7**: 459.

Ou, L., Li, D., Lv, J. et al. (2018). Pan-genome of cultivated pepper (Capsicum) and its use in gene presence–absence variation analyses. *New Phytologist* **220** (**2**): 360–363.

Paten, B., Novak, A.M., Eizenga, J.M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research* **27** (**5**): 665–676.

Patterson, E.L., Pettinga, D.J., Ravet, K. et al. (2017). Glyphosate resistance and EPSPS gene duplication: convergent evolution in multiple plant species. *Journal of Heredity* **109** (**2**): 117–125.

Paul, S., Bhardwaj, A., Bag, S.K. et al. (2015). PanCoreGen — profiling, detecting, annotating protein-coding genes in microbial genomes. *Genomics* **106** (**6**): 367–372.

Pedersen, T.L., Nookaew, I., Wayne Ussery, D., and Månsson, M. (2017). PanViz: interactive visualization of the structure of functionally annotated pangenomes. *Bioinformatics* **33** (**7**): 1081–1082.

Piegu, B., Guyot, R., Picault, N. 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** (**10**): 1262–1269.

Pinosio, S., Giacomello, S., Faivre-Rampant, P. et al. (2016). Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular Biology and Evolution* **33** (**10**): 2706–2719.

Plissonneau, C., Hartmann, F.E., and Croll, D. (2018). Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biology* **16** (**1**): 5.

Qi, J., Liu, X., Shen, D. et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nature Genetics* **45**: 1510.

Rasko, D.A., Rosovitz, M., Myers, G.S.A. et al. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology* **190** (**20**): 6881–6893.

Redon, R., Ishikawa, S., Fitch, K.R. et al. (2006). Global variation in copy number in the human genome. *Nature* **444** (**7118**): 444–454.

Rishishwar, L., Mariño-Ramírez, L., and Jordan, I.K. (2016). Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics* **18** (**6**): bbw072.

Roulin, A., Piegu, B., Wing, R.A., and Panaud, O. (2007). Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus Oryza. *The Plant Journal* **53** (**6**): 950–959.

Roulin, A., Piegu, B., Fortune, P.M. et al. (2009). Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in Poaceae. *BMC Evolutionary Biology* **9** (**1**): 58.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

Sakai, H., Kanamori, H., Arai-Kichise, Y. et al. (2014). Construction of pseudomolecule sequences of the aus rice cultivar Kasalath for comparative genomics of Asian cultivated rice. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* **21** (**4**): 397–405.

Saxena, R.K., Edwards, D., and Varshney, R.K. (2014). Structural variations in plant genomes. *Briefings in Functional Genomics* **13** (**4**): 296–307.

Schatz, M.C., Maron, L.G., Stein, J.C. et al. (2014). Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biology* **15** (**11**): 506.

Schlötterer, C. (2015). Genes from scratch–the evolutionary fate of de novo genes. *Trends in genetics: TIG* **31** (**4**): 215–219.

Schmid, K.J., Ramos-Onsins, S., Ringys-Beckstein, H. et al. (2005). A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169** (**3**): 1601–1615.

Schnable, P.S., Ware, D., Fulton, R.S. et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* **326** (**5956**): 1112–1115.

Snipen, L., Almøy, T., and Ussery, D.W. (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics* **10**: 385.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* **16** (**8**): 472–482.

Springer, N.M., Ying, K., Fu, Y. et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics* **5** (**11**): e1000734.

Stein, J.C., Yu, Y., Copetti, D. et al. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nature Genetics* **50** (**2**): 285–296.

Sun, C., Hu, Z., Zheng, T. et al. (2017). RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Research* **45** (**2**): 597–605.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. *Science* **278** (**5338**): 631–637.

Tettelin, H., Masignani, V., Cieslewicz, M.J. et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences* **102** (**39**): 13950–13955.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* **11** (**5**): 472–477.

Tetz, V.V. (2005). The pangenome concept: a unifying view of genetic information. *Medical Science Monitor* **11** (**7**): HY24–HY29.

Vos, M. and Eyre-Walker, A. (2017). Are pangenomes adaptive or not? *Nature Microbiology* **2** (**12**): 1576.

Wang, W., Mauleon, R., Hu, Z. et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557** (**7703**): 43–49.

Wicker, T., Sabot, F., Hua-Van, A. et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**: 973.

Würschum, T., Boeven, P.H.G., Langer, S.M. et al. (2015). Multiply to conquer: copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genetics* **16** (**1**): 96.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

Xiao, H.A.L. and Wenfei, A.N.D.L. (2009). A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS One* **4** (**2**): 1–12.

Xu, K., Xu, X., Fukao, T. et al. (2006). Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* **442** (**7103**): 705–708.

Xu, X., Liu, X., Ge, S. et al. (2011). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology* **30**: 105.

Xue, W., Xing, Y., Weng, X. et al. (2008). Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nature Genetics* **40**: 761.

Yao, W., Li, G., Zhao, H. et al. (2015). Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biology* **16**: 187.

Y-h, L., Zhou, G., Ma, J. et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology* **32** (**10**): 1045–1052.

Zhang, D.X. and Hewitt, G.M. (2003). Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology* **12** (**3**): 563–584.

Zhang, D., Iyer, L.M., and Aravind, L. (2012). Bacterial GRAS domain proteins throw new light on gibberellic acid response mechanisms. *Bioinformatics (Oxford, England)* **28** (**19**): 2407–2411.

Zhao, Q., Feng, Q., Lu, H. et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics* **50** (**2**): 278–284.

Zhou, P., Silverstein, K.A.T., Ramaraj, T. et al. (2017). Exploring structural variation and gene family architecture with de novo assemblies of 15 medicago genomes. *BMC Genomics* **18** (**1**): 261.

Annual Plant Reviews Online, Volume 2. Edited by Jeremy Roberts.

7- Perspectives and Conclusion

Doc, I'm from the future. I came here in a time machine that you invented. Now I need your help to get back to my time – Marty Mcfly, Back to the Future

Beyond reviewing the state-of-the-art of pangenomic analyses particularly in plants, the review looked at the prospects and challenges we will face (in 2019) after highlighting a number of key points in this exciting new era of paradigm shift from one genome to pangenome that we are still experiencing today. Even if the review did not explicitly mention this paradigm shift, it pointed out that a single reference genome was insufficient to capture all diversity and that by focusing diversity analyses on SNPs and a single reference genome, a large part of the diversity was ignored. Relatively obvious today, the role of structural variations as a major player in diversity has begun to be showed up, first through studies targeting specific metabolic pathways or genes affected by structural variations (Gamuyao et al. 2012; Wang, Xiong, et al. 2015; Xu, Xu, et al. 2006) and then progressively through larger-scale studies, whether by resequencing individuals, comparing new genomes assembled or pangenomic analysis (Gao et al. 2019; Montenegro et al. 2017; Yao et al. 2015). In addition, although it would have merited its own in-depth section, we nevertheless cited several examples of phenotypic traits associations related to structural variations, which are usually excluded from "classical" diversity analyses based on SNPs alone. Thus, pangenomics enables discovery of a large number of structural within a population, previously hidden by the use of a single reference and reveals how structural variations have shaped the genomic landscape of (pan)genome.

In this review, we also have summarized the current whole of knowledge and approaches to decipher both the structure of pangenomes within a species or across species and the dynamics of pangenomes as well as how pangenome structure is shaped by evolutionary process and events. Pangenomics provides also a more comprehensive way to better understanding of the evolutionary mechanisms that allow organisms to adapt to new environments and to investigate the process of adaptation and selection not only within a species but also between species or even at the level of a given genus. This will lead to other questions related to the notion of species such as what is a species from a pangenomic point of view.

The pangenome concept, combined with third-generation sequencing technologies and conformation capture methods such as Hi-C, offers new opportunities to address biological questions from a new perspective. However, pangenomics will face new challenges in analyzing, storing and visualizing the mass of data generated by this newly developing approach. All these ideas will be further developed in the last chapter.

To conclude

Writing this review with François Sabot and Mathieu Rouard has been very enriching, allowing us to confront our ideas on this new field (but not only) and to approach this concept from different angles : methodology, biology, genetics (a little) and evolution. I also immersed myself in this relatively new concept but so exciting, opening a new way to better understand plant and animal diversity.

Based on this state-of-art, we then developed an approach to build the first african rice panreference, which we describe in the next chapter.

II How to build the African Rice Pangenome ?

8- Strategy and main results

If you know you are on the right track, if you have this inner knowledge, then nobody can turn you off... no matter what they say – Barbara McClintock

8.1 Developing a tool for building an eukaryotic pangenome

Pangenomics offers a more comprehensive way to study genetic diversity in a population across species or genera. To date, pangenomics studies in plants have revealed a surprising variation in gene content from 7% to 81% of dispensable genes (Table 3.1, p. 44, Torkamaneh, Lemay, and Belzile 2021, Yang, Liu, et al. 2022) although this ratio may be impacted by factors such as the population sampling and the number of samples, as discussed in the part II (section 5.2.3, page 53).

One of the key steps in performing a pangenomic analysis is to detect the full content of variations in a population, *i.e.* to build the pangenome, in order to define what is core or variable in a second step. This primordial step will be even more difficult if one wants to detect all variations (genic or not), over a large number of individuals. If, in addition, this analysis is carried out on Eukaryotes, which have large and complex genomes (eg. high repeat content, polyploidy), the construction of the pangenome is an even more of a complex and time-consuming process. These last years, third-generation sequencing technologies, such as PacBio and Oxford Nanopore, combined with Hi-C or BioNano approaches, have enabled the sequencing and assembly of larger, complex genomes at high resolution (Cheng et al. 2021) such as the 14.66-Gb high-quality assembly of a South african bread wheat (Athiyannan et al. 2022), the 3.1-Gb high-quality haplotype-resolved assembly of an autotetraploid potato genome (Sun, Jiao, et al. 2022) or the 25.4-Gb high-quality genome of Chinese pine (Niu et al. 2022). These technologies cope well with large structural variations, but remain expensive for plants to characterize the pangenome from a population of hundreds of individuals, which have large genomes and require high sequencing depth. In addition, the large volumes of short reads data available on many species offer opportunities to perform large-scale pangenomics analyses but very few tools are available to perform such analyses especially on eukaryotes.

As described in the part I (section 3.2, page 40), pangenome construction from short read sequencing technology is mainly based on two approaches:

- The ***assemble-then-map approach*** starts with the *de novo* assembly of several genomes followed by genomes comparison (*i.e.* mapping of contigs against reference genome or pair-wise genome comparison) (Gao et al. 2019; Gordon et al. 2017; Hu et al. 2017);
- The ***map-then-assemble approach*** is based on the short-reads mapping followed by the *de novo* assembly of the unmapped reads (Golicz, Bayer, et al. 2016; Hufnagel et al. 2021; Sherman et al. 2019).

Very few tools are publicly available to perform all the steps at once, being either developed for bacteria (Ding, Baumdicker, and Neher 2018; Laing et al. 2010; Page et al. 2015), or based on the *de novo* 'assemble-then-map' approach (Hu et al. 2017). The method we developed is based on the 'map-then-assemble' approach which consists in:

1. identifying large fragments absent from a reference genome using short reads data;
2. placing these variations on the reference;
3. building a pan-reference.

From an approach to a generic tool

The FrangiPANe tool was implemented to easily apply our method based on the 'map-then-assemble' approach and to create an accurate panreference for any organism from short reads data. Our method and frangiPANe were validated using the whole genome sequencing data from 248 African Rice accessions both cultivated and wild, as a proof-of-concept to build the first pan-reference for African rice.

The analysis process is divided into five steps. After pair-ended short reads mapping against a reference genome, unmapped reads are *de novo* assembled for each sample and all contigs are pooled after filtering. The last two steps consist in reducing the sequence redundancy at intra- and inter-population level and into placing the non-redundant sequences on the reference genome (Figure 8.1, page 88).

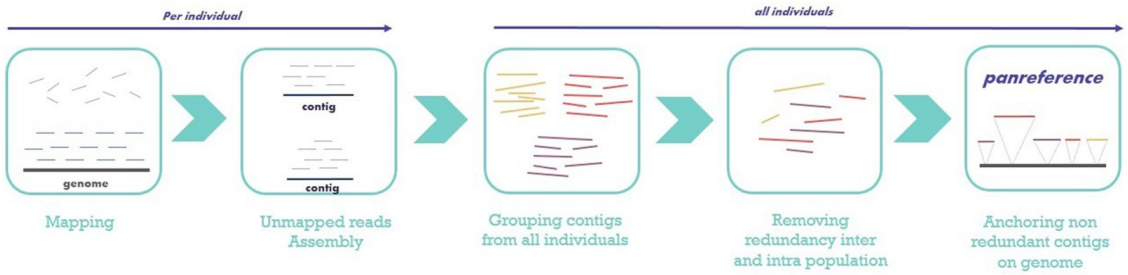

Figure 8.1: Summary of the approach 'Map-then-assemble' implemented in FrangiPANe.
Raw pair-ended short reads are mapped to the reference genome, separately for each sample, and unmapped reads are assembled. Next, contigs from all individuals are pooled and clustered to reduce redundancy. Non-redundant contigs are finally anchored on the genome. Figure from Tranchant-Dubreuil et al. 2023.

FrangiPANe simplifies the construction of a pangenome by providing both the entire process to build it from its own data, *i.e.* from mapping to the anchoring of contigs on the genome, and all the required bioinformatic softwares, within a single virtual machine (List of main tools in appendix, Table 14.2, page 149). In addition, an unique interface allows to perform each step of the analysis and progressively analyze the data displayed in different ways such as tables or plots, through a well-documented jupyter notebook. FrangiPANe is available on github, IRD dataverse or on the BioSphere Cloud of the French Institute of Bioinformatics (Appliance frangiPANe).

8.2 Building the first panreference for African Rice

To identify sequences absent from the *Oryza glaberrima* genome, we took advantage of short reads sequencing data from 248 African Rice accessions, with sequencing depths greater than 20 X, that were previously described in Cubry et al. 2018 and Monat, Pera, et al. 2016 (Table 8.1, page 88). These samples consisted of 164 domesticated and 84 wild relatives that were representative of the genetic diversity within the two African rice species *O. glaberrima* and *O. barthii* (Orjuela et al. 2014).

Species	#samples	#reads	#Gb	X	Ref.
O. glaberrima	162	20 999 547 976	2,074	35	Cubry et al., 2018
O. barthii	84	9 940 764 450	982	28	Cubry et al., 2018
O. glaberrima	1 (CG14)	207 601 236	20.76	60	Monat et al., 2016
O. glaberrima	1 (TOG5681)	289 581 328	24.910	72	Monat et al., 2016

Table 8.1: List of Illumina sequencing data used from 248 African rice accessions.
For each dataset, the species, the number of samples (#samples), the number of reads (#reads), the total number of gigabases (#Gb), the average depth sequencing (X) and the reference are provided. Table From Tranchant-Dubreuil et al. 2023.

In addition, we relied on two genomes, part of the 248 accessions sequenced using short reads:

- An improved reference genome of the cultivar CG14 from *Oryza glaberrima* generated by the OMAP consortium (GCA 000147395);
- A new whole genome assembly of the cultivar TOG5681 from *O. glaberrima* sequenced in long reads (ONT).

Main results of the African Rice panreference building

A precise synthesis of the main outcomes were presented hereafter but all the material and methods as well as all the results are fully detailed in part 9 (page 91) which contains the article published in *NAR Genomics and bioinformatics* (Tranchant-Dubreuil et al. 2023). Following the approach described in the previous section (Figure 8.1, page 88), the main results were:

- A high mean mapping rate of the resequencing data of the 248 samples on the CG14 genome : 96% and 97.8% for *O. barthii* and *O glaberrima*. This mapping rate decreased respectively to 93.7% and 96.2% considering only mapped in pairs (Figure 14.5 in appendix, page 150);
- After assembling the unmapped reads and filtering contigs according to several criteria such as minimal size, 8 Mb of sequences were obtained per individual, totaling 1.65 Gb and 1,306,706 sequences (Table 8.2, page 89);
- After reducing redundancy, **we identified 513.5 Mb (484,394 contigs)** with an average sequence size of 1,060 bp ranging from 301 bp up to 83,704 bp. (Figure 14.6 in appendix, page 150);
- **31.5% of the non-redundant contigs (152,411) were placed at a unique position** on the reference genome (145 Mb; Figure 8.2, page 90) whereas 8% (39,630) of the contigs were placed at multiple positions (31 Mb) and finally, 60.3% of the contigs remained unplaced.

Species	#raw ctgs	#raw ctgs sample	#filtered ctgs	#filtered ctgs sample	Tot. length (Mb)	Tot. length sample	seq size (bp)
O. barthii	5,424,759	64,580	763,176	9,085	740	10.6	1,192
0. glaberrima	4,427,624	27,210	543,500	3,334	917	5.5	1,355
TOTAL	9,887,127	39,867	1,306,676	5,290	1,657	8	

Table 8.2: Assembly summary.
This table provides statistics about the contigs (ctgs) assembled by abyss and the contigs keeped after filtering steps. The statistics include the contigs number (#raw ctgs, #filtered ctgs), the average number of contigs per sample (#raw ctgs per sample, #filtered ctgs per sample), the total length of sequence assembled, the average length of sequence assembled by sample and the average sequence size. Table from Tranchant-Dubreuil et al. 2023.

The cultivar TO581 sequenced both in long and short reads were used as support to valid the approach. 5,318 contigs (7.9 Mb) were assembled from TOG5681 short reads data and 97.7% of them were recovered on the corresponding long reads assembled genome. 1,696 contigs (31.9%) from TOG5681 were placed at a unique position on the CG14 genome and 95.1% of these placed contigs also aligned to the TOG5681 genome with a coverage of 100%. We also realigned on the CG14 genome the TOG5681 1kbp-flanking sequences surrounding the aligned contigs and 92.5% of them were found at the same position on the CG14 genome, thus validating the anchoring approach.

Panreference annotation

In total, **52.1% of the panreference was annotated as repetitive elements**, including retrotransposons (25.3%), DNA transposons (16.3%) and unclassified elements (10.5%) (Figure in appendices 14.9, page 153). The transposable elements content ratio was twice higher in contigs (67.6%) than in the genome reference (29.2%).

Transferring the Nipponbare genome annotation, 95.5% of the genes annotated on the Asian rice genome (36,159 genes) were successfully mapped on the panreference (Figure in appendices 14.8, page

152) with an average sequence identity in exons of mapped genes of 96% and an average alignment coverage of 98% (Figure 14.7 in appendix, page 14.7). In total, **we identified 3,252 new genes transferred from the Asian Rice genome that were absent from the African reference genome**.

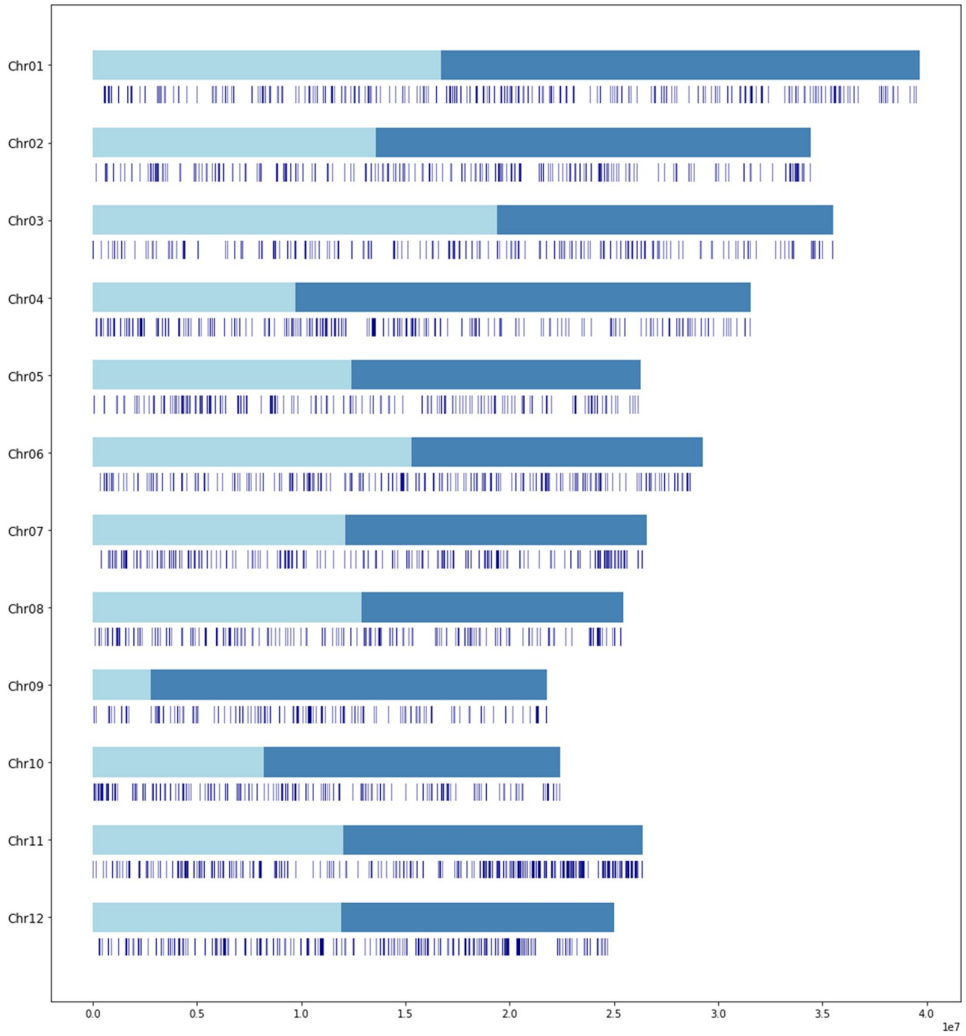

Figure 8.2: Contigs location on the 12 chromosomes of CG14.
A total of 152,411 sequences were uniquely anchored, representing 31.5% of the total number of contigs. Figure from Tranchant-Dubreuil et al. 2023.

9- Article "FrangiPANe, a tool for creating a pan-reference using left behind reads"

I have a bad feeling about this – Han Solo

Published online 20 February 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 1 **1**
https://doi.org/10.1093/nargab/lqad013

FrangiPANe, a tool for creating a panreference using left behind reads

Tranchant-Dubreuil Christine [1,*,†], Chenal Clothilde[1,2,3,†], Blaison Mathieu[1], Albar Laurence[4], Klein Valentin[1], Mariac Cédric[1], Wing Rod A.[5], Vigouroux Yves[1,*] and Sabot Francois [1,*]

[1]DIADE, Univ Montpellier, CIRAD, IRD, 911 Avenue Agropolis 34934, 34830 Montpellier Cedex 5, France, [2]MIVEGEC, Univ Montpellier, CNRS, IRD, 911 Avenue Agropolis 34934, 34830 Montpellier Cedex 5, France, [3]ISEM, Univ Montpellier, CNRS, IRD, EPHE, CIRAD, INRAP, 1093-1317 Route de Mende, 34090 Montpellier, France, [4]PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France and [5]Center for Desert Agriculture, Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

Received July 08, 2022; Revised December 21, 2022; Editorial Decision January 20, 2023; Accepted February 02, 2023

ABSTRACT

We present here FrangiPANe, a pipeline developed to build panreference using short reads through a map-then-assemble strategy. Applying it to 248 African rice genomes using an improved CG14 reference genome, we identified an average of 8 Mb of new sequences and 5290 new contigs per individual. In total, 1.4 G of new sequences, consisting of 1 306 676 contigs, were assembled. We validated 97.7% of the contigs of the TOG5681 cultivar individual assembly from short reads on a newly long reads genome assembly of the same TOG5681 cultivar. FrangiPANe also allowed the anchoring of 31.5% of the new contigs within the CG14 reference genome, with a 92.5% accuracy at 2 kb span. We annotated in addition 3252 new genes absent from the reference. FrangiPANe was developed as a modular and interactive application to simplify the construction of a panreference using the map-then-assemble approach. It is available as a Docker image containing (i) a Jupyter notebook centralizing codes, documentation and interactive visualization of results, (ii) python scripts and (iii) all the software and libraries requested for each step of the analysis. We foreseen our approach will help leverage large-scale illumina dataset for pangenome studies in GWAS or detection of selection.

INTRODUCTION

Nowadays, an increasing number of studies highlights the limit of using a single individual genome to assess genomic diversity within a species (1–4). For instance, in plants, between 8% and 27% of genes varied in presence/absence across individuals from the same species (5–7). Pangenomics offers an alternative way to study gene content variations and more broadly the whole genomic variations within a population. Initially introduced in bacteria by Tettelin *et al.* (8), the pangenome concept refers to the replete genomic content of a species, consisting in (i) the core genome, shared among all individuals, and (ii) the dispensable genome, shared only in a subset of individuals. With the decrease in sequencing costs, pangenomics analyses are more and more frequent in plants (9–12) and animals (13–16).

The pivotal step in any pangenomic analysis is the construction of a panreference that captures the (almost) full diversity of a large set of genomes. However, the pangenome construction remains a cumbersome and challenging process, especially for Eukaryotes due to their large genome size and complexity (e.g. repeat content or polyploidy). Although long reads sequencing technologies are increasingly used to directly detect large structural variations, generally through reassembly of genomes (12,17–19), short reads ones currently remain less expensive and is still widely used. In addition, the numerous short reads datasets already available on many organisms provide an important source of data to perform large-scale pangenomic analyses. Two approaches were mainly used for the pangenome construction from individuals sequencing with such short reads:

*To whom correspondence should be addressed. Tel: +33 467416334; Fax: +33 467416222; Email: christine.tranchant@ird.fr
Correspondence may also be addressed to Sabot Francois. Email: francois.sabot@ird.fr
Correspondence may also be addressed to Vigouroux Yves. Email: yves.vigouroux@ird.fr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Table 1. List of Illumina sequencing data used from 248 African rice accessions. For each dataset, the species, the number of samples (#samples), the number of reads (#reads), the total number of gigabases (#Gb), the average depth sequencing (X) and the reference are provided

Species	#samples	#reads	#Gb	X	Ref
O. glaberrima	162	20 999 547 976	2,074	35	Cubry *et al.*, 2018
O. barthii	84	9 940 764 450	982	28	Cubry *et al.*, 2018
O. glaberrima	1 (CG14)	207 601 236	20.76	60	Monat *et al.*, 2016
O. glaberrima	1 (TOG5681)	289 581 328	24.910	72	Monat *et al.*, 2016

(i) *de novo* genome assembly followed by genome comparison (here referred as the 'assemble-then-map' approach; 10,20), and (ii) the 'map-then-assemble' approach, based on the mapping of resequencing short reads followed by the *de novo* assembly of the unmapped reads (6,13,21).

Very few tools are publicly available to perform all the steps at once, being either developed for bacteria (22–24), or based on the *de novo* 'assemble-then-map' approach (25).

We present here a method based on the 'map-then-assemble' approach to (i) identify large fragments absent from a reference genome using short reads data, (ii) locate these variations on the reference, and (iii) build a pan-reference. For that purpose, we developed frangiPANe, a pipeline tool to easily apply this approach and to create an accurate panreference for any organism using short reads data.

To validate our method and frangiPANe, we used the resequencing data from 248 genomes (26,27) of the cultivated African rice (*Oryza glaberrima*) and of its closest wild relative (*O. barthii*) as a proof-of-concept to build the first pan-reference for African rice.

MATERIALS AND METHODS

Sample sequencing

Short-read sequencing data. We used whole genome sequencing data (Table 1) from 248 African rice accessions previously described in Cubry *et al.* (26) and Monat *et al.* (27) (Illumina technology TrueSeqv3, 100–150 bp paired-end reads), including 164 domesticated and 84 wild relative individuals. These samples covered the full range of genetic diversity in the two African rice species *O. glaberrima* and *O. barthii* (28).

TOG5681 long reads sequencing and assembly. DNA was extracted following the protocol from Serret *et al.* (29) using an adapted CTAB-lysis approach to ensure high molecular weight DNA. DNA quality and concentration were controlled using PFGE and Qubit, and subject to a LSK-109 library as recommended by suppliers (Oxford Nanopore Technology, Inc, Oxford, United Kingdom). The library was loaded on two 9.4.1 flowcells, raw FAST5 base-called using Guppy 4.0.5 (hac model) with a cut-off at PhredQ 7, and FASTQ data were controlled using NanoPlot 1.38.1 (30). FASTQ were then assembled using Flye 2.8 (31) with the –nano-raw mode and standard options. Initial polishing was ensured by three turns of Racon 1.3 (32) under standard conditions using the initial set of nanopore reads and mapping performed by Minimap2 v2.10 (33) in -x map-ont mode. Final polishing was performed using Medaka 1.2 (https://Github.com/Nanoporetech/Medaka) with the standard model. Contamination was checked using Blobtools 1.1 (34) and remapping of short reads, as recommended, in the same way as described below. Final chromosome-scale scaffolding was done using RagTag 2.1 (35) using the CG14 OMAPv2 as reference sequence. BUSCO v5.0 (36) with *Viridiplantae* database v10 was used for computing the gene space completion, and all basic statistics on contigs and scaffolds were obtained using QUAST 5.0 (37).

'Map-then-assemble' approach

This approach starts with mapping of resequencing pair-ended short reads on a reference genome, followed by *de novo* assembly of unmapped reads for each sample. Next, all contigs are pooled after filtering. The last two steps consist in reducing the sequence redundancy at intra- and inter-population level and into placing the non-redundant sequences on the reference genome. Figure 1 provides an overview of the 'map-then-assemble' approach.

Alignment to the reference genome. The CG14 reference genome was downloaded from the European Nucleotide Archive (ENA) (Accession GCA_000147395, https://www.ebi.ac.uk/ena/browser/view/GCA_000147395). For each accession, pair-ended short reads were aligned to the CG14 reference with bwa aln (option -$n = 5$) and bwa sampe (version 0.7.15) (38). Mapping results were sorted with Picard-tools sortSam (version 2.6.0). Samtools view (version 1.3.1) (39) was used to extract unaligned reads (option -F 2).

Unmapped reads assembly and filtering. For each sample, unmapped pair-ended short reads were assembled with Abyss-pe (version 2.0.2) (40). We first optimized k-mer size for assembly in the AA accession (Supplementary notes and Supplementary Table S1), and chose a k-mer size (option -k 64) maximizing N50 and minimizing both contigs number and L50. Contigs shorter than 300 bp were excluded.

We screened contigs for vector sequences using Vecscreen (https://www.ncbi.nlm.nih.gov/Tools/Vecscreen/about/) and the NCBI UniVec_core database (V.build 10.0, http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Contigs were aligned with blastn against the NCBI NT nucleotide database (27 October 2019) and the rice organites genomes (mitochondrial and chloroplast). Contigs with best hits from outside the green plants taxon (Viridiplantae) or on rice organites genomes were removed.

Reducing redundancy. Contigs from all individuals were clustered using CD-HIT (version 4.6, options -c 0.80 -s 0.95) (41). Only the longest sequence for each cluster was conserved (Supplementary Notes and Supplementary Table S2).

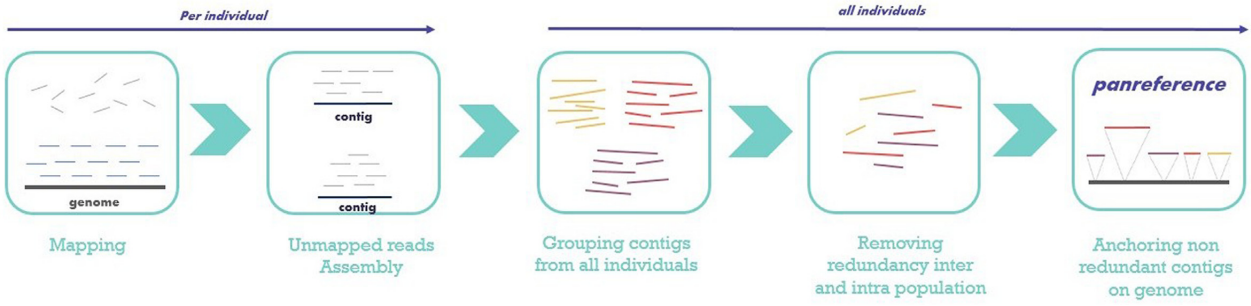

Figure 1. Summary of the approach 'Map-then-assemble' implemented in FrangiPANe. Raw pair-ended short reads are mapped to the reference genome, separately for each sample, and unmapped reads are assembled. Next, contigs from all individuals are pooled and clustered to reduce redundancy. Non-redundant contigs are finally anchored on the genome.

Contigs position on the genome reference. Pair-ended shorts reads from each individual were remapped to a new cumulative reference formed of the CG14 assembly and of the non-redundant contigs. Mapping to this panreference was performed using bwa aln and sampe with the same parameters as described before. Pair-ended short reads mapping on both a contig and a chromosome were used to anchor contigs (9; 13). Briefly, all reads aligned within the first or last 300 bases of a contig and for which mates mapped on a CG14 chromosome were pre-selected. Contigs position within chromosomes was considered as valid if: (i) at least 10 reads with MAPQ 10 are aligned on the same contig and their mates on the same chromosome, and (ii) the positions of the 10 mate reads on the chromosome are all located in a span shorter than 2kb.

Assembly validation and position validation on chromosomes. We used the TOG5681 genome assembly based on long reads sequencing to validate (i) our contigs assembly from TOG5681 pair-ended short reads and (ii) their position on chromosomes. We used the nucmer tool (MUMmer version 4.0beta3) (42) and kept only alignment showing 90% identity and 80% coverage of the contig, with a minimum aligned sequence length of 300 bp.

Panreference annotation

Transposable element identification. Transposable elements were detected using RepeatMasker (version 4.0.7) (43) with the RiTE-db (version 1.1) (44) and the RepBase (version 23.11, *Oryza* section) (45) databases.

Genes mapping. Annotation of the panreference was performed using Liftoff (version 1.6.1) (46) with annotated genes from the Nipponbare reference (*Oryza sativa* ssp. *japonica* cv. Nipponbare, IRGSP-1.0–1-2021–11-11 release). Genes were considered as successfully mapped if a minimum of 50% of the Nipponbare gene was aligned to the panreference with a sequence identity higher than 50% (options -s 0.5 -a 0.5). Gene copies were annotated using a minimum of sequence identity threshold of 95% (options -copies -sc 0.95).

Gene ontology annotation. Genes sequences were aligned to the NCBI NR protein database (9 September 2021, *Viridiplantae* section) using blastx (options -e-value 1e-6). Genes with protein domain signatures were recovered using InterProScan (version 5.53.87; options –goterms –iprlookup –pathway) (47). GO annotation and enrichment analysis were carried out through the Blast2GO (version 0.3, with default options, Fisher' exact test with a cutoff of *P*-value 0.05) (48).

A tool to build panreference from scratch

FrangiPANe was developed as a modular and interactive application to simplify the construction of a panreference using the map-then-assemble approach (Figure 1). It is available as a Docker image containing (i) a Jupyter notebook centralizing codes, documentation and interactive visualization of results, (ii) python scripts and (iii) all the softwares and libraries requested for each step of the analysis. Supplementary Table S3 presents the main list of tools required by FrangiPANe.

The code, documentation, installation manual and test data are available under the GPLv3 and CC4.0 BY-NC license at https://github.com/tranchant/frangiPANe. A dedicated virtual machine is also available on the BioSphere Cloud of the French Institute of Bioinformatics (Appliance frangiPANe, https://biosphere.france-bioinformatique.fr/catalogue/appliance/201/).

RESULTS

The CG14 and TOG5681 genomes

We relied on an improved reference genome of the cultivar CG14 from *Oryza glaberrima* from the OMAP consortium (Accession GCA_000147395, https://www.ebi.ac.uk/ena/browser/view/GCA_000147395) and on a new whole genome assembly of the cultivar TOG5681 (see below for details), both accessions being themselves part of the 248 accessions sequenced using short reads (27).

TOG5681 control genome. We obtained 509 485 ONT long reads of minimal PhredQ 7, for 6.612 Gb of data (18x) with a N50 of 23.8 kb. After assembly and polishing, the final dataset represents 148 contigs, for a total assembly size of 348 131 590 bases, a N50 of 15 386 152 bases (L50 of 9), and 99.5% of the assembly being comprised in contigs larger than 50 kb. The BUSCO score for this assembly is 95%,

Table 2. Assembly summary. This table provides statistics about the contigs (ctgs) assembled by abyss and the contigs keeped after filtering steps. The statistics include the contigs number (#raw ctgs, #filtered ctgs), the average number of contigs per sample (#raw ctgs per sample, #filtered ctgs per sample), the total length of sequence assembled, the average length of sequence assembled by sample and the average sequence size

Species	#raw ctgs	#raw ctgs per sample	#filtered ctgs	#filtered ctgs per sample	Total length (Mb)	Total length per sample	seq size (bp)
O. barthii	5 424 759	64 580	763 176	9085	740	10.6	1192
O. glaberrima	4 427 624	27 210	543 500	3334	917	5.5	1355
	9 887 127	39 867	1 306 676	5290	1657	8	

including 2.1% of duplicated target genes. Blobtools indicated only three contamination contigs, representing less than 0.01% of the total size. After removal of these contaminated contigs, RagTag was used on the remaining 145 contigs to scaffold the TOG5681 genome using the CG14 one as reference, with 99.4% of the bases placed, leading to a final chromosome scale assembly of 59 contigs (12 chromosomes + 47 unplaced contigs) representing 348 140 190 bases.

Building african rice panreference

To identify sequences absent from the CG14 genome, we used short reads sequencing data of 164 domesticated and 84 wild relatives, all of which exceeding a sequencing depth of 20× (Table 1). The mean mapping rate of these 248 genomes was high, with 96% and 97.8% for *O. barthii* and *O. glaberrima*, respectively. The mapping rate decreased respectively to 93.7% and 96.2% considering only reads correctly mapped in pairs (Supplementary Figure S1).

Unmapped reads assembly produced a total of 2.9 Gb of sequences and 9 887 127 contigs. After filtering for adapter (<1% of sequences), alien sequences (0.01%) and minimal size (86.7%), we ended up with 1.65 Gb and 1 306 676 contigs. On average, 8 Mb of sequences and 5290 contigs were obtained per individual (from 1.4 to 25.2 Mb assembled per individual and a contigs number ranging from 1008 and 49 949 per individual, Table 2). The exception was CG14, for which we assembled a few 633 contigs, each with a very small size (303 bp on average).

After reducing redundancy, we identified 513.5 Mb (484 394 contigs) with an average sequence size of 1060 bp (ranging from 301 bp up to 83 704 bp, Supplementary Figure S2). 56.4% of these non-redundant sequences were identified as singleton (Supplementary Table S4, Supplementary Figure S2).

Contigs anchoring on the reference genome. We remapped all pair-ended short reads on the panreference consisting of the cumulation of the CG14 genome and of the newly deduplicated assembled sequences (484 394 contigs). We increased the mapping rate by 0.9 and 2.3% for the domesticated and the wild relative accessions, respectively (98.7% and 98.3%).

Using the pair-end mapping information, we accurately placed 31.5% of the non-redundant contigs (152 411 contigs) at a unique position on the reference genome (145 Mb; Figure 2). A total of 39 630 contigs (8.2%) were placed at multiple positions, on the same chromosome or not (31 Mb). Finally, 292 353 contigs (60.3%) remained unplaced (representing a total of 337 Mb).

The assembled contigs from TOG5681 short reads data (7.9 Mb, 5318 contigs) were recovered at 97.7% on the corresponding long reads assembled genome. A total of 1696 contigs (31.9%) from this accession were placed with a high confidence at a unique position on the CG14 genome. 95.1% of these 1696 contigs also mapped against the TOG5681 genome with a coverage of 100%. We realigned on the CG14 genome the TOG5681 1kbp-flanking sequences surrounding the aligned contigs and 92.5% of them were found at the same position on the CG14 genome, thus validating the anchoring approach (Supplementary Figure S3).

Panreference annotation

In total, 52.1% of the panreference was annotated as repetitive elements, including retrotransposons (25.3%), DNA transposons (16.3%) and unclassified elements (10.5%) (Supplementary Table S5). The transposable elements (TEs) content ratio was twice higher in contigs (67.6%) than in the genome reference (29.2%). We also observed a higher percentage of DNA transposons within the contigs (34%) than in the reference genome (26.5%), especially regarding the ones being anchored on the genome (42.5%) (Supplementary Table S5 and Supplementary Figure S4). We also observed a higher divergence of TEs in the contigs than in the reference genome (Supplementary Figure S5).

Out of 37 864 genes annotated from the Nipponbare genome, 95.5% of genes (36 159 genes) were successfully mapped on the panreference, including 35 252 genes on chromosomes and 907 on all non redundant contigs. The average sequence identity in exons of mapped genes was 96% and the average alignment coverage was 98% (Supplementary Table S6 and Supplementary Figure S6).

98.7% of these genes are placed on the same chromosome on the Nipponbare and the CG14 genomes respecting the co-linearity of gene order between genomes (Supplementary Figure S7).

In addition to the successfully mapped genes, we found 2631 additional gene copies, 281 and 2345 on the CG14 genome and on contigs sequences, respectively.

Genes present in the contigs were enriched in GO related to detoxification, response to chemical and response to toxic substance (Supplementary Table S7).

DISCUSSION

Understanding plant genomic diversity requires reliable tools to rapidly build up pangenome sequences. We present here a framework to develop such an approach and apply it on 248 African rice genomes.

Overall, the results of FrangiPANe on African rice are in agreement with the ones from its cousin species the Asian

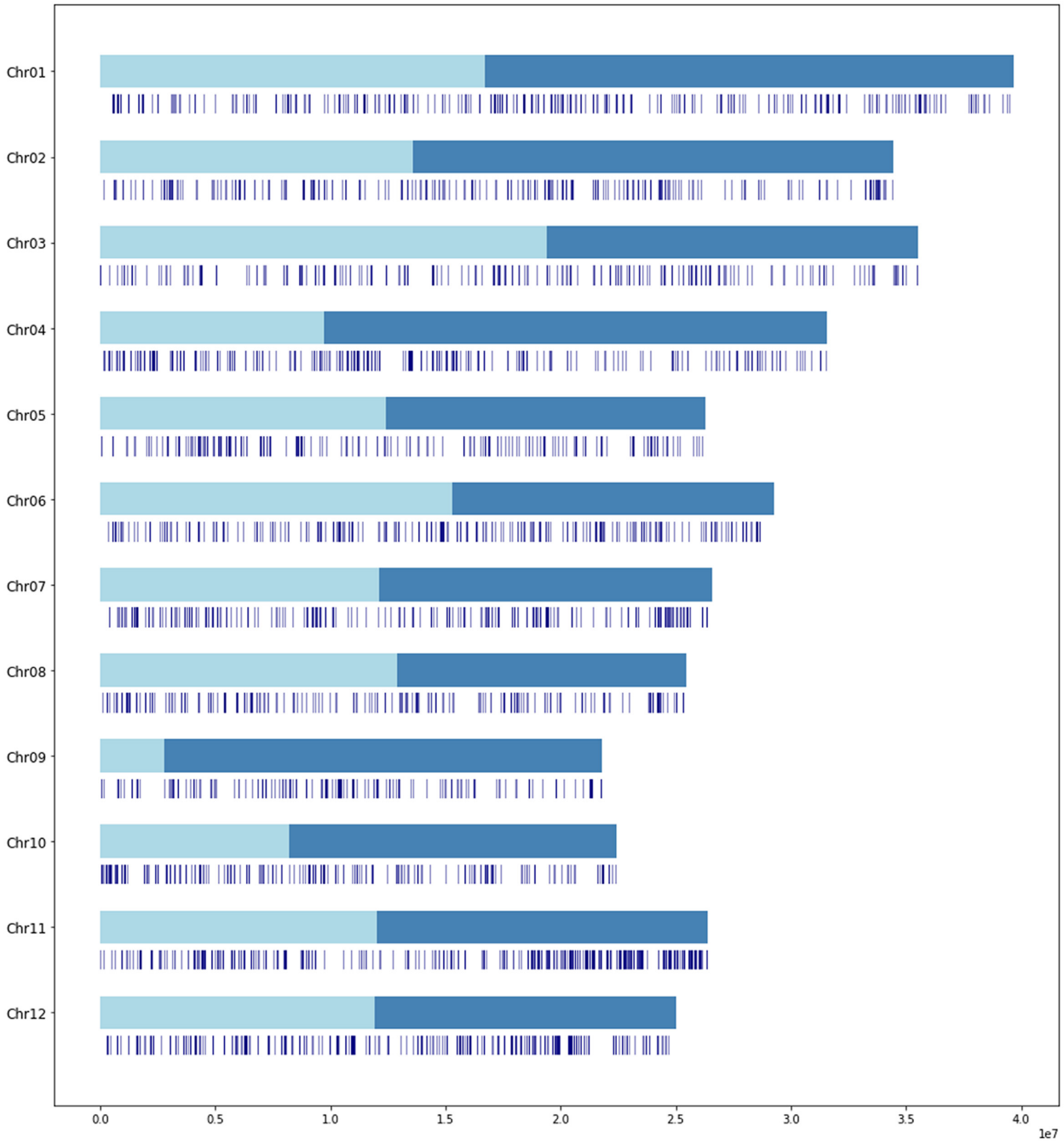

Figure 2. Contigs location on the 12 chromosomes of CG14. A total of 152 411 sequences were uniquely anchored, representing 31.5% of the total number of contigs.

rice (2, 12). In total, we identified 513 Mb of new sequences, in addition to the 344 Mb reference genome. The new sequenced part is in accordance with the Asian rice one, which ranges from 268 Mb (2) to 1.3 Gbp (12). The lowest value (268 Mb) is based on short reads from 3010 Asian rice genomes (2), an approach similar to ours. We found twice as many new sequences for a smaller set of individuals, but we also included wild rice species. Generally, newly assembled dispensable genomic sequences are generally enriched in TEs. For instance in Asian rice, 52.7% of the newly assembled sequence were TEs (12), compared to an expected number of 35% in the reference genome (49). Our re-assembled sequence using short reads data was composed of 53% of TEs in African rice, almost identical to the one estimated with a long reads approach on Asian rice (52.7%; (12)). In terms of gene number, using 66 Asian rice, Zhao *et al.* (2018) found 10 872 new genes (50), so, roughly in average, 165 genes per individual. Using the 3010 Asian rice genomes

(2), a total of 12 465 novel full-length genes were detected, representing an average of 4 per genome. Here, we found 13 genes per genome (3252 genes in total), three times more than the 3010 genome study and 10 times less than in the 66 ones. The large disparity between these estimations might lay in the stringency of gene calling and in the procedure of annotation. In our case, we certainly underestimate the number of genes, as we only used a transfer of annotation. *De novo* annotation should thus allow identification of additional new genes specific to the African rice.

Our tool presents several improvements compared to other available tools. These were developed primarily for bacterial species (24, 51, 52) using short reads sequencing data (53,54) such as PanSeq (22), PGAP (55), roary (23) or PanX (24). They are mostly gene-oriented tools, however, and were specifically designed and tested on the small and simple bacterial genomes. In Eukaryotes, HUPAN (25), a command line tool, has been developed and applied to rice and human (25, 20). This tool starts with *de novo* genome assembly of each individual, followed by mapping of contigs upon the reference genome, and finally clustering of all unaligned contigs ('assemble-then-map' approach). However, such assembly based on short reads lead to missing regions and repeat compression.

We proposed here FrangiPANe as a new solution relying on a massively parallelizable approach, based on the 'map-then-assemble' pathway. Our tool proved to be particularly accurate with 97.7% of assembled contigs from the TOG5681 accession also present in a new long reads genome assembly of the same TOG5681 individual.

FrangiPANe also provides a complete environment for panreference creation through an unique and interactive interface, without requiring huge programming skills or the installation of numerous bioinformatic softwares. Based on Docker (https://docs.docker.com/get-docker) and Jupyter (https://jupyter.org/), it streamlines the whole process involving multiple analysis steps and the data visualization in different way (e.g. tables or plot) within a single well-documented notebook.

While long reads *de novo* genome assembly offers new opportunities to perform pangenome analysis (11,18,19), the vast majority of currently available datasets are from short reads Illumina technology, and are generally very large in terms of number of individuals. FrangiPANe offers opportunities to take advantage of these datasets to gain a better understanding of plant and animal genomic diversity, and also to carry out large-scale pangenomic studies to detect selection or perform association with phenotype (GWAS).

DATA AVAILABILITY

frangiPANe is freely available in the GitHub repository https://github.com/tranchant/frangiPANe, under the double licence CeCiLL-C (http://www.cecill.info/licences/Licence_CeCILL-C_V1-en.html) and GNU GPLv3.

A virtual machine is also available at the BioSphere service of the French Institute of Bioinformatics (Appliance frangiPANe, https://biosphere.france-bioinformatique.fr/catalogue/appliance/201/). The sequences (fasta file) and their placement on the reference genome (csv file) have been deposited in the IRD dataverse: Tranchant, Christine; Chenal, Clothilde; Blaison, Mathieu; Albar, Laurence; Klein, Valentin; Mariac, Cédric; Wing, Rod; Vigouroux, Yves; Sabot, Francois, 2022, 'Supporting data for the African Rice Panreference produced by the frangiPANe software', DataSuds, V1, https://doi.org/10.23708/93OQMD.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors acknowledge Ndomassi Tando and the ISO 9001 certified IRD itrop HPC (member of the South Green Platform) at IRD Montpellier as well as the TGCC platform for providing HPC resources that have contributed to the research results reported within this paper (URL: https://bioinfo.ird.fr/ - http://www.southgreen.fr). They also thank Christophe Blanchet and the French Institute of bioinformatics (IFB) to provide access to the appliance frangiPANe through the Biosphere cloud (https://biosphere.france-bioinformatique.fr/cloud/).

FUNDING

France Genomique French National infrastructure and funded as part of "Investissement d'avenir" [ANR-10-INBS-09]; IRIGIN project.
Conflict of interest statement. None declared.

REFERENCES

1. Springer,N.M., Ying,K., Fu,Y., Ji,T., Yeh,C.-T., Jia,Y., Wu,W., Richmond,T., Kitzman,J., Rosenbaum,H. *et al.* (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.*, **5**, e1000734.
2. Wang,W., Mauleon,R., Hu,Z., Chebotarov,D., Tai,S., Wu,Z., Li,M., Zheng,T., Fuentes,R.R., Zhang,F. *et al.* (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.
3. Tranchant-Dubreuil,C., Rouard,M. and Sabot,F. (2019) Plant Pangenome: impacts on Phenotypes and Evolution. In: *Annual Plant Reviews online*. Vol. **2**, pp. 453–478.
4. 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.
5. Schatz,M.C., Maron,L.G., Stein,J.C., Wences,A., Gurtowski,J., Biggers,E., Lee,H., Kramer,M., Antoniou,E., Ghiban,E. *et al.* (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.*, **15**, 506.
6. Gao,L., Gonda,I., Sun,H., Ma,Q., Bao,K., Tieman,D.M., Burzynski-Chang,E.A., Fish,T.L., Stromberg,K.A., Sacks,G.L. *et al.* (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.*, **51**, 1044–1051.
7. Hübner,S., Bercovich,N., Todesco,M., Mandel,J.R., Odenheimer,J., Ziegler,E., Lee,J.S., Baute,G.J., Owens,G.L., Grassa,C.J. *et al.* (2019) Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants*, **5**, 54–62.
8. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13950–13955.
9. Golicz,A.A., Bayer,P.E., Barker,G.C., Edger,P.P., Kim,H., Martinez,P.A., Chan,C.K.K., Severn-Ellis,A., McCombie,W.R., Parkin,I.A.P. *et al.* (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.*, **7**, 13390.

10. Gordon,S.P., Contreras-Moreira,B., Woods,D.P., Des Marais,D.L., Burgess,D., Shu,S., Stritt,C., Roulin,A.C., Schackwitz,W., Tyler,L. *et al.* (2017) Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.*, **8**, 2184.
11. Liu,Y., Du,H., Li,P., Shen,Y., Peng,H., Liu,S., Zhou,G.-A., Zhang,H., Liu,Z., Shi,M. *et al.* (2020) Pan-genome of wild and cultivated soybeans. *Cell*, **182**, 162–176.
12. Qin,P., Lu,H., Du,H., Wang,H., Chen,W., Chen,Z., He,Q., Ou,S., Zhang,H., Li,X. *et al.* (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, 3542–3558.
13. Sherman,R.M., Forman,J., Antonescu,V., Puiu,D., Daya,M., Rafaels,N., Boorgula,M.P., Chavan,S., Vergara,C., Ortega,V.E. *et al.* (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.*, **51**, 30–35.
14. Gerdol,M., Moreira,R., Cruz,F., Gómez-Garrido,J., Vlasova,A., Rosani,U., Venier,P., Naranjo-Ortiz,M.A., Murgarella,M., Greco,S. *et al.* (2020) Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biol.*, **21**, 275.
15. Li,R., Fu,W., Su,R., Tian,X., Du,D., Zhao,Y., Zheng,Z., Chen,Q., Gao,S., Cai,Y. *et al.* (2019) Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. *Front. Genet.*, **10**, 1169.
16. Tian,X., Li,R., Fu,W., Li,Y., Wang,X., Li,M., Du,D., Tang,Q., Cai,Y., Long,Y. *et al.* (2020) Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Sci. China. Life Sci.*, **63**, 750–763.
17. Song,J.M., Guan,Z., Hu,J., Guo,C., Yang,Z., Wang,S., Liu,D., Wang,B., Lu,S., Zhou,R. *et al.* (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat. Plants*, **6**, 34–45.
18. Jayakodi,M., Padmarasu,S., Haberer,G., Bonthala,V.S., Gundlach,H., Monat,C., Lux,T., Kamal,N., Lang,D., Himmelbach,A. *et al.* (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, **588**, 284–289.
19. Walkowiak,S., Gao,L., Monat,C., Haberer,G., Kassa,M.T., Brinton,J., Ramirez-Gonzalez,R.H., Kolodziej,M.C., Delorean,E., Thambugala,D. *et al.* (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature*, **588**, 277–283.
20. Hu,Z., Sun,C., Lu,K.C., Chu,X., Zhao,Y., Lu,J., Shi,J. and Wei,C. (2017) EUPAN enables pan-genome studies of a large number of eukaryotic genomes. *Bioinformatics*, **33**, 2408–2409.
21. Hufnagel,B., Soriano,A., Taylor,J., Divol,F., Kroc,M., Sanders,H., Yeheyis,L., Nelson,M. and Péret,B. (2021) Pangenome of white lupin provides insights into the diversity of the species. *Plant Biotechnol. J.*, **19**, 2532–2543.
22. Laing,C., Buchanan,C., Taboada,E.N., Zhang,Y., Kropinski,A., Villegas,A., Thomas,J.E. and Gannon,V.P.J. (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinf.*, **11**, 461.
23. Page,A.J., Cummins,C.A., Hunt,M., Wong,V.K., Reuter,S., Holden,M.T.G., Fookes,M., Falush,D., Keane,J.A. and Parkhill,J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691.
24. Ding,W., Baumdicker,F. and Neher,R.A. (2018) panX: pan-genome analysis and exploration. *Nucleic. Acids. Res.*, **46**, e5.
25. Duan,Z., Qiao,Y., Lu,J., Lu,H., Zhang,W., Yan,F., Sun,C., Hu,Z., Zhang,Z., Li,G. *et al.* (2019) HUPAN: a pan-genome analysis pipeline for human genomes. *Genome Biol.*, **20**, 149.
26. Cubry,P., Tranchant-Dubreuil,C., Thuillet,A.-C., Monat,C., Ndjiondjop,M.-N., Labadie,K., Cruaud,C., Engelen,S., Scarcelli,N., Rhoné,B. *et al.* (2018) The rise and fall of African rice cultivation revealed by analysis of 246 new genomes. *Curr. Biol.*, **28**, 2274–2282.
27. Monat,C., Pera,B., Ndjiondjop,M.-N., Sow,M., Tranchant-Dubreuil,C., Bastianelli,L., Ghesquière,A. and Sabot,F. (2016) De novo assemblies of three *Oryza glaberrima* accessions provide first insights about pan-genome of African rices. *Genome Biol. Evol.*, **9**, 1–6.
28. Orjuela,J., Sabot,F., Chéron,S., Vigouroux,Y., Adam,H., Chrestin,H., Sanni,K., Lorieux,M. and Ghesquière,A. (2014) An extensive analysis of the African rice genetic diversity through a global genotyping. *Theor. Appl. Genet.*, **127**, 2211–2223.
29. Serret,J., Couderc,M., Mariac,C., Albar,L. and Sabot,F. (2021) From low cost plant HMW DNA extraction to MinION sequencing. *protocols.io.*, dx.doi.org/10.17504/protocols.io.bu3vnyn6.
30. De Coster,W., D'Hert,S., Schultz,D.T., Cruts,M. and Van Broeckhoven,C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**, 2666–2669.
31. 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.
32. 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.
33. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
34. Laetsch,D.R. and Blaxter,M.L. (2017) BlobTools: interrogation of genome assemblies. *F1000Research*, **6**, 1287.
35. Alonge,M., Soyk,S., Ramakrishnan,S., Wang,X., Goodwin,S., Sedlazeck,F.J., Lippman,Z.B. and Schatz,M.C. (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.*, **20**, 224.
36. Manni,M., Berkeley,M.R., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.
37. Mikheenko,A., Prjibelski,A., Saveliev,V., Antipov,D. and Gurevich,A. (2018) Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, **34**, i142–i150.
38. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
39. 1000 Genome Project Data Processing Subgroup, Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R.(2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
40. Simpson,J.T., Wong,K., Jackman,S.D., Schein,J.E., Jones,S.J.M. and Birol,I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–1123.
41. 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.
42. 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.
43. Smit,A.F.A., Hubley,R. and Green,P. (1999) RepeatMasker. 1999. http://repeatmasker.org, (23 December 2022, date last accessed).
44. Copetti,D., Zhang,J., El Baidouri,M., Gao,D., Wang,J., Barghini,E., Cossu,R.M., Angelova,A., Maldonado,L.C.E., Roffler,S. *et al.* (2015) RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics*, **16**, 538.
45. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.
46. Shumate,A. and Salzberg,S.L. (2021) Liftoff: accurate mapping of gene annotations. *Bioinformatics*, **37**, 1639–1643.
47. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
48. Götz,S., García-Gómez,J.M., Terol,J., Williams,T.D., Nagaraj,S.H., Nueda,M.J., Robles,M., Talón,M., Dopazo,J. and Conesa,A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.*, **36**, 3420–3435.
49. Matsumoto,T., Wu,J., Kanamori,H., Katayose,Y., Fujisawa,M., Namiki,N., Mizuno,H., Yamamoto,K., Antonio,B.A., Baba,T. *et al.* (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
50. Zhao,Q., Feng,Q., Lu,H., Li,Y., Wang,A., Tian,Q., Zhan,Q., Lu,Y., Zhang,L., Huang,T. *et al.* (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.*, **50**, 278–284.
51. Freschi,L., Vincent,A.T., Jeukens,J., Emond-Rheault,J.G., Kukavica-Ibrulj,I., Dupont,M.J., Charette,S.J., Boyle,B. and Levesque,R.C. (2019) The *Pseudomonas aeruginosa* Pan-genome

provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biol. Evol.*, **11**, 109–120.
52. Davies,M.R., McIntyre,L., Mutreja,A., Lacey,J.A., Lees,J.A., Towers,R.J., Duchêne,S., Smeesters,P.R., Frost,H.R., Price,D.J. *et al.* (2019) Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nat. Genet.*, **51**, 1035–1043.
53. Vernikos,G.S. (2020) A review of pangenome tools and recent studies. In: Tettelin,H. and Medini,D. (eds). *The Pangenome*. Springer, Cham, pp. 89–112.
54. Bonnici,V., Maresi,E. and Giugno,R. (2021) Challenges in gene-oriented approaches for pangenome content discovery. *Brief. Bioinform.*, **22**, bbaa198.
55. Zhao,Y., Wu,J., Yang,J., Sun,S., Xiao,J. and Yu,J. (2012) PGAP: pan-genomes analysis pipeline. *Bioinformatics*, **28**, 416–418.

10- Discussion and Conclusion

And so! You and I are going to fly... a little break in the suspense... THE MOON !!! – Gru, Despictable Me

Exploring population diversity through a pangenomic approach requires, in the first instance, robust tools to construct pangenomes. We have developed an approach and one of the first all-in-one tools that simplify the complex task of building a eukaryotic pangenome from multiple, individual short-read-based genome sequencing. Taking advantage of 248 rice genomes resequenced, we have validated our approach on the African rice. We identified a total of 515 Mb of new sequences, this part of novel sequences being within the range of values found in the Asian rice, varying between 268 Mb (Wang, Mauleon, et al. 2018) and 1.3 Gbp (Qin et al. 2021). If we compared our result with the lowest value based on a similar approach carried out on 3,010 Asian rice genomes, we found twice as many sequences but in our study, wild species were integrated. As observed in other pangenomic studies (Golicz, Bayer, et al. 2016; Gordon et al. 2017), these sequences are enriched in transposable elements, with a total TEs rate of 52.1% in the panreference including the genome and the novel sequences. That result is similar to the 52.7% observed with a long reads approach on Asian Rice (Qin et al. 2021). It would be interesting to annotate TEs accurately in families and both to detect complete copies and to look at their insertion site and at nearby.

A "basic" and minimalist annotation of the panreference was performed by a simple transfer of the annotation from the Asian reference genome, *Oryza sativa japonica* and 95.5% of genes were transferred, including 3,252 new genes. This number was underestimated, as it did not contain, for example, genes specific to both cultivated and wild African rice. Annotation is one of the critical steps in the construction of a pangenome, as with a reference genome. Indeed, the majority of subsequent analyses will be based on this annotation, from which the genes will then be classified as core and dispensable, or structural variations will be associated with genes or all gene enrichment analyses will be derived, for example. A *de novo* annotation of the 515 Mb of novel sequences was then carried out and was integrated into the third part of this project which focused on the impact of domestication on the pangenome architecture and dynamics.

FrangiPANe offers interesting functionalities compared to other publicly available tools. Most of them are based on a gene-oriented approach and specific to bacterial genomes. HUPAN is a command line tool used to build human and rice pangenomes (Duan et al. 2019) and is based on the 'assemble-then-map' approach from short reads, the latter leading to missing region and repeat compression. Available as a virtual machine containing the complete environment, FrangiPANe allows to progressively create its own pangenome through an interactive and a single user-friendly interface, step by step with visualization of the results in different ways such as tables or plots. The main objective was to provide a turnkey protocol allowing standardisation and traceability of the pangenomic analysis, with the aim of being able to repeat the same procedure when integrating new individuals or creating pangenomes of new species. It would be interesting to integrate new functionalities and post analysis steps such as the characterisation of the core and dispensable pangenome or the identification of genes under selection. Currently, one of the outputs of FrangiPANe is the panreference provided in the form of a fasta file and a table of contig positions on the reference genome. Generating the panreference in the form of a graph (format gfa) would be a real bonus, this panreference could be displayed through pangenomic visualization tools such as Panache (Durant et al. 2021) for instance. Although *de novo* genome assembly from long reads coupled with a graph approach provides new opportunities for pangenomics analyses, there are still several technological limitations that need to be overcome before these analyses can be performed on large numbers of individuals. During this

transition, FrangiPANe offers the opportunity to take advantage of the huge volume of data available to explore plant and animal diversity and identify genes under selection or potential traits associated with structural variation.

To conclude

The idea to develop FrangiPANe was born during a poster presentation with Clothilde Chenal, also PhD student in the DIADE unit, at the JOBIM conference in 2020. In this poster entitled "*How to choose a reference genome (while waiting for a pangenome) ?*", I presented the map-then-assemble approach that we had defined with François Sabot and Yves Vigouroux and the first results of assembling structural variations. C. Chenal, as part of her PhD project , presented the preliminary analysis for choosing the mosquito genome to be used to apply the same approach and the scripts developed at the beginning of my PhD project. After several questions about whether we would develop a common and available tool, we decided to implement FrangiPANe which was used by C. Chenal for building the mosquitoe pangenome. Once structural variations within the two Rice species were identified and the panrefence built, the next steps focused on exploring the inter- and intra-species pangenomes and what we could learn about pangenome structure and dynamics and more broadly about the evolutionary history of the African rice.

IV What can we learn about the African Rice Pangenome ?

11- Background and key scientific outcomes

May the Force be with you – Jedi Master Yoda, Star Wars

11.1 When Pangenome concept meets Population Genetics

High-throughput sequencing technologies have paved the way for the comparison of the cultivated genomes with those of their wild relatives. Numerous diversity analyses have shown that domestication has shaped genomes with a reduction in diversity, mainly allelic diversity (Huang et al. 2012; Hufford, Xu, et al. 2012; Lin, Zhu, et al. 2014; Qi, Liu, et al. 2013). However, domestication consequences on the pangenome have not yet been discussed, although these are emerging issues that have recently been addressed in last plant pangenomics reviews (Khan et al. 2019; Petereit et al. 2022).

After setting up a tool to build the African rice pangenome as proof-of-concept (Tranchant-Dubreuil et al. 2023), we then explored then pangenome diversity during the domestication of the African rice, focusing on how domestication reshapes the African rice pangenome and how selection has acted on its organization and dynamics.

11.2 Key scientific outcomes

We listed here after the main scientific results of an article in progress, a draft of which is presented in the following part 12 (page 109):

- 22,765 genes annotated out the 513.5Mb of new sequences, i.e. a total of 63,318 genes in the panreference comprising the reference genome and the 484,394 new sequences (Table 11.1, page 105);

	Total	Genome Reference	Sequences absents from reference
Total Mb	857	344	513
Genes number	63,318	40,553	22,765
Gene length max.	84,687	84,687	32,209
Gene length mean	2,521	3,188	1,331

Table 11.1: Annotation summary.
This table provides the size (in Mb) of the CG14 Genome Reference and the newly sequences identified by Tranchant-Dubreuil et al. 2023 as well as the total size of the pangenome. It also includes information on genes : number of genes, maximum and average length.

- In our analysis, the African rice pangenome is composed of 64.2% of core genes (present in at least 95% of accessions) and 35.8% of dispensable genes (Figure 11.1a, page 106);
- We observed a variation in the wild and cultivated pangenome size with a total of 60,110 and 57,497 genes in *O. barthii* and *O. glaberrima* respectively. Sharing the same number of core genes (about 39,000 genes), we showed a large number of dispensable genes in the wild species, i.e.

2,613 additional genes. That led to a dispensable ratio of 34.8 and 30.8 in the wild and cultivated accessions (Figure 11.1b, page 106);

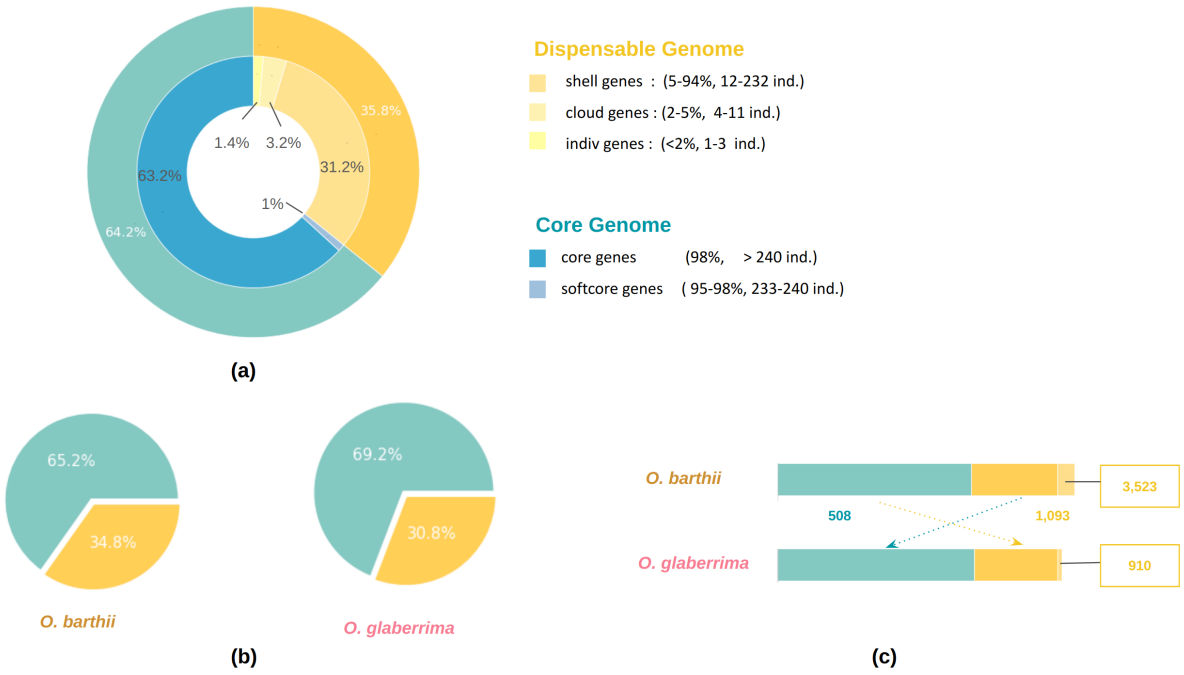

Figure 11.1: Gene-based Pangenomes of the African Rice.
(a) Gene Number in all rice accessions, including domesticated and wild-relative accessions. This nested pie chart displays the number of core and dispensable genes for the 247 accessions (outer ring, teal blue color for core genes and yellow golden for dispensable genes). The inner ring shows the proportion of core genes divided into core and soft genes (blue colors), and dispensables genes divided into shell, cloud and individual genes (yellow colors). (b) Gene Number within each species. The two pie charts indicates the number of core and dispensable genes in the *O. barthii* (left) and *O. glaberrima* (right) accessions. (c) Dynamic of African rice pangenomes by comparing wild and cultivated pangenomes. Each pangenome is represented by a horizontal bar displaying core and dispensable genes, distinctly in golden yellow and teal blue. This image shows that the pangenome of *O.barthii* (top bar) is larger than that of *O. glaberrima* (bottom bar). The last compartiment, at the end of each bar in lighter yellow, indicates the number of genes specific to each species, i.e., 3,523 and 910 genes only present in *O. barthii* and *O. glaberrima* respectively. The transition of genes between the core and dispensable compartments of the pangenomes of the two species is represented by the dashed arrows: 508 core genes in *O. barthii* are classified as dispensable in *O. glaberrima*, while 1,093 dispensable genes in *O. barthii* are found as core in *O. glaberrima*.

- Of the dispensable genome, 3,523 and 910 genes were specific to *O. barthii* and *O. glaberrima*. Functional enrichment analysis identified that wild-specific genes were significantly enriched in polysaccharide and carbohydrate binding, as well as in the defense response to biotic stress, while cultivated-specific ones were enriched in molecular function such as the calmodulin binding;
- Genes were identified as core in the pangenome of one species and dispensable in the pangenome of the other species :
  - 508 core genes in *O. barthii* were identified as dispensable in *O. glaberrima* pangenome;
  - 1,093 dispensable genes in the *O. barthii* pangenome were found to be core genes in *O. glaberrima* pagenome.

  Among these genes switching between core and dispensable compartments according to the species, some were found significantly enriched in pathways such as the nitrogen compound transport, NAD(P)H, quinone, or Peptidyl-prolyl cis/trans isomerases.
- Using the read coverage of the 484,394 sequences to perform a principal component analysis (PCA),

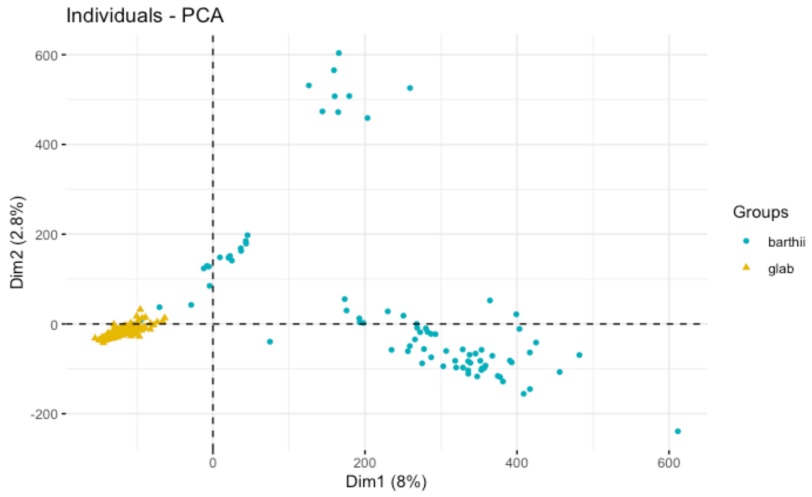

Figure 11.2: Principal Component analysis on new sequences.
This figure shows the PCA based on 484,394 newly sequences identified as absent in the *Oryza glaberrima* reference genome.

we showed that our PCA analysis based on sequence coverage recapitulates the structuration of individuals done with SNPs (Cubry et al. 2018, Figure 11.2, page 107). The wild accessions were clearly separated from domesticated accessions and could be divided into 3 groups (Figure 11.2, page 107).

- Based on our analysis for detection of outliers contributing overwhelmingly to the differentiation between wild and cultivated samples, we identified 7,579 contigs as putatively selected during domestication, of which 683 corresponds potentially to new genes.
- Interestingly, we were able to find selection in the *PROG1* gene, a gene selected during African and Asian rice domestication and associated with a major deletion during African rice domestication. In our analysis, *PROG1* was also specific to wild species as expected.
- In addition to a significant variation in the number of genes, we also observed that structural variation overlapped 30% of the genes in the reference genome and if we considered genes as well as the 5kbp-flanking sequence, this ratio reached 70%.

All the results as well as the material and methods part are detailed in a first version of the article in the next part (part 12, page 109).

12- Article "Domestication reshapes the pangenome in African Rice"

If you keep trying, you will eventually succeed. So the more it fails, the more likely it is to work – Shadocks

Domestication reshapes the pangenome in African Rice

Christine Tranchant-Dubreuil[a,1], **Yves Vigouroux**[a,1], **and Francois Sabot**[a,1]

This manuscript was compiled on April 27, 2023

African rice | pangenome | domestication | structural variations

Rice is the staple food of more than half of humanity (FAO, (1, 2)). To cope with ongoing climate change, breeding rice for a hotter climate is a challenge (3–5). Using the greater genetic diversity found in wild rice is particularly interesting in the current context. Among the 23 species of the genus Oryza, there are today only 2 cultivated species: one in Asia, *Oryza sativa*, and another in Africa, *Oryza glaberrima*. Both species have independent histories and domestication processes (6). The cultivated species *O. glaberrima* was domesticated from the wild rice *O. barthii* in the inner delta of the Niger River in Mali 3,500 years ago (7, 8). Although Asia is the world's leading rice producer and exporter, producing rice with generally preferred taste qualities, *O. glaberrima* nonetheless has interesting agronomic characteristics such as greater resistance to biotic stresses, and better tolerance to drought, salinity and submersion (9).

Domestication reshapes diversity in cultivated crops compared to their wild relatives (10–13), but the consequences on the pangenome structure have not yet been studied. Studies of loss-of-function alleles are over-represented in detecting selection associated with domestication ((14–17)), and the extreme loss of function is reflected in the presence and absence of genes, a feature found when we study pangenomes. How pangenome structures were modified during domestication might bring new insight in the study of plant domestication. Indeed, gene deletions have been found such as the *sh1* gene involved in loss of seed dispersal in rice (18), or the deletion of the *GmCHX1* gene in soybean associated with salt tolerance (19). Gene duplications have also been observed as for the *GL7* gene contributing to grain variations in rice (20), or in tomato, a duplication of *SUN* involved in fruit shape variation (21). Pangenome studies help identify variations between cultivated and wild accessions such as seed coat color in soybean (22), or fruit color in the strawberry (23). Through an extensive pangenome study based on 725 wild and cultivated tomato accessions, Gao et al. observed a large variation in the number of genes within the two species, notably highlighting 4,873 genes absent from the reference genome but present in both cultivated and wild accessions (24). Over the past decade, pangenomics has facilitated access to the vast unexplored content of structural variations, previously generally inaccessible with classical diversity analyses based on a single reference genome. We developed a method to facilitate the study of pangenomes (25). In this study, we built upon this approach to understand how the pangenome was reshaped during rice domestication and identify how selection acted on the pangenomes.

Results

Characterization of genes in the African Rice pangenome. We previously retrieved 513 Mb of assembled sequences identified as absent in the *Oryza glaberrima* reference genome (25). We identified 22,765 additional genes, leading to a total of 63,318 annotated genes (Table 1). A total of 56% and 37% of the 22,765 newly genes could be annotated with gene ontology or Pfam domains respectively, a figure similar in proportions to those observed with genes from the *Oryza glaberrima* reference genome (56% and 39 % respectively).

We then map all 247 resequenced African rice genomes (8, 25) against the pangenome: the 344 Mb *Oryza glaberrima* reference genome (ENA Accession : GCA_000147395) and the new 513 Mb sequences (25). Mapping rate was very high with 98.7% and 98.3% for the domesticated and the wild relative accessions, respectively (Supplementary Table S1). Using SGSgeneloss (26), we could further study 60,110 genes classified (27), according to the percentage of accessions in which they were present (Figure 1a, Supplementary Figure S1). A total of 64.2% of the genes were present in at least 95% of the accessions (233 accessions), of which 63.2% were core genes (frequency higher than 98% of accessions) and 1% softcore genes (frequency between 95% and 98% of accessions). The 35.8% remaining were defined as dispensable genes, consisting of 31.2% of shell genes (between 5% and 94% of accessions), 3.2% cloud genes (frequency between 2% and 5% of accessions) and 1.4% individual genes (frequency of less 2% accessions). Dispensable genes are on average shorter with fewer exon (Supplementary Table S2 and Supplementary Figure S2). Gene Ontology enrichment analysis showed that dispensable genes were significantly enriched notably in recognition of pollen, polysaccharide and carbohydrate binding, multicellular organism development (Supplementary Table S3 and

[a]DIADE, Univ Montpellier, CIRAD, IRD - 911 Avenue Agropolis 34934/34830, Montpellier, FRANCE
Please provide details of author contributions here.
[1]A.O.(Author One) contributed equally to this work with A.T. (Author Two) (remove if not applicable).
[2]To whom correspondence should be addressed. E-mail: author.twoemail.com
Please declare any competing interests here.

Supplementary Figure S3).

If analyzed per species, the pangenome of *O. barthii* was composed of 65.2% core and 34.8% dispensable genes, and not different from the core genome of *O. glaberrima* (fisher test, p-value of 0.6521), reaching 69.2%, and consequently a dispensable genome of 30.8% (Figure 1b). While the number of core genes was similar in both species (about 39,000 genes), the difference was mainly due to a larger number of dispensable genes in the wild species. A total of 2,613 additional genes was observed in *O. barthii* and thus a total gene number of 60,110 and 57,497 genes were observed in *O. barthii* and *O. glaberrima*, respectively (Supplementary Table S4).

Among the dispensable genome, 3,523 and 910 genes were identified as specific to *O. barthii* and *O. glaberrima*, respectively (Figure 1c). For wild-specific genes, its frequency in accessions ranged from 1.2% to 85.7%, with an average of 25% (Supplementary Figure S4), while for cultivated-specific genes, their frequency varies from 0.6% to 55.8% with an average of 4.9% (Supplementary Figure S5). Functional enrichment analysis identified that wild-specific genes were significantly enriched in polysaccharide and carbohydrate binding, as well as in the defense response to biotic stress, while the cultivated-specific ones were enriched in molecular function such as the calmodulin binding (Supplementary Table S5). Interestingly, we also observed that a total of 508 core genes in *O. barthii* were identified as dispensable in *O. glaberrima* pangenome, and 1,093 dispensable genes in the *O. barthii* pangenome were found to be core genes in *O. glaberrima* (Figure 1c). In the *O. glaberrima* dispensable genome, the frequency of the switching-genes varied from 0.6% to 93.9% with an average of 76.7% (Supplementary Figure S7), while in the *O. barthii* dispensable genome, its frequency ranged from 19% to 92.8% with an average of 73.5% (Supplementary Figure S6). For genes present in the dispensable genome of *O. glaberrima* and in the core of *O. barthii*, they were detected as significantly enriched in pathways such as the nitrogen compound transport or NAD(P)H, quinone. For the genes present in the dispensable genome of *O. barthii* and in the core genome of *O. glaberrima*, gene ontology enrichment analysis showed a strong enrichment in Peptidyl-prolyl cis/trans isomerases (Supplementary Table S6).

The analysis per individual showed the number of genes varied between 41,606 and 47,884 genes (Figure 2). A lower average of 44,106 was observed for the cultivated species *O. glaberrima* (Wilcoxon test, p-value $<$ 10-15), compared to an average of 46,130 genes in the wild species *O. barthii* accessions (Figure 2, Supplementary Table S7).

Structural variations associated with the 513Mb assembled sequences. Assembly of pangenome led to 513 Mb in 484,394 sequences. Among them, we identified 22,765 new genes, with 95% of them without similarity with known genes. We were able to anchor 33% of them on the reference genome (25). A total of 81.8% of these anchored sequences (124,636 sequences) had a gene in their vicinity (±5 kbp of coding sequence). In addition, a total of 79% of the genes present in the genome had at least one structural variation placed inside or its flanking regions of 5 kbp. Finally, 30% of genes (*i.e.* 12,666 genes) had a structural variation directly within their sequence (Figure 3) and of these 12,666 genes, 56% have a variation in an exon (Supplementary Table S8).

Selection of PAVs during domestication. We used the read coverage of the 484,394 sequences to perform a principal component analysis (PCA). We showed that our PCA analysis based on sequence coverage (Figure 4) recapitulates the structuration of individuals done with SNPs (8). The wild accessions were clearly separated from domesticated accessions and could be divided into 3 groups (Figure 4). Based on our analysis for detection of outliers contributing overwhelmingly to the differentiation between wild and cultivated samples, we identified 7,579 contigs as putatively selected during domestication. We could link these contigs to 1476 genes in the reference genome in which the contigs were anchored and to 683 new genes annotated on new sequences. We previously showed the *PROG1* gene was present in wild *O. Barthii* and absent from *O. glaberrima* (8). The complete deletion of this gene was associated with African rice domestication [REF, (8)]. Our PCA approach was effectively able to recover a putative selection signature on a *PROG1* re-assembled contigs (Supplementary Table S7), as expected. We end up founding two different contigs with an identity higher than 99% (Supplementary Table S7). When comparing their sequences, both contigs aligned well with each other, except for a region of less than 1000 bp (Supplementary Figure S8a). One contig has a G->A mutation that causes a cyteine-to-tyrosine change in the protein-coding-region (Supplementary Figure S8b). It is unclear if two *PROG1* genes exist in wild *O. barthii*, or if there are different assembled alleles. Neither the less, finding selection signature in the *PROG1* gene validated our approach based on coverage and PCA. Among the 683 genes identified, 237 had unknown function and 446 had putative function (Supplementary Table S9). Among the 446 genes potentially under selection and with a putative function, 227 genes were identified as belonging to the same compartment in both species, including 3 and 224 genes in the core and dispensable compartments respectively. 7 genes were identified both in the core of one species and in the dispensable genome of another species, including 1 and 6 present respectively in the dispensable genome of *O. barthii* and *O. glaberrima* respectively. For example, among thes genes switching between core and dispensable compartments of bioth species, a gene encoding a G-type lectin S-receptor-like serine/threonine protein kinase was found as dispensable in the pangenome of the wild species whereas it was core in the pangenome of the cultivated species. In contrast, a gene annotated as coding for a phototropin was identified as core in the wild and dispensable in the cultivated. Finally, 211 genes present only in one of the two species were identified as potentially under selection, including 31 genes specific to cultivated accessions and 180 genes specific to wild accessions. As an example, the PROG1 gene described in previous studies (8) was found to be one of the specific genes present only in wild accessions.

Discussion

Pangenomic analysis of both wild and cultivated species can help to better characterize the extensive gene pool and variability within the two species, and thus better understanding the impact of both natural and artificial selection and genetic drift on the pangenome of each species. Taking advantage of a publication of an African Rice panreference (25), we report here the pangenome of the cultivated African rice *O. glaberrima* and its wild relative *O. barthii*, based on 248 rice genomes (8, 25) and this panreference . Our results estimated an african rice super pangenome of 63,318 genes, of which 22,765 are absent from the reference genome. This result is slightly similar to the 19,319 new genes reported in the analysis based on 111 Asian rice long-read sequenced genomes, including

wild accessions. Using a similar approach to ours, Zhao et al. found a lower number of new genes (10,872 genes), but from a smaller number of cultivated and wild accessions, 3 times less, than in our study (28). The African Rice pangenome was composed of 64.2% of core genes and 35.8% of dispensable genes. Different analyses describing pangenomes of cultivated and wild Asian rice (*O. sativa* and *O. rufipogon*) estimated the ratio of core genes between 42.6% and 61.9% (28–30). While our results are in agreement with the higher ratio of 61.7% described by Zhao et al., it is quite different from the ratio of 42.6% and 51.5% estimated respectively in a study using 251 accessions (including 28 asian wild relatives) (29) and in another analysis based on 111 accessions (6 wild) (30). These differences might depend on the sampling based on *O. Rufipogon* which has a higher genetic diversity than African rices (29), or on the approaches used such as the annotation or core/dispensable classification procedure for instance. In addition to detecting a surprising number of novel genes, our study showed a significant variation in the number of genes between cultivated and wild. While it is expected that the gene pool of a wild type has a higher allelic diversity, our results highlighted that the pangenome of O. barthii is larger with a total size of 60,110 genes compared to 57,497 genes in that of the cultivated. In a super pan-genome of 214 asian rice accessions, Shang et al. described similar results with a greater number of genes in the pangenome of the wild species *O. rufipogon* than in the cultivated species (29). We also provide first insight into how domestication has shaped the African Rice pangenome and the pangenome dynamics of each species in interaction with its environment and under the effect of evolutionary force. Thus, our results showed a surprising number of species-specific genes or genes switching between core and dispensable compartments of both species. As expected, we also assembled the PROG1 gene, which was absent from the reference and our analysis confirmed that it was a wild-specific gene or otherwise completely absent in all cultivated rice accessions (ref). As described in other plant pangenomes (26, 30, 31), genes significantly enriched in the defense response to biotic stress were identified among wild-specific genes. Among the genes significantly enriched switching between core and dispensable compartments according to the species, we also identified a peptidyl-prolyl cis/trans isomerase. If these genes have not been well characterized in plants, evidence suggests they are associated with host/pathogen interaction in eucaryotes (32). Moreover, among the genes switching between the two pangenome compartments of species, we found new genes with a selection signature that could be particularly interesting such as genes involved in the tolerance to salt stress (G-type lectin S-receptor-like serine/threonine protein, (33)) or in plant growth (Phototropins, (34–36)) for instance. In addition to a significant variation in the number of genes, we also observed that structural variation overlapped 30% of the genes in the reference genome and if we considered genes as well as the 5kbp-flanking sequence, this ratio reached 70%. This result is quite consistent with what was observed in Asian rice (29), where 35.4% of the genes were affected by structural variations in the rice pangenome based on 251 accessions sequenced in long read. Through their insertion nearby genes, these variations might influence gene expression by modifying gene structure or by impacting on cis-regulatory sequences (37). This work is a step forward in better understanding how the pangenomes of a group of closely related species have been shaped by evolutionary processes such as domestication. Pangenomes offer robust opportunities to access unexplored content of structural variations, including new genes and increased gene diversity, potentially interesting for future crop improvement in the climate changes.

Methods

Sample sequencing.

Genome sequencing of 247 African rice accessions. Whole-genome sequencing data from 247 African rice accessions used in this study were described previously in (8, 38) (Illumina technology, 100-150 bp paired-end reads with an average depth of sequencing of 28X minimal). Sequencing data included 163 domesticated and 84 wild relative individuals which represented the species-wide genetic diversity for both species.

Genes content characterization.

Distribution of contigs placed on genes in reference genomes. We identified contigs placed within genes or in a flanking region (+- 5 kbp of coding sequence) using the software bedtools intersect (v2.29.2) (39) with annotation data and contigs positions on the genome (doi:10.23708/93OQMD).

Contigs annotation. The newly assembled sequences (doi:10.23708/93OQMD) were annotated using MAKER2 (v. 2.31.9) (40) with *de novo* gene prediction performed by Augustus and Snap, alignments evidence with rice ESTs and proteins libraries (uploaded from NCBI (Jan 27, 2022)). Genes were annotated by aligning their protein sequences to NCBI non redundant sequences (M

Alignment of 247 rice genome resequencing against the pangenome.

Mapping. The Whole Genome Sequencing reads were mapped according to the protocol described in (8, 25). Raw reads from each accession were aligned to the pangenome using bwa (v. 0.7.15) (41) and the properly mapped-in pair reads were kept with SAMtools view (v1.3.1) (42).

Reads count on contigs. Using only reads mapped with a quality higher than 10, the number of reads mapped is calculated both for each contig and for each accession, followed by a normalization step taking account of the sequencing depth of each accession.

Detection of selection signature. A principal component analysis was carried out on the contigs (based on the normalized matrix of read counts), using the R package FactoMineR ((43)).We performed genome scans for signatures of local adaptation using the R package pcadapt ((44), v4.3.3, k=1) with PAVs and only PAVs with an FDR < 0.05 (Mahalanobis distance, Benjamini-Hochberg method) were considered as outliers.

Gene Presence/Absence Variation analysis.

SGSgeneloss analysis The presence or absence of each gene in each accession was determined using the software SGSGeneLoss v0.1 ((26)). Using the mapping results, a given gene was considered as present in a given accession if more than 20% of its exon regions were covered by at least two reads (minCov = 2, lostCutoff = 0.20), otherwise it was considered as absent.

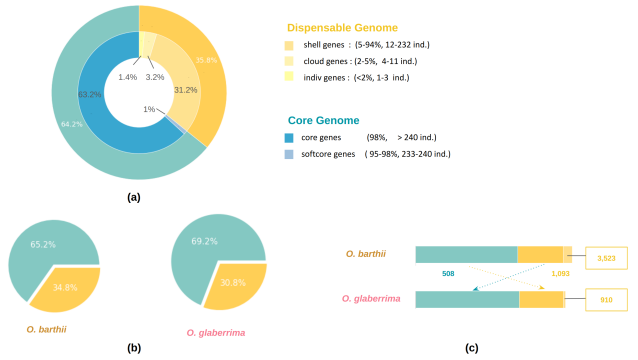

Fig. 1. Gene-based Pangenomes of the African Rice. (a) Gene Number in all rice accessions, including domesticated and wild-relative accessions. This nested pie chart displays the number of core and dispensable genes for the 247 accessions (outer ring, teal blue color for core genes and yellow golden for dispensable genes). The inner ring shows the proportion of core genes divided into core and soft genes (blue colors), and dispensables genes divided into shell, cloud and individual genes (yellow colors). (b) Gene Number within each species. The two pie charts indicates the number of core and dispensable genes in the *O. barthii* (left) and *O. glaberrima* (right) accessions. (c) Dynamic of African rice pangenomes by comparing wild and cultivated pangenomes. Each pangenome is represented by a horizontal bar displaying core and dispensable genes, distinctly in golden yellow and teal blue. This image shows that the pangenome of *O.barthii* (top bar) is larger than that of *O. glaberrima* (bottom bar). The last compartiment, at the end of each bar in lighter yellow, indicates the number of genes specific to each species, i.e., 3,523 and 910 genes only present in *O. barthii* and *O. glaberrima* respectively. The transition of genes between the core and dispensable compartments of the pangenomes of the two species is represented by the dashed arrows: 508 core genes in *O. barthii* are classified as dispensable in *O. glaberrima*, while 1,093 dispensable genes in *O. barthii* are found as core in *O. glaberrima*.

Core and dispensable pangenes Pangenes were divided into 5 categories based on the number of accessions containing a gene : (1) core genes present in more than 240 accessions ($>$ 98%), (2) softcore genes present in 233-240 accessions (95-98%), (3) shell genes present in 12-232 accessions (5-94%), (4) cloud genes present in 4-11 samples (2-5%) and individual genes present in 1-3 samples ($<$2%).

Comparison of core and variable genes Core and variable genes were compared respectively to gene length, exon number

Gene ontology terms analysis. The frequencies of GO terms in core genes or dispensable genes were compared with those of all genes annotated in the pangenome, using the Fisher exact test (classic and weighted) implemented in the R package TopGO ref, and those with p-value lower than 0.05 were considered as significantly enriched.

Analysis of PROG1. The protein sequence of the *PROG1* gene was aligned to all newly assembled sequences (25) using BLAST (45, 46) and only alignments covering at least 90% of the gene length were kept with an e-value threshold of 10e-30. The newly contigs identified by the blast analysis were aligned against each other with each other using the NUCMER software (47) and after filtering according to a minimum alignment length of more than 300 bp and an identity ratio of 90%, alignments were visualized with mummerplot. Multiple alignment of the contigs and the gene was performed with the MUSCLE software (48).

Tables.

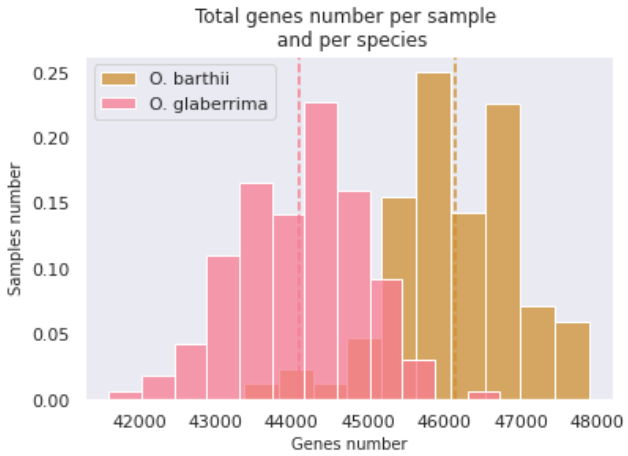

Fig. 2. Distribution of total genes number over *O. barthii* accessions (brown) and *O. glaberrima* (red).

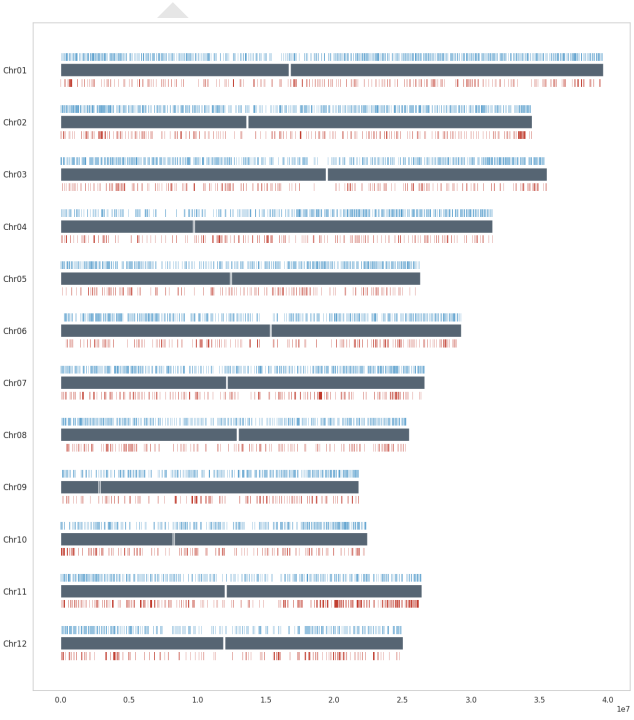

Fig. 3. Genes and PAV location on the 12 chromosomes of CG14. Each chromosome is represented with two tracks : (a) Genes annotated on the genome (blue), (b) Genes with a PAV, representing 30% of the total number of genes.

Supporting Information Appendix (SI).

Acknowledgments

This work was supported by a grant from the France Genomique French National infrastructure and funded as part of “Investissement d’avenir” (ANR-10-INBS-09) and the IRIGIN project. The authors acknowledge Ndomassi Tando and the ISO 9001 certified IRD itrop HPC (member of the South Green Platform) at IRD montpellier as well as the TGCC platform for providing HPC resources

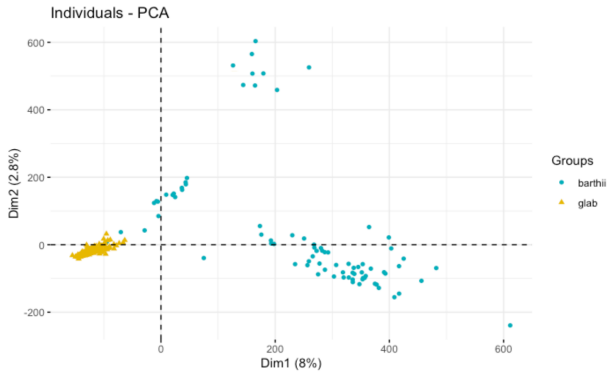

Fig. 4. Principal Component analysis based on 484,394 newly sequences identified as absent in the *Oryza glaberrima* reference genome. This figure shows ...

Table 1. Annotation summary.

	Total	Refere Genome	New Sequences
Total Mb	857	344	513
Genes number	63,318	40,553	22,765
Gene length max.	84,687	84,687	32,209
Gene length mean	2,521	3,188	1,331

This table provides the size (in Mb) of the CG14 Genome Reference and the newly sequences identified by (25) as well as the total size of the pangenome. It also includes information on genes : number of genes, maximum and average length.

that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/- http://www.southgreen.fr.

1 NK Fukagawa, LH Ziska, Rice: Importance for Global Nutrition. *J Nutr Sci Vitaminol* **65**, 2–3 (2019).
2 PA Seck, A Diagne, S Mohanty, MCS Wopereis, Crops that feed the world 7: Rice. *Food Secur.* **4**, 7–24 (2012).
3 C Zhao, et al., Plausible rice yield losses under future climate warming. *Nat. Plants* **3**, 16202 (2016).
4 C Chen, et al., Global warming and shifts in cropping systems together reduce China's rice production. *Glob. Food Secur.* **24**, 100359 (2020).
5 S Saud, et al., Comprehensive Impacts of Climate Change on Rice Production and Adaptive Strategies in China. *Front. Microbiol.* **13**, 2254 (2022).
6 DA Vaughan, H Morishima, K Kadowaki, Diversity in the Oryza genus. *Curr. Opin. Plant Biol.* **6**, 139–146 (2003).
7 M Wang, et al., The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. *Nat. genetics* **46**, 982–8 (2014).
8 P Cubry, et al., The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. *Curr. Biol.* (2018).
9 Y Agnoun, SSH Biaou, M Sié, RS Vodouhè, A Ahanchédé, The African Rice Oryza glaberrima Steud: Knowledge Distribution and Prospects. *Int. J. Biol.* **4** (2012).
10 T Lin, et al., Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**, 1220 (2014).
11 X Huang, et al., A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
12 MB Hufford, et al., Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
13 J Qi, et al., A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* **45**, 1510 (2013).
14 MT Sweeney, MJ Thomson, BE Pfeil, S McCouch, Caught Red-Handed: Rc Encodes a Basic Helix-Loop-Helix Protein Conditioning Red Pericarp in Rice. *The Plant Cell* **18**, 283–294 (2006).
15 L Tan, et al., Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* **40**, 1360–1364 (2008).
16 W Wu, et al., A single-nucleotide polymorphism causes smaller grain size and loss of seed shattering during African rice domestication. *Nat. Plants* **3**, 17064 (2017).
17 KT Win, et al., A single base change explains the independent origin of and selection for the nonshattering gene in African rice domestication. *New Phytol.* **213**, 1925–1935 (2017).
18 Z Lin, et al., Parallel domestication of the Shattering1 genes in cereals. *Nat. Genet. 2012 44:6* **44**, 720–724 (2012).
19 X Qi, et al., Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nat. Commun. 2014 5:1* **5**, 1–11 (2014).
20 Y Wang, et al., Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet. 2015 47:8* **47**, 944–948 (2015).
21 H Xiao, N Jiang, E Schaffner, EJ Stockinger, E Van Der Knaap, A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319**, 1527–1530 (2008).
22 JM Song, et al., Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat. Plants* **6**, 34–45 (2020).
23 Q Qiao, et al., Evolutionary history and pan-genome dynamics of strawberry (Fragaria spp.). *Proc. Natl. Acad. Sci.* **118**, e2105431118 (2021).
24 L Gao, et al., The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019).
25 TD Christine, et al., FrangiPANe, a tool for creating a panreference using left behind reads. *NAR Genomics Bioinforma.* **5** (2023).
26 AA Golicz, et al., Gene loss in the fungal canola pathogen Leptosphaeria maculans. *Funct. & Integr. Genomics 2014 15:2* **15**, 189–196 (2014).
27 SP Gordon, et al., Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
28 Q Zhao, et al., Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* p. 1 (2018).
29 L Shang, et al., A super pan-genomic landscape of rice. *Cell research* **32**, 878–896 (2022).
30 F Zhang, et al., Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome research* **32**, 853–863 (2022).
31 Y Liu, et al., Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
32 M Steinert, Editorial: Peptidyl-prolyl cis/trans isomerases (PPIases) in host-pathogen interactions. *Front. Cell. Infect. Microbiol.* **12** (2022).
33 XL Sun, et al., GsSRK, a G-type lectin S-receptor-like serine/threonine protein kinase, is a positive regulator of plant tolerance to salt stress. *J. Plant Physiol.* **170**, 505–515 (2013).
34 J Wang, et al., Phototropin 1 Mediates High-Intensity Blue Light-Induced Chloroplast Accumulation Response in a Root Phototropism 2-Dependent Manner in Arabidopsis phot2 Mutant Plants. *Front. Plant Sci.* **12** (2021).
35 K Sakamoto, WR Briggs, Cellular and Subcellular Localization of Phototropin 1 . *The Plant Cell* **14**, 1723–1735 (2002).
36 E Demarsy, et al., Phytochrome Kinase Substrate 4 is phosphorylated by the phototropin 1 photoreceptor. *The EMBO J.* **31**, 3457–3467 (2012).
37 C Feschotte, Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
38 C Monat, et al., de novo assemblies of three Oryza glaberrima accessions provide first insights about pan-genome of African rices. *Genome Biol. Evol.* p. evw253 (2016).
39 AR Quinlan, IM Hall, Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
40 C Holt, M Yandell, MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinforma.* **12**, 1–14 (2011).
41 H Li, R Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. (Oxford, England)* **25**, 1754–60 (2009).
42 H Li, et al., The Sequence Alignment/Map format and SAMtools. *Bioinforma. (Oxford, England)* **25**, 2078–9 (2009).
43 S Lê, J Josse, F Husson, Factominer: An r package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
44 F Privé, K Luu, BJ Vilhjálmsson, MG Blum, Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. *Mol. Biol. Evol.* (2020).
45 A SF, G W, M W, M EW, L DJ, Basic local alignment search tool. *J. molecular biology* **215**, 403–410 (1990).
46 SF Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
47 G Marçais, et al., MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, 1–14 (2018).
48 RC Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

Supplementary Figures and Tables for Domestication reshapes the pangenome in African Rice

Christine Tranchant-Dubreuil, Yves Vigouroux, Francois Sabot

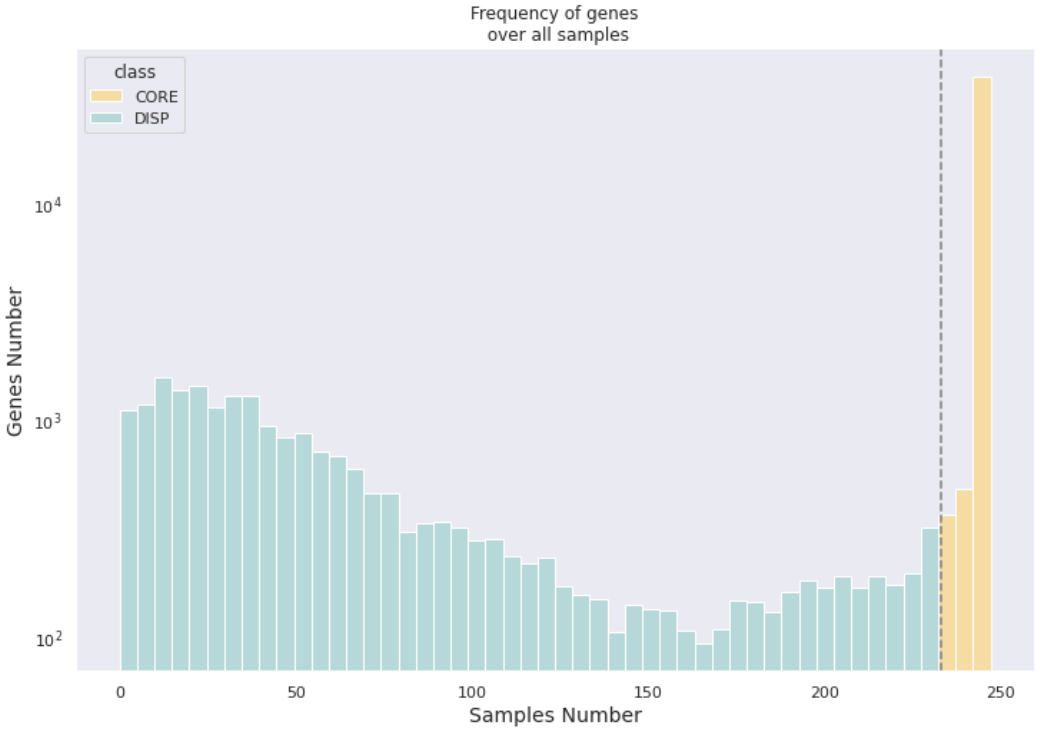

Fig. 1. Gene frequency distribution found across the 247 African Rice accessions (wild and cultivated species). This histogram displays the frequency of genes present in the 247 accessions. The color of the bar indicates whether this number of genes was found in more than 95% of the accessions (232 accessions, golden yellow) or less than 95% (teal blue color).

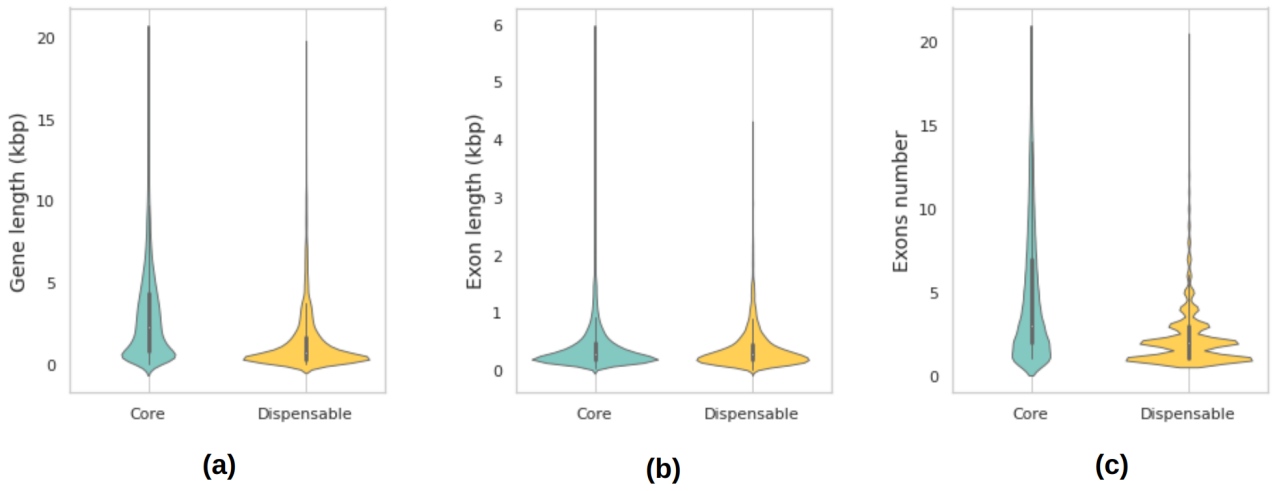

Fig. 2. Statistics about genes of the African Rice pangenome, including both species. (a) Violin plots showing the gene length distribution in the core and dispensable compartments. (b) Violin plots showing the exon length distribution in the core and dispensable compartments. (c) Violin plots showing the exon number distribution in the core and dispensable compartments.

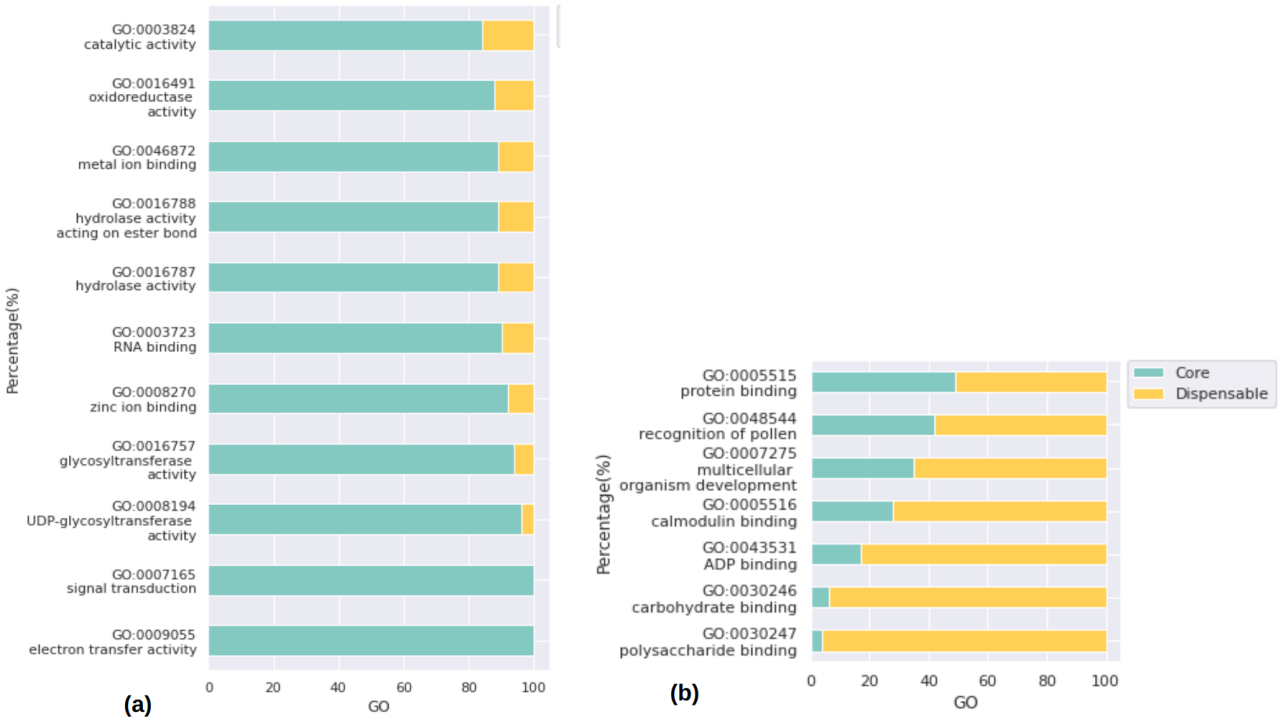

Fig. 3. GO term enriched in the pangenome of the African rice (cultivated and wild species). These graphs show the proportion of genes with an enriched GO term in the core genome (a, left) and in dispensable compartments (b, right). In each graph, the proportion of genes with an enriched GO term is colored blue or yellow depending on whether they are part of the core or dispensable compartments. Only GO-term identified as enriched with a p-value lower than 1e-05 are shown.

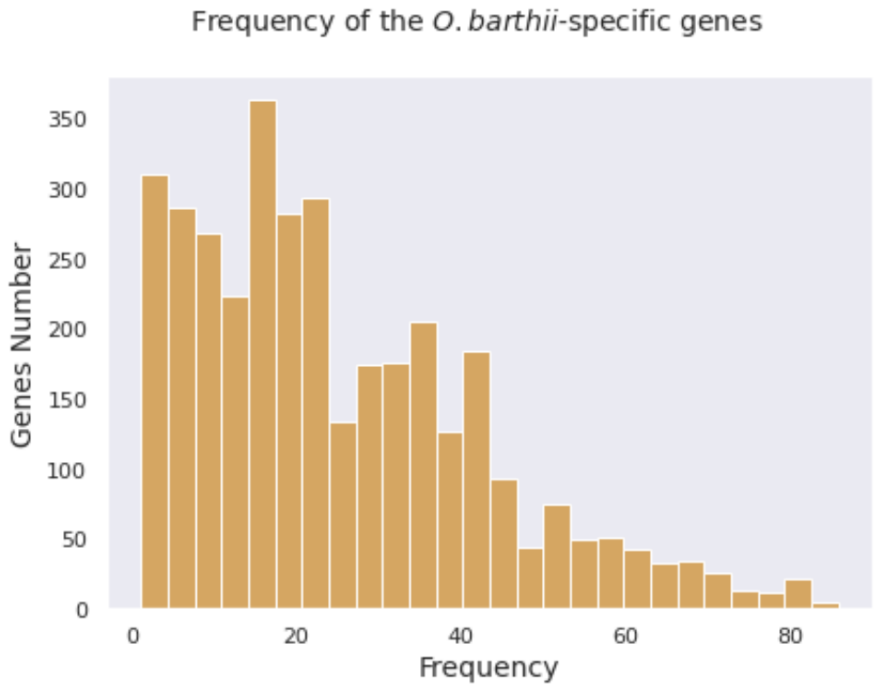

Fig. 4. Frequency of the wild-specific gene. This plot shows the distribution of the 3,523 *O. barthii*-specific genes frequencie across all *O. barthii* accessions.

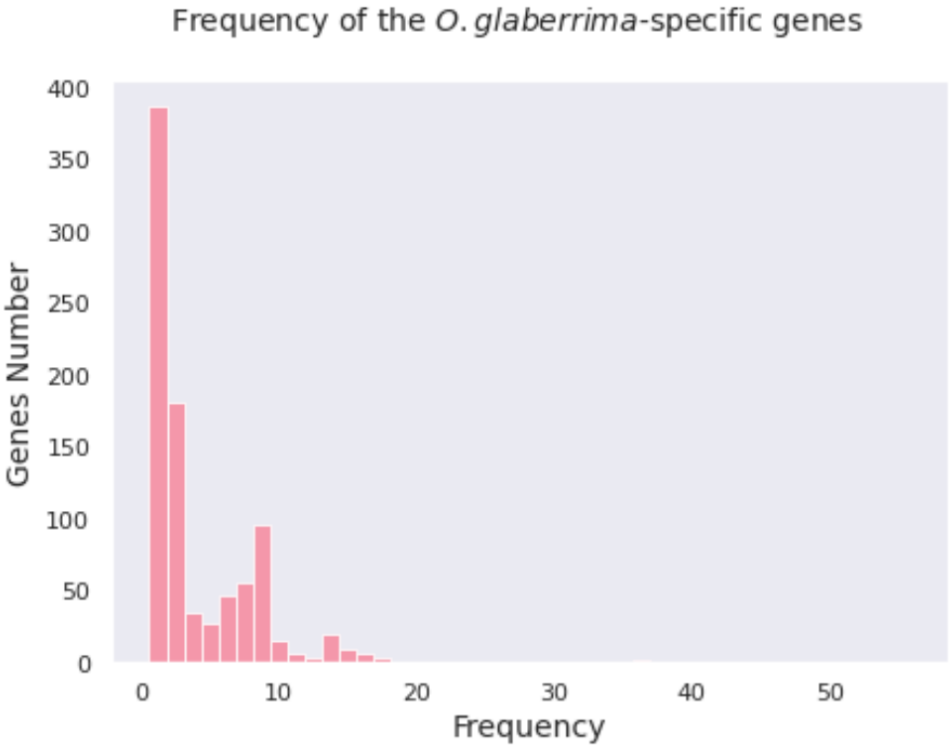

Fig. 5. Frequency of the cultivated-specific gene. This plot shows the distribution of the 910 *O. glaberrima*-specific genes frequencies across all *O. glaberrima* accessions.

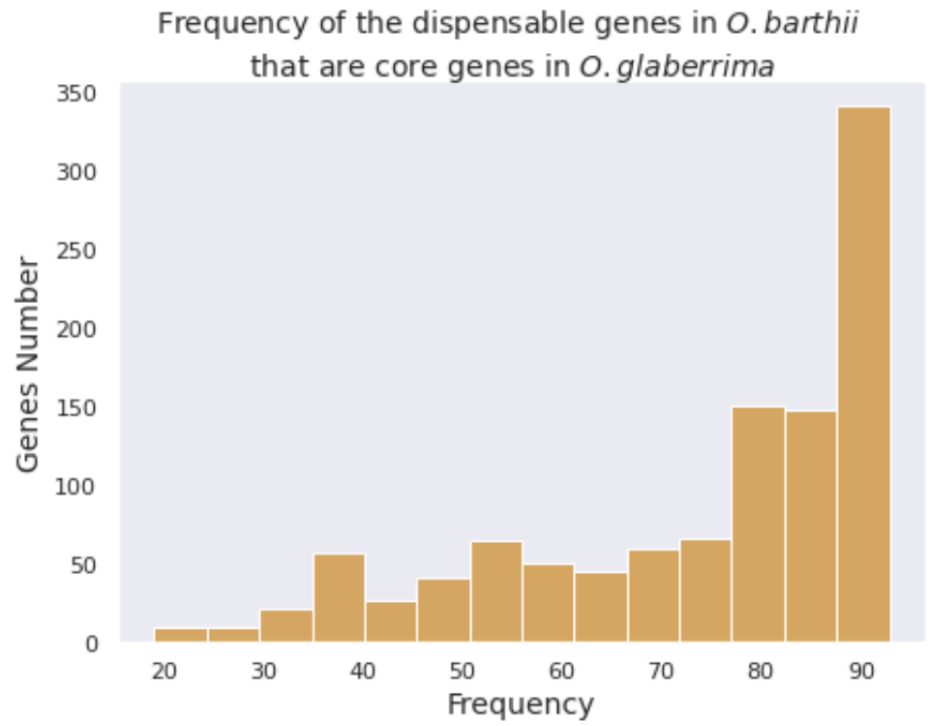

Fig. 6. Frequency of the switching-genes in the African rice wild species. The plot shows the distribution of the 1,093 dispensable genes in *O. barthii* that were identified as core in *O. glaberrima* pangenome.

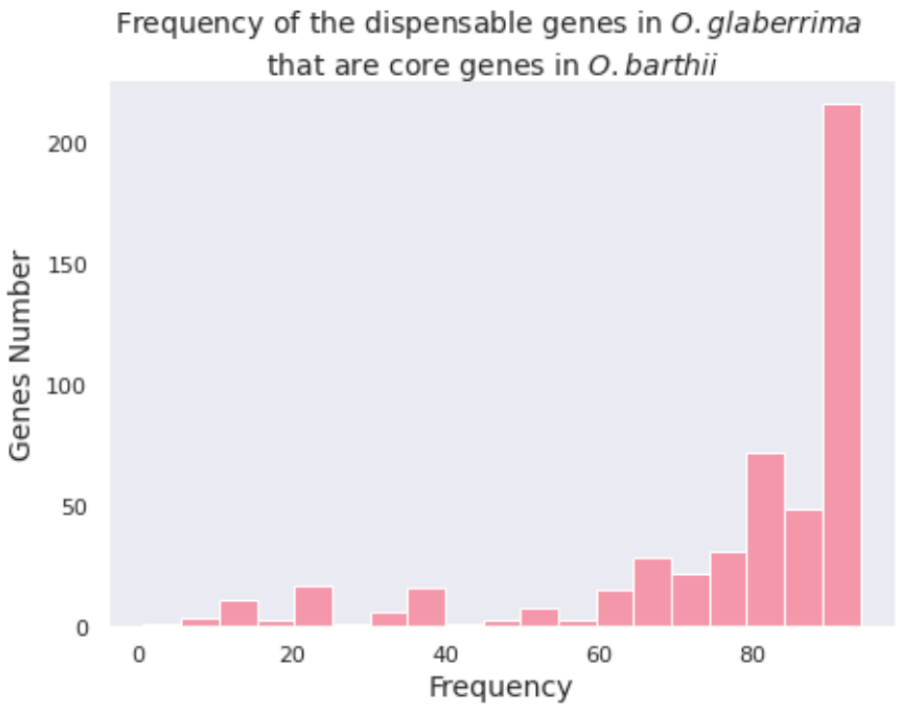

Fig. 7. Frequency of switching-genes in the African rice cultivated. The plot shows the distribution of the 508 dispensable genes in *O. glaberrima* that were identified as core in *O. barthii* pangenome.

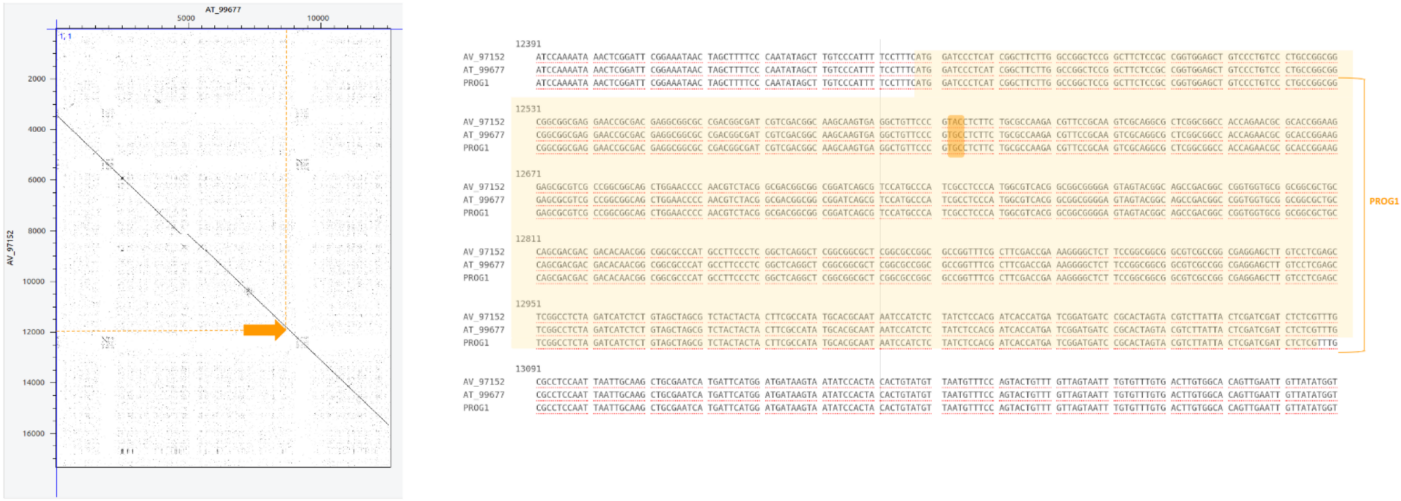

Fig. 8. ***PROG1*** **identification.** (a) Dotter alignment of the two contigs identified with PROG1 gene. The PROG1 gene is placed on the two contigs in orange. (b) Multiple alignment of the PROG1 gene from O. sativa (chromosome 7) and from the two contigs. The yellow block in light yellow corresponds to the gene region and a G->A mutation identified is in orange..

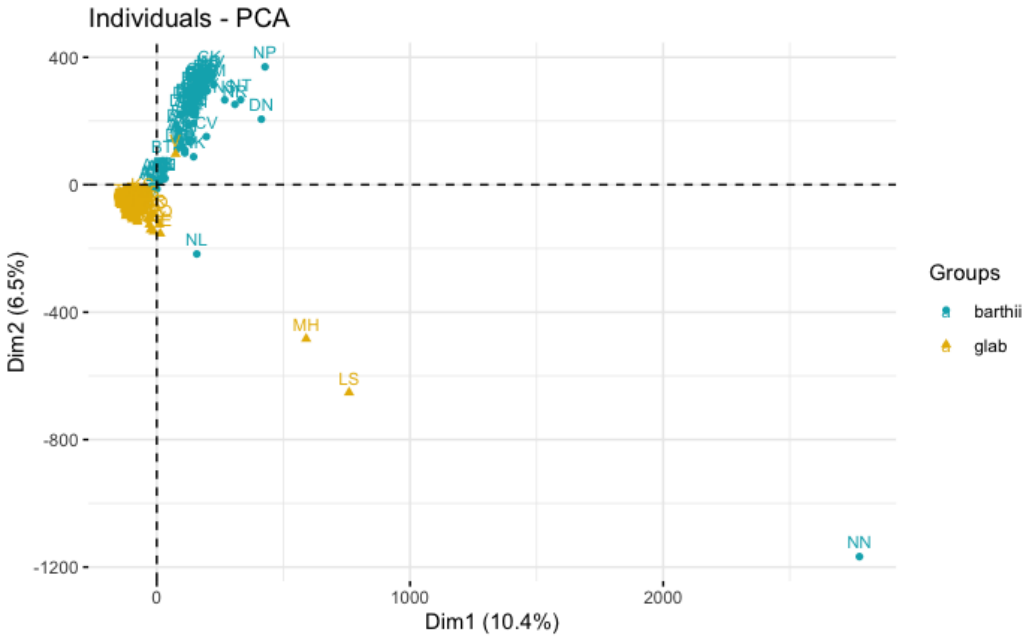

Fig. 9. Principal Component analysis based on 484,394 newly sequences. The accessions NN, LS and MH were considered as outliers and removed from further analysis.

	All species	*O. barthii*	*O. glaberrima*
All Mapped Reads			
min-max	97.1 - 99.1	97.3 - 99.1	97.1 - 99.1
mean	98.5	98.3	98.7
Reads correctly mapped in pair			
min-max	82.1 - 97.64	82.1 - 97.5	88.2 - 97.6
mean	96.2	95.2	96.7

Table 1. Percentage of reads mapping for all 248 African cultivated and wild rice accessions.

For each group of reads (raw reads mapped and reads correctly mapped in pair) and for each accession group (both species together, *O. barthii* and *O. glaberrima*), the minimum and maximum mapping rate on the reference genome as the mean percentage of mapped reads are provided.

	Core	Dispensable
Gene length max.	51,855	84,687
Gene length mean	3,201	1,442
Exon number max	78	25
Exon number mean	5.5	2.3
Exon size max	18,096	4,218
Exon size mean	410	383

Table 2. Annotation summary of the genes of the African Rice pangenome.

This table provides statistics on core and dispensables genes of the African Rice pangenome, including information on both genes (maximum and average length) and exons (maximum and average number of exons, maximum and average exon length).

Comp.	Class	GO id	GO term	Genes	P-value (classic)	P-value (weight)
core	MF	GO:0003723	RNA binding	320	<1e-30	< 1e-30
core	MF	GO:0008270	zinc ion binding	232	< 1e-30	< 1e-30
core	MF	GO:0008194	UDP-glycosyltransferase activity	178	< 1e-30	< 1e-30
core	MF	GO:0016787	hydrolase activity	476	< 1e-30	< 1e-30
core	MF	GO:0016788	hydrolase activity, acting on ester bonds	266	< 1e-30	< 1e-30
core	MF	GO:0046872	metal ion binding	489	< 1e-30	1.3e-27
core	MF	GO:0016491	oxidoreductase activity	338	< 1e-30	5.2e-21
core	MF	GO:0003824	catalytic activity	1905	< 1e-30	1.9e-19
core	MF	GO:0016757	glycosyltransferase activity	272	< 1e-30	2.7e-17
core	MF	GO:0009055	electron transfer activity	49	1.2e-14	1.2e-14
core	BP	GO:0009733	response to auxin	59	0,0026	0,0026
core	BP	GO:0000160	phosphorelay signal transduction system	27	0,0251	0,0251
core	BP	GO:0007165	signal transduction	49	0,0011	0,0439
dispensable	MF	GO:0043531	ADP binding	1806	< 1e-30	< 1e-30
dispensable	MF	GO:0030247	polysaccharide binding	497	< 1e-30	< 1e-30
dispensable	MF	GO:0030246	carbohydrate binding	675	< 1e-30	< 1e-30
dispensable	BP	GO:0007275	multicellular organism development	35	4.3e-13	5.6e-14
dispensable	MF	GO:0005516	calmodulin binding	101	1,40E-06	1.40E-06
dispensable	MF	GO:0005515	protein binding	2895	2,40E-05	2.10E-06
dispensable	BP	GO:0048544	recognition of pollen	13	3.4e-05	3.4e-05

Table 3. List of Gene Ontology Terms identified as enriched in the core and dispensable compartments of the African rice pangenome (wild and cultivated species).

This table provides, for each GO term identified as enriched, the compartiment (core or dispensable) in which it was found enriched, the ontology class to which it belongs (Biological Process or BP, Molecular Function or MF) , the identifier and name of the GO term, the number of genes annotated with this GO term, the classical P-value and the weighted P-value.

	Core genes	Dispensable genes	Total pangenome genes
Oryza barthii	39,191	20,919	60,110
Oryza glaberrima	39,776	17,721	57,497

Table 4. Composition of the pangenome of *O. barthii* and *O. Glaberrima*.

This table provides the number of core and dispensable genes as well as the total gene number in the pangenomes of *O. barthii* and *O. glaberrima*.

Species	Cat	GO term ID	GO term	P-value (classic)	P-value (weight)
O. barthii	MF	GO:0030247	polysaccharide binding	< 1e-30	< 1e-30
O. barthii	MF	GO:0030246	carbohydrate binding	< 1e-30	1.2e-05
O. barthii	MF	GO:0043531	ADP binding	8.3e-07	8.3e-07
O. barthii	BP	GO:0098542	defense response to other organism	0.00016	0.00016
O. barthii	MF	GO:0008168	methyltransferase activity	0.00022	0.0027
O. barthii	MF	GO:0016836	hydrolyase activity	0.00526	0.0053
O. barthii	BP	GO:0006511	ubiquitin-dependent protein catabolic process	0.01075	0.01075
O. barthii	BP	GO:0045927	positive regulation of growth	0.02782	0.02782
O. barthii	BP	GO:0015074	DNA integration	0.04147	0.04147
O. glaberrima	MF	GO:0005516	calmodulin binding	1.5e-08	1.5e-08
O. glaberrima	MF	GO:0008061	chitin binding	2.5e-06	2.5e-06
O. glaberrima	MF	GO:0043531	ADP binding	3.5e-05	3.5e-05
O. glaberrima	MF	GO:0004386	helicase activity	0.00013	0.00013
O. glaberrima	BP	GO:0006629	lipid metabolic process	0.00013	0.00013
O. glaberrima	MF	GO:0005515	protein binding	0.00073	0.02185
O. glaberrima	MF	GO:0005507	copper ion binding	0.00592	0.00592

Table 5. List of the Gene Ontology Term for species-specific genes found significantly enriched in *O. barthii* and *O. glaberrima*.

This table provides the ontology class, the GO term identifier, the GO term, the classical and weighted P-value. The ontology is divided into three classes: Biological Process (BP), Molecular Function (MF), Cellular Component (CC). Only GO-term identified as enriched with a p-value lower than 1e-05 are shown.

Dispensable	Class	GO term ID	GO term	P-value (classic)	P-value (weight)
O. glaberrima	BP	GO:0071705	nitrogen compound transport	2.1e-08	2.1e-08
O. glaberrima	BP	GO:0006810	transport	2.4e-05	1.00
O. glaberrima	MF	GO:0043531	ADP binding	0.0012	0.0012
O. glaberrima	MF	GO:0042626	ATPase-coupled transmembrane transporter activity	0.0339	0.0339
O. glaberrima	MF	GO:0016655	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	0.0405	0.0405
O. barthii	MF	GO:0043531	ADP binding	3.6e-05	3.6e-05
O. barthii	MF	GO:0003755	peptidyl-prolyl cis-trans isomerase activity	0.0025	0.0025

Table 6. List of the Gene ontology Term for genes switching between the core and the dispensable compartments of *O. barthii* and *O. glaberrima* and found as significantly enriched in the dispensable genome.

This table provides the ontology class, the GO term identifier, the GO term, the classical and weighted P-value. The ontology is divided into three classes: Biological Process (BP), Molecular Function (MF), Cellular Component (CC). The five first lines describe the GO enrichment analysis of genes that switched from the core compartment of *O. barthii* to the dispensable compartment of *O. glaberrima*. The two last rows indicate the GO enrichment analysis of genes that switched between the core genome of *O. glaberrima* and the dispensable genome of *O. barthii*. Only GO-term identified as enriched with a p-value lower than 1e-05 are shown.

Number of genes per genome	Minimum	Maximum	Mean
Oryza barthii	43,368	47,884	46130
Oryza glaberrima	41606	46721	44,106

Table 7. Genes number per accession. This table provides the number of minimum, maximum and mean number of genes found among the genomes of *O. barthii* and *O. glaberrima*.

	Contigs number	Total contigs	Gene number
within genes	56,467	56,467	12,270
+/- 500bp	14,227	70,694	15,519
+/- 1kb	25,857	82,324	18,580
+/- 5kb	68,169	124,636	31,889

Table 8. Location of Structural Variations in close proximity to or within genes.

This table provides respectively the number of structural variations placed within a gene or in a region surrounding the gene up to 5 kbp and the number of genes involved.

	Genes Number	Total
PUTATIVE FUNCTION		**446**
African rice pangenome		*227*
Core genome	3	
Dispensable genome	224	
Dispensable compartment		*8*
O. barthii	7	
O. glaberrima	1	
Species-specific genes		*211*
O. barthii	181	
O. glaberrima	30	
UNKNOWN PROTEIN		**237**
TOTAL		**683**

Table 9. Statistics about genes under selection.

This tables provides the distribution of the 683 genes in the African Rice pangenome putatively identified under selection. It gives the number of genes that belonged to the core or dispensable compartments in the African Rice pangenome comprising both species, that were identified in both the core compartment of one species and in the dispensable compartment of another specie , or that are present in only one of the two species.

13- Discussion and Conclusion

You don't know the power of the dark side! I must obey my master – Darth Vader's internal struggle between the light side and the dark side, Star Wars

Initial results from our latest (and still ongoing) study show the potential of pangenomic approaches to explore diversity within a crop, the African rice, and how its diversity has been reshaped during its domestication. We start to get a first glimpse of the role that these structural variations have played in gene composition and adaptation, and how evolutionary forces have shaped the organisation and dynamics of the (pan)genome.

We annotated 22,765 new genes across all individuals, increasing the number of genes (40,553 in the reference) by 56%. This result is slightly similar to the 19,319 new genes observed in a pangenomic analysis from 111 Asian rice long-read sequenced genomes, including wild accessions, but 3 times more than the 10,872 genes found with similar data and approach on the Asian rice, but from a smaller number of cultivated and wild accessions (Zhao et al. 2018).

Our results also showed a reduction of the dispensable genome of the cultivated rice compared to the wild relative species, from 34.8% to 30.8%, and conversely an increase of the core genome. Similar trends were observed in Tomato, from 20.98% to 18.6% (Gao et al. 2019). While it is expected that the gene pool of a wild type has a higher allelic diversity, our analysis highlighted a larger pangenome size in the wild accessions with 60,110 genes compared to 57,497 genes in *O. glaberrima*, mainly due to 2,613 additional genes in the wild accession. Similar results were reported in a super pan-genome of 214 Asian rice accessions with a greater number of genes in the pangenome of the wild species *O. rufipogon* than in the cultivated species (Shang et al. 2022).

By comparing more deeply the contents of their pangenome and of their compartments, we also explored the structure of pangenome for each species and their dynamics. Thus, we observed 3,523 genes specific to *O. barthii*, which could be explained by a loss of genes under the action of genetic drift or by founder effect during the *O glaberrima* domestication. More surprisingly, we found 910 genes specific to *O. glaberrima* which could be explained by the effect of genetic drift on the *O. barthii* gene pool. In addition, our results identified 1,601 genes switching between core and dispensable compartments of both species, including 508 core and 1,093 dispensable genes in *O. barthii* (and conversely in *O. glaberrima*). It would be interesting to have a closer look on the gene function of these switching genes as the specific genes. We thus found new switching genes with a selection signature in potential interesting pathways such as the tolerance to salt stress or plant growth for instance.

More globally, by a complementary approach, we have also identified 683 candidate genes of 7,7579 contigs identified with a signature of selection. Further analyses could be interesting to complement the current results :

- integrate RNAseq data into our analysis.
- Investigate whether SVs (gene or not) are associated with phenotypic traits
- Focus on certain genes known to be involved in domestication or selection

In addition to a significant variation in the number of genes, we also observed that structural variation overlapped 30% of the genes in the reference genome, and that ratio reached 70% if we considered genes as well as their 5kbp-flanking sequence. It would be interesting to to look more closely at whether these contigs contain TEs, annotate more precisely TEs and to search for complete copies to detect a potential role in gene regulation.

Finally, it would be interesting to characterise the pangenome and its compartments not only on the basis of genes but more broadly at the level of the entire DNA sequences.

Although not all SV were fully investigated because of the limitation of the short read technologies, our work provides a first view of the SV landscape in African rice and the impact of domestication on its pangenome. Altogether, our approaches help to reshape our thinking about the consequences of domestication on diversity, associated both to a loss of diversity and genes.

To conclude

This part of the project has allowed me to concretely start linking all the results obtained since the beginning of my PhD project (finally!), from the identification of SVs and their annotation to the genotyping of SVs at the population level. There are still analyses to be integrated but we already have got a first overview of the domestication impact on the pangenome.

It was also an opportunity to use RNAseq data produced ten years ago, that were no longer in use... In fact, we have a lot of sequencing and re-sequencing data that are valuable resources and we should not be forgotten...

V Discussion and Conclusion

14- Discussion and Outlook

Science can never solve one problem without raising ten more problems – George Bernard Shaw

The publication of the Arabidopsis and rice genome in the early 2002 marked the beginning of a new era with dramatic advances in sequencing technologies over the past 10 years. The progressive publication of high-quality reference genomes has opened a new era that greatly facilitated genetic and genomic studies in plants. Thus, genomic variability mainly based on single nucleotides (SNPs) has been extensively characterized through mapping short reads to a single reference (Atwell et al. 2010; Lai et al. 2012; Xu, Liu, et al. 2012; Zhou, Jiang, et al. 2015). Then, as seen in previous sections, an increasing number of studies have shown that structural variations, including CNVs and PAVs, plays a major role in genetic diversity. This was first observed for specific genes and pathways (Gamuyao et al. 2012; Wang, Xiong, et al. 2015; Xu, Xu, et al. 2006) and then on a genome-wide scale with short and long read data (Muñoz-Amatriaín et al. 2013; Springer et al. 2009; Swanson-Wagner et al. 2010). Although a reference genome is a valuable resource, it is very limited as it is based on a single individual and a large diversity present in a species was previously ignored. Pangenomes have been proposed as a more comprehensive way of exploring genetic diversity and capturing the full range of diversity in a group of individuals (e.g. a population, cultivars, species, or a clade).

Pangenomics coupled to high-throughput sequencing technologies has ushered in an exciting new era with a paradigm shift from one genome to pangenome, and is revealing how structural variations have shaped the genomic landscape of (pan)genome. In the last five years, plant pangenomic studies have shown that the variability at the gene level can be reflected through alleles of genes with different types of structural variations (deletion, insertion, but also through a variation of genes number with CNVs. An extreme form of CNV is the PAV, corresponding to the absence of the gene. This was illustrated by , for instance, two studies in wheat and soybean, in which 12,150 and 27,175 genes were absent from the reference genome but present in the resequenced cultivars (Liu, Du, et al. 2020; Montenegro et al. 2017). Today, the notion of reference needs to be rethought and the pangenome, which can be considered as a mixture of genomes grouping all the variations of a group, is on the way to becoming this new "reference" if the current methodological limits are overcome.

The first crop pangenomes were built from short read sequencing data. Two main approaches have been developed to build a pangenome, including assemble-then-map and map-then-assemble.The assemble-then-map approach is based on comparison of whole genomes after their *de novo* assembly (Gao et al. 2019; Gordon et al. 2017; Hu et al. 2017). The map-then-assemble approach consists of extracting unmapped reads which are then assembled *de novo* (Golicz, Bayer, et al. 2016; Hufnagel et al. 2021). In plants, pangenomics analysis has emerged as an approach that could greatly explore SVs previously hidden, and also facilitate improvements in various plants by uncovering numerous associations between SVs and key agronomic traits. Numerous examples were described including yield, fruit shape and flavor in tomato (Alonge et al. 2020; Li, He, et al. 2023), production in soybean (Liu, Du, et al. 2020), fruit maturity and shape in peach (Guo et al. 2020), heat tolerance in mil (Yan et al. 2023), fruit shape, flowering time or root growth in cucumber (Li, Wang, et al. 2022 or grain color in sorgho (Tao et al. 2021) By providing access to an unexplored reservoir of (gene)diversity within species or clades, the pangenome is of particular interest in the context of ongoing climate change to develop more resilient crops. Overall, the study of pangenomes will certainly allow to a better understanding of how evolutionary forces have shaped diversity. However, we certainly are still ill equipped in terms of population genetic analysis to

take advantage of this new diversity. In our study, we pioneered the use for example PCAadapt using coverage to recover selection signature, and the literature on pangenomes is still limited in using such population genomic tools. Moreover, in our study based on 247 african rice cultivated and wild, 22,765 new genes were annotated and we showed that domestication was associated with a loss of genes, a likely consequences of an increase of drift during the domestication process. Thus, the first results of our latest ongoing study show some opportunities offered by pangenomics to explore this diversity.

Although very promising, pangenomics still has many challenges and limitations to overcome in order to move from an emergent approach to a more classical one integrated in genetic and genomics studies. One of the first limitations is related to the construction of pangenome from short read data only. Whatever approach is used, we can assemble 100-200 bp reads into contigs relatively short and thus identify small structural variations, but cannot resolve large and complex structural variations. One way to overcome this limitation has been to focus the analysis on genes, because simple sequencess are easier to assemble. But this focus let aside non-genic variations such as cis-regulatory elements (Feschotte 2008, Figure 2.2, page 36). In tomato and rice, pangenomes constructed from short and long reads from the same accessions were compared, highlighting the limitation of the short read technologies for investigating large SVs.

In the last years, third-generation sequencing became more cost-effective, accessible and robust, and adequate computational methods were implemented. In order to exploit long reads data (and short reads also) and integrate all SVs detected, a new approach based on the pangenome graph has recently emerged in a continuous and rapid development (Garrison et al. 2018; Li, Feng, and Chu 2020; Sirén et al. 2021), using a graph representing all variations (Figure 14.1, page 128).

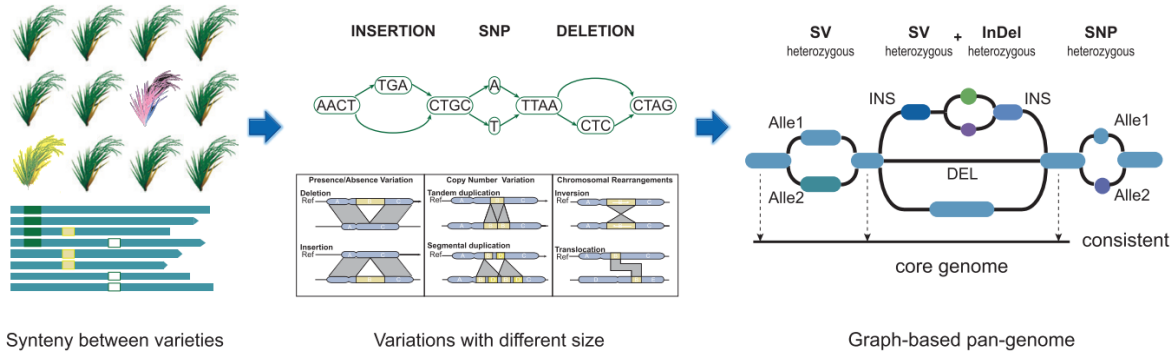

Figure 14.1: General process of graph-based pan-genome construction.
(Left) Selection of representative varieties and comparison of linear genomes between varieties. (Middle) Search for variation between varieties, and division of the variants into small structural variants and large structural variants according to the size of the variant, where the length of small structural variants was <50 bp, including SNPs (single bases), insertions and deletions, and the length of large structural variants was >50 bp, including presence/absence variations (insertions and deletions), copy number variants (tandem duplications and segmental duplications), and chromosomal rearrangements (inversions and translocations). (Right) Construction of a graph-based pan- genome based on variation information to graphically display the genome and variations between varieties. Alle, allele; DEL, deletion; INS, insertion; InDel, insertion/deletion; SNP, single nucleotide polymorphism; SV, structural variation. Figure from Wang, Qian, et al. 2023.

Pangenome graphs can be constructed mainly by two types of approaches. The first is based on a reference genome and a catalog of previously identified structural variations (in a vcf file). It is suitable for identifying structural variations, including SNPs, INDELs and large SVs (>50bp), but does not detect more complex or nested SVs. For structural variation detection, the first step is either to compare whole genomes or to align reads on a genome with tools such as MUMMER (Marçais et al. 2018) or minimap2 (Li 2018). Then, the alignments are analysed by tools such as the Syri pipeline (Goel et al. 2019) or Assemblytics (Nattestad and Schatz 2016) to extract information about the SVs (vcf file). The graph can

then be constructed from this list of variants and the reference genome with software like vg (Garrison et al. 2018). Minigraph (Li, Feng, and Chu 2020) used an alternative approach by aligning all assemblies to a reference to incrementally build a graph-based pangenome. It uses contigs with a minimum length of 100 kb and identifies SVs ranging from 20kb to 100kb but is not set up to identify SNPs. The cactus-minigraph has integrated the two approaches used by minigraph and vg to incorporate both SNPs and SVs.

Combined with third generation high-throughput sequencing technologies, pangenome graphs are very promising, even if they are still in their infancy. In addition, such non-linear representation raise new and naive questions for a non-expert of graph (which I am), such as :

- in the case of the minigraph method, does the order of alignment of the genomes have an impact and how to define the order of the alignments ?
- the graph seems to be able to group all the structural variations but the software currently has different resolutions to detect SV. Can we integrate all the variations in the same graph or in different graphs depending on the analysis ?
- How can the quality of a graph be assessed ?

Although more and more pangenome graphs are being created (Li, Liu, et al. 2022; Liu, Van Eck, et al. 2002; Qin et al. 2021; Zhou, Zhang, et al. 2022), they will have to be tested on larger and higher complex genomes. All these questions will no longer be relevant when the different tools to build a graph-based pangenome would have been tested on several models, these analyses benchmarked, standardised and best practices generalised after being tested on a large scale.

Once the pangenome graph has been constructed with a good accuracy, the next step is to use this graph to, for instance:

- define what is core or dispensable in the population or what, in our case, is specific to cultivated or wild accessions;
- to perform population-scale SVs genotyping i.e. tagged SV variations in individual (not having long read sequencing data) sequenced with short read sequencing

Thus, downstream analysis requires the development of new robust and efficient algorithms (and tools) to, for example, align short reads against a pangenome graph. These softwares are under development and are progressively available, like Giraffe software which were used to genotype SVs in 5,202 human genomes sequenced in short reads (Figure 14.2, page 130).

Other interesting challenges will be related to the annotation of these (pan)genomes, whether in transposable element or in gene, and more broadly to integrate and link all other biological information available (RNAseq, methylome, phenotype, epigenome) to the pangenome.

More than 20 years after the sequencing of the first plant genome, pangenomics opens a new and very promising era with many challenges to face. It will also lead us to rethink our knowledge of evolutionary processes such as domestication by understanding for instance, why genes have been potentially lost, what they code for and what impact this variability has on gene regulation. More broadly, it can be an opportunity to push scientists to think outside the box to deepen and rethink all the knowledge acquired.

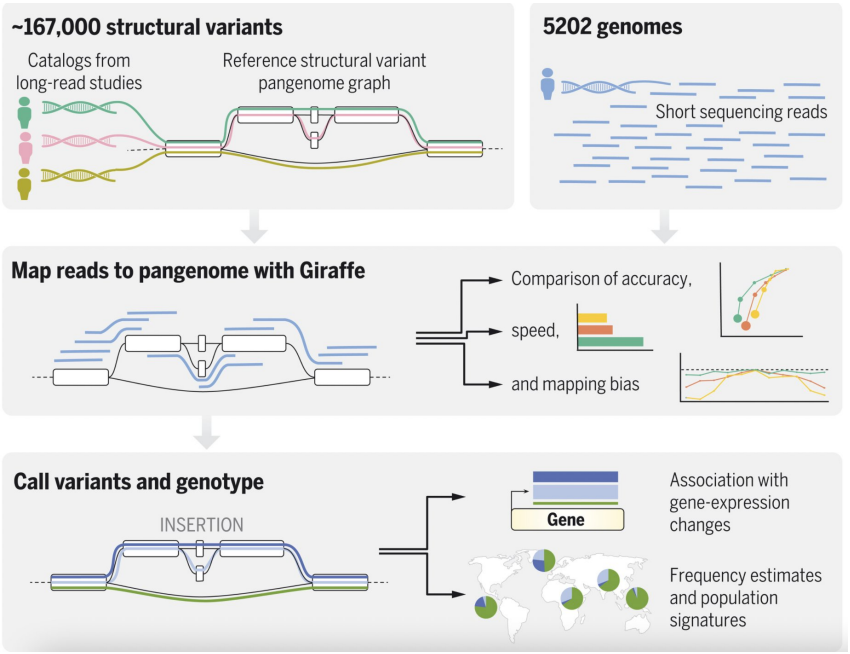

Figure 14.2: Overview of the experiments.
Variant calls from long read–based and large-scale sequencing studies were used to construct pangenome reference graphs (top). Giraffe (and competing mappers) mapped reads to the graph or to linear references, and mapping accuracy, allele coverage balance, and speed were evaluated (middle). Then, mapped reads were used for variant calling, and variant call accuracy was evaluated (bottom). Structural variant calls were analyzed alongside expression data to identify eQTLs and population frequency estimates. From Sirén et al. 2021.

Bibliography

Shahal Abbo, Simcha Lev-Yadun, and Avi Gopher. "Agricultural Origins: Centers and Noncenters; A Near Eastern Reappraisal". In: *Critical Reviews in Plant Sciences* 29.5 (2010), pages 317–328. DOI: 10.1080/07352689.2010.502823. URL: https://doi.org/10.1080/07352689.2010.502823 (cited on page 28).

Karine Alix et al. "Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants". In: *Annals of Botany* 120.2 (2017), pages 183–194. ISSN: 0305-7364. DOI: 10.1093/AOB/MCX079. URL: https://academic.oup.com/aob/article/120/2/183/3959620 (cited on page 33).

Robin Allaby. "Integrating the processes in the evolutionary system of domestication". In: *Journal of Experimental Botany* 61.4 (2010), pages 935–944. ISSN: 0022-0957. DOI: 10.1093/jxb/erp382. eprint: https://academic.oup.com/jxb/article-pdf/61/4/935/1445402/erp382.pdf. URL: https://doi.org/10.1093/jxb/erp382 (cited on page 29).

Michael Alonge et al. "Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato". In: *Cell* 182.1 (2020), 145–161.e23. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2020.05.021 (cited on page 127).

C Andrews. "The Hardy-Weinberg Principle". In: *Nature Education Knowledge* 10.3 (2010), page 65. URL: https://www.nature.com/scitable/knowledge/library/the-hardy-weinberg-principle-13235724/ (cited on page 34).

Rudi Appels et al. "Shifting the limits in wheat research and breeding using a fully annotated reference genome". In: *Science* 361.6403 (2018), eaar7191. DOI: 10.1126/science.aar7191. URL: https://www.science.org/doi/abs/10.1126/science.aar7191 (cited on page 34).

Motoyuki Ashikari et al. "Cytokinin Oxidase Regulates Rice Grain Production". In: *Science* 309.5735 (2005), pages 741–745. DOI: 10.1126/science.1113373. URL: https://doi.org/10.1126/science.1113373 (cited on page 32).

Naveenkumar Athiyannan et al. "Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning". In: *Nature Genetics 2022 54:3* 54.3 (2022), pages 227–231. ISSN: 1546-1718. DOI: 10.1038/s41588-022-01022-1. URL: https://www.nature.com/articles/s41588-022-01022-1 (cited on page 87).

Susanna Atwell et al. "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines". In: *Nature 2010 465:7298* 465.7298 (2010), pages 627–631. ISSN: 1476-4687. DOI: 10.1038/nature08800. URL: https://www.nature.com/articles/nature08800 (cited on pages 4, 39, 127).

Lorenzo Barchi et al. "Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding". In: *The Plant Journal* 107.2 (2021), pages 579–596. ISSN: 1365-313X. DOI: 10.1111/TPJ.15313. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.15313 (cited on page 43).

André Beló et al. "Allelic genome structural variations in maize detected by array comparative genome hybridization". In: *Theoretical and Applied Genetics* 120.2 (2010), pages 355–367. ISSN: 1432-2242. DOI: 10.1007/s00122-009-1128-9. URL: https://doi.org/10.1007/s00122-009-1128-9 (cited on page 38).

J. D. Berger et al. “Domestication bottlenecks limit genetic diversity and constrain adaptation in narrow-leafed lupin (Lupinus angustifolius L.)” In: *Theoretical and Applied Genetics* 124.4 (2012), pages 637–652. ISSN: 00405752. DOI: 10.1007/S00122-011-1736-Z/TABLES/5. URL: https://link.springer.com/article/10.1007/s00122-011-1736-z (cited on pages 3, 29, 51, 52).

Eugenio Butelli et al. “Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges ”. In: *The Plant Cell* 24.3 (2012), pages 1242–1255. ISSN: 1040-4651. DOI: 10.1105/tpc.111.095232. URL: https://doi.org/10.1105/tpc.111.095232 (cited on page 34).

Cheng Chang et al. “Copy Number Variation of Cytokinin Oxidase Gene Tackx4 Associated with Grain Weight and Chlorophyll Content of Flag Leaf in Common Wheat”. In: *PLOS ONE* 10.12 (2015), e0145970. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0145970. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0145970 (cited on page 37).

Haoyu Cheng et al. “Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm”. In: *Nature Methods* 18.2 (2021), pages 170–175. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01056-5. URL: https://doi.org/10.1038/s41592-020-01056-5 (cited on page 87).

Bruno Contreras-Moreira et al. “Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species”. In: *Frontiers in Plant Science* 8 (2017). ISSN: 1664-462X. DOI: 10.3389/fpls.2017.00184. URL: https://www.frontiersin.org/articles/10.3389/fpls.2017.00184 (cited on pages 5, 43, 51).

Philippe Cubry et al. “The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes”. In: *Current Biology* (2018). ISSN: 09609822. DOI: 10.1016/j.cub.2018.05.066 (cited on pages 8, 10, 29, 31, 45, 88, 107).

Charles Darwin. “On the Origin of Species by Means of Natural Selection”. In: (1859) (cited on pages 29, 33).

Charles Darwin. “The Variation of Animals and Plants Under Domestication”. In: (1868) (cited on pages 27, 29).

Aurora Díaz et al. “Copy Number Variation Affecting the Photoperiod-B1 and Vernalization-A1 Genes Is Associated with Altered Flowering Time in Wheat (Triticum aestivum)”. In: *PLoS ONE* 7.3 (2012). Edited by Samuel P. Hazen, e33234. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0033234. URL: http://dx.plos.org/10.1371/journal.pone.0033234 (cited on pages 37, 38).

Wei Ding, Franz Baumdicker, and Richard A. Neher. “panX: pan-genome analysis and exploration”. In: *Nucleic Acids Research* 46.1 (2018), e5. DOI: 10.1093/NAR/GKX977. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5758898/ (cited on pages 8, 87).

Rodolfo Dirzo and Peter H Raven. “Global State of Biodiversity and Loss”. In: *Annual Review of Environment and Resources* 28.1 (2003), pages 137–167. DOI: 10.1146/annurev.energy.28.050302.105532. URL: https://doi.org/10.1146/annurev.energy.28.050302.105532 (cited on pages 3, 26).

John F. Doebley, Brandon S. Gaut, and Bruce D. Smith. “The molecular genetics of crop domestication.” In: *Cell* 127.7 (2006), pages 1309–21. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.12.006. URL: http://www.cell.com/article/S0092867406015923/fulltext (cited on pages 3, 25, 27–29, 31, 32).

Zhongqu Duan et al. “HUPAN: a pan-genome analysis pipeline for human genomes”. In: *Genome Biology 2019 20:1* 20.1 (2019), pages 1–11. ISSN: 1474-760X. DOI: 10.1186/S13059-019-1751-Y. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1751-y (cited on page 101).

Éloi Durant et al. “Panache: a web browser-based viewer for linearized pangenomes”. In: *Bioinformatics* 37.23 (2021), pages 4556–4558. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab688. URL: https://doi.org/10.1093/bioinformatics/btab688 (cited on page 101).

Adam Eyre-Walker et al. “Investigation of the bottleneck leading to the domestication of maize”. In: *Proceedings of the National Academy of Sciences of the United States of America* 95.8 (1998), pages 4441–4446. ISSN: 00278424. DOI: 10.1073/PNAS.95.8.4441. URL: https://www.pnas.org/doi/abs/10.1073/pnas.95.8.4441 (cited on pages 3, 29).

Cédric Feschotte. “Transposable elements and the evolution of regulatory networks”. In: *Nature Reviews Genetics* 9.5 (2008), pages 397–405. ISSN: 1471-0064. DOI: 10.1038/nrg2337. URL: https://doi.org/10.1038/nrg2337 (cited on pages 34, 36, 53, 128).

Anne Frary et al. “fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size”. In: *Science* 289.5476 (2000), pages 85–88. DOI: 10.1126/science.289.5476.85. URL: https://doi.org/10.1126/science.289.5476.85 (cited on page 32).

Dorian Q. Fuller et al. “Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.17 (2014), pages 6147–6152. ISSN: 10916490. DOI: 10.1073/pnas.1308937110 (cited on pages 3, 25, 26, 29).

Iulian Gabur et al. “Connecting genome structural variation with complex traits in crop plants”. In: *Theoretical and Applied Genetics 2018 132:3* 132.3 (2018), pages 733–750. ISSN: 1432-2242. DOI: 10.1007/S00122-018-3233-0. URL: https://link.springer.com/article/10.1007/s00122-018-3233-0 (cited on pages 4, 33, 38).

Rico Gamuyao et al. “The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency”. In: *Nature* 488.7412 (2012), pages 535–539. ISSN: 1476-4687. DOI: 10.1038/nature11346. URL: https://doi.org/10.1038/nature11346 (cited on pages 83, 127).

Lei Gao et al. “The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor”. In: *Nature Genetics* 51.6 (2019), pages 1044–1051. ISSN: 1061-4036. DOI: 10.1038/s41588-019-0410-2. URL: http://www.nature.com/articles/s41588-019-0410-2 (cited on pages 4, 40, 43, 52, 83, 87, 123, 127).

Erik Garrison et al. “Variation graph toolkit improves read mapping by representing genetic variation in the reference”. In: *Nat. Biotechnol.* 36.9 (2018), pages 875–879. ISSN: 15461696. DOI: 10.1038/nbt.4227 (cited on pages 12, 128, 129).

Manish Goel et al. “SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies”. In: *Genome Biology* 20.1 (2019), page 277. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1911-0. URL: https://doi.org/10.1186/s13059-019-1911-0 (cited on page 128).

Agnieszka A. Golicz, Jacqueline Batley, and David Edwards. “Towards plant pangenomics”. In: *Plant Biotechnology Journal* 14.4 (2016), pages 1099–1105. ISSN: 14677644. DOI: 10.1111/pbi.12499. URL: http://doi.wiley.com/10.1111/pbi.12499 (cited on pages 5, 51).

Agnieszka A. Golicz, Philipp E. Bayer, et al. “The pangenome of an agronomically important crop plant Brassica oleracea”. In: *Nature Communications* 7 (2016), page 13390. ISSN: 2041-1723. DOI: 10.1038/ncomms13390. URL: http://www.nature.com/doifinder/10.1038/ncomms13390 (cited on pages 5, 7, 41, 43, 51, 52, 54, 87, 101, 127).

Enrique Gonzalez et al. “The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility”. In: *Science* 307.5714 (2005), pages 1434–1440. DOI: 10.1126/science.1101160. URL: https://www.science.org/doi/abs/10.1126/science.1101160 (cited on page 36).

Sean P. Gordon et al. “Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure”. In: *Nature Communications* 8.1 (2017), page 2184. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02292-8. URL: http://www.nature.com/articles/s41467-017-02292-8 (cited on pages 4, 5, 7, 39, 41–43, 51–54, 87, 101, 127).

Dariusz Grzebelus. “The Functional Impact of Transposable Elements on the Diversity of Plant Genomes”. In: *Diversity* 10.2 (2018). ISSN: 1424-2818. DOI: 10.3390/d10020018. URL: https://www.mdpi.com/1424-2818/10/2/18 (cited on pages 33, 34).

Jian Guo et al. “An integrated peach genome structural variation map uncovers genes associated with fruit traits”. In: *Genome Biology* 21.1 (2020), pages 1–19. ISSN: 1474760X. DOI: 10.1186/S13059-020-02169-Y/FIGURES/5. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02169-y (cited on page 127).

Karl Hammer. “Das domestikationssyndrom”. In: *Kulturpflanze* 32 (1984), pages 11–34 (cited on page 27).

P. J. Hastings et al. “Mechanisms of change in gene copy number”. In: *Nature Reviews Genetics 2009 10:8* 10.8 (2009), pages 551–564. ISSN: 1471-0064. DOI: 10.1038/nrg2593. URL: https://www.nature.com/articles/nrg2593 (cited on page 147).

Yoko Hattori et al. “The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water”. In: *Nature* 460.7258 (2009), pages 1026–1030. ISSN: 0028-0836. DOI: 10.1038/nature08258. URL: http://www.nature.com/articles/nature08258 (cited on pages 5, 52).

William J Haun et al. “The Composition and Origins of Genomic Variation among Individuals of the Soybean Reference Cultivar Williams 82 ”. In: *Plant Physiology* 155.2 (2010), pages 645–655. ISSN: 0032-0889. DOI: 10.1104/pp.110.166736. URL: https://doi.org/10.1104/pp.110.166736 (cited on page 38).

Gordon c. Hillman and M. Stuart Davies. “6. Domestication rates in wild-type wheats and barley under primitive cultivation”. In: *Biological Journal of the Linnean Society* 39.1 (2008), pages 39–78. ISSN: 0024-4066. DOI: 10.1111/j.1095-8312.1990.tb01611.x. eprint: https://academic.oup.com/biolinnean/article-pdf/39/1/39/14071744/j.1095-8312.1990.tb01611.x.pdf. URL: https://doi.org/10.1111/j.1095-8312.1990.tb01611.x (cited on page 28).

Zhiqiang Hu et al. “EUPAN enables pan-genome studies of a large number of eukaryotic genomes”. In: *Bioinformatics* 33.15 (2017), pages 2408–2409. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTX170. URL: https://academic.oup.com/bioinformatics/article/33/15/2408/3091809 (cited on pages 7, 8, 41, 54, 87, 127).

Xuehui Huang et al. “A map of rice genome variation reveals the origin of cultivated rice”. In: *Nature* 490.7421 (2012), pages 497–501. ISSN: 1476-4687. DOI: 10.1038/nature11532. URL: https://doi.org/10.1038/nature11532 (cited on pages 10, 105).

Sariel Hübner et al. “Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance”. In: *Nature Plants* 5.1 (2019), pages 54–62. ISSN: 20550278. DOI: 10.1038/s41477-018-0329-0. URL: http://www.nature.com/articles/s41477-018-0329-0 (cited on pages 41–43).

Matthew B Hufford, Xun Xu, et al. “Comparative population genomics of maize domestication and improvement”. In: *Nature Genetics* 44.7 (2012), pages 808–811. ISSN: 1546-1718. DOI: 10.1038/ng.2309. URL: https://doi.org/10.1038/ng.2309 (cited on pages 10, 105).

Matthew B. Hufford, Arun S. Seetharam, et al. “De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes”. In: *Science* 373.6555 (2021), pages 655–662. ISSN: 10959203. DOI: 10.1126/SCIENCE.ABG5289. URL: https://www.science.org/doi/10.1126/science.abg5289 (cited on page 44).

Bárbara Hufnagel et al. “Pangenome of white lupin provides insights into the diversity of the species”. In: *Plant Biotechnology Journal* 19.12 (2021), pages 2532–2543. ISSN: 1467-7652. DOI: 10.1111/PBI.13678. URL: https://doi.org/10.1111/pbi.13678 (cited on pages 43, 52, 87, 127).

Bhavna Hurgobin et al. “Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus”. In: *Plant Biotechnology Journal* 16.7 (2018), pages 1265–1274. ISSN: 14677644. DOI: 10.1111/pbi.12867. URL: http://doi.wiley.com/10.1111/pbi.12867 (cited on page 43).

David L. Hyten et al. “Impact of genetic bottlenecks on soybean genome diversity”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.45 (2006), pages 16666–16671. ISSN: 00278424. DOI: 10.1073/PNAS.0604379103/SUPPL_FILE/04379DATASET3.XLS. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0604379103 (cited on pages 3, 29).

Enrique Ibarra-Laclette et al. “Architecture and evolution of a minute plant genome”. In: *Nature* 498.7452 (2013), pages 94–98. ISSN: 1476-4687. DOI: 10.1038/nature12132. URL: https://doi.org/10.1038/nature12132 (cited on page 34).

The Arabidopsis Genome Initiative. “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. In: *Nature* 408.6814 (2000), pages 796–815. ISSN: 1476-4687. DOI: 10.1038/35048692. URL: https://doi.org/10.1038/35048692 (cited on pages 4, 39).

Diwash Jangam, Cédric Feschotte, and Esther Betrán. “Transposable Element Domestication As an Adaptation to Evolutionary Conflicts”. In: *Trends in Genetics* 33.11 (2017), pages 817–831. ISSN: 0168-9525. DOI: https://doi.org/10.1016/j.tig.2017.07.011. URL: https://www.sciencedirect.com/science/article/pii/S0168952517301282 (cited on page 34).

Yinping Jiao, Hainan Zhao, et al. “Genome-wide genetic changes during modern breeding of maize.” In: *Nature genetics* 44.7 (2012), pages 812–5. ISSN: 1546-1718. DOI: 10.1038/ng.2312. URL: http://www.nature.com/ng/journal/v44/n7/full/ng.2312.html#access (cited on page 38).

Yuannian Jiao, Norman J Wickett, et al. “Ancestral polyploidy in seed plants and angiosperms”. In: *Nature* 473.7345 (2011), pages 97–100. ISSN: 1476-4687. DOI: 10.1038/nature09916. URL: https://doi.org/10.1038/nature09916 (cited on page 33).

Aamir W. Khan et al. “Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement”. In: *Trends in Plant Science* (2019). ISSN: 13601385. DOI: 10.1016/j.tplants.2019.10.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S136013851930281X (cited on pages 10, 105).

Shozo Kobayashi, Nami Goto-Yamamoto, and Hirohiko Hirochika. “Retrotransposon-induced mutations in grape skin color.” In: *Science (New York, N.Y.)* 304.5673 (2004), page 982. ISSN: 1095-9203 (Electronic). DOI: 10.1126/science.1095011 (cited on page 34).

Saeko Konishi et al. “An SNP Caused Loss of Seed Shattering During Rice Domestication”. In: *Science* 312.5778 (2006), pages 1392–1396. DOI: 10.1126/science.1126410. URL: https://doi.org/10.1126/science.1126410 (cited on page 32).

Kaitao Lai et al. “Single nucleotide polymorphism discovery from wheat next-generation sequence data”. In: *Plant Biotechnology Journal* 10.6 (2012), pages 743–749. DOI: https://doi.org/10.1111/j.1467-7652.2012.00718.x (cited on pages 4, 39, 127).

Chad Laing et al. “Pan-genome sequence analysis using Panseq: An online tool for the rapid analysis of core and accessory genomic regions”. In: *BMC Bioinformatics* 11.1 (2010), pages 1–14. ISSN: 14712105. DOI: 10.1186/1471-2105-11-461/TABLES/3. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-461 (cited on pages 8, 87).

Greger Larson et al. “Current perspectives and the future of domestication studies”. In: 111.17 (2014), pages 6139–6146. ISSN: 10916490. DOI: 10.1073/pnas.1323964111 (cited on pages 25, 30).

Jennifer A. Lee, Claudia M.B. Carvalho, and James R. Lupski. “A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders”. In: *Cell* 131.7 (2007), pages 1235–1247. ISSN: 00928674. DOI: 10.1016/j.cell.2007.11.037. URL: http://www.cell.com/article/S0092867407015413/fulltext (cited on page 147).

Tong Geon Lee, Indrajit Kumar, et al. “Evolution and selection of Rhg1, a copy-number variant nematode-resistance locus”. In: *Molecular Ecology* 24.8 (2015), pages 1774–1791. DOI: https://doi.org/10.1111/mec.13138. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13138 (cited on page 38).

Changbao Li, Ailing Zhou, and Tao Sang. “Genetic analysis of rice domestication syndrome with the wild annual species, Oryza nivara”. In: *New Phytologist* 170.1 (2006), pages 185–194. ISSN: 0028-646X. DOI: https://doi.org/10.1111/j.1469-8137.2005.01647.x. URL: https://doi.org/10.1111/j.1469-8137.2005.01647.x (cited on page 32).

Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18 (2018), pages 3094–3100. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTY191. arXiv: 1708.01492. URL: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778 (cited on page 128).

Heng Li, Xiaowen Feng, and Chong Chu. “The design and construction of reference pangenome graphs with minigraph”. In: *Genome Biology* 21.1 (2020). ISSN: 1474760X. DOI: 10.1186/S13059-020-02168-Z (cited on pages 12, 128, 129).

Hongbo Li, Shenhao Wang, et al. “Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber”. In: *Nature Communications 2022 13:1* 13.1 (2022), pages 1–14. ISSN: 2041-1723. DOI: 10.1038/s41467-022-28362-0. URL: https://www.nature.com/articles/s41467-022-28362-0 (cited on pages 43, 127).

Ning Li, Qiang He, et al. “Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species”. In: *Nature Genetics* (2023). ISSN: 1546-1718. DOI: 10.1038/s41588-023-01340-y. URL: https://doi.org/10.1038/s41588-023-01340-y (cited on page 127).

Wei Li, Jianan Liu, et al. “Plant pan-genomics: recent advances, new challenges, and roads ahead”. In: *Journal of Genetics and Genomics* 49.9 (2022), pages 833–846. ISSN: 18735533. DOI: 10.1016/J.JGG.2022.06.004 (cited on pages 6, 51, 129).

Ying-huib Li, Guangyu Zhou, Jianxin Ma, et al. “De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits”. In: *Nature Biotechnology* 32.10 (2014), pages 1045–1052. ISSN: 1087-0156. DOI: 10.1038/nbt.2979 (cited on pages 5, 7, 43, 51, 53, 54).

Yiyuan Li, Jianhui Xiao, et al. “A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation”. In: *New Phytologist* 196.1 (2012), pages 282–291. ISSN: 0028-646X. DOI: https://doi.org/10.1111/j.1469-8137.2012.04243.x. URL: https://doi.org/10.1111/j.1469-8137.2012.04243.x (cited on page 38).

Tao Lin, Guangtao Zhu, et al. “Genomic analyses provide insights into the history of tomato breeding”. In: *Nature Genetics 2014 46:11* 46.11 (2014), pages 1220–1226. ISSN: 1546-1718. DOI: 10.1038/ng.3117. URL: https://www.nature.com/articles/ng.3117 (cited on pages 10, 105).

Zhongwei Lin, Xianran Li, et al. “Parallel domestication of the Shattering1 genes in cereals”. In: *Nature Genetics 2012 44:6* 44.6 (2012), pages 720–724. ISSN: 1546-1718. DOI: 10.1038/ng.2281. URL: https://www.nature.com/articles/ng.2281 (cited on page 39).

Olga F Linares. “African rice (Oryza glaberrima): History and future potential”. In: *Proceedings of the National Academy of Sciences* 99.25 (2002), pages 16360–16365. DOI: 10.1073/pnas.252604599. URL: https://www.pnas.org/doi/abs/10.1073/pnas.252604599 (cited on page 46).

Jiping Liu, Joyce Van Eck, et al. “A new class of regulatory genes underlying the cause of pear-shaped tomato fruit”. In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pages 13302–13306. ISSN: 00278424. DOI: 10.1073/PNAS.162485999. URL: https://www.pnas.org/doi/abs/10.1073/pnas.162485999 (cited on pages 32, 129).

Yucheng Liu, Huilong Du, et al. “Pan-genome of wild and cultivated soybeans”. In: *Cell* 182.1 (2020), 162–176.e13. ISSN: 10974172. DOI: 10.1016/j.cell.2020.05.023 (cited on pages 43, 127).

James R. Lupski. “Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits”. In: *Trends in Genetics* 14.10 (1998), pages 417–422. ISSN: 01689525. DOI: 10.1016/S0168-9525(98)01555-8. URL: http://www.cell.com/article/S0168952598015558/fulltext (cited on page 147).

Emma Mace et al. “The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes”. In: *BMC Plant Biology* 14.1 (2014), page 253. ISSN: 1471-2229. DOI: 10.1186/s12870-014-0253-z. URL: https://doi.org/10.1186/s12870-014-0253-z (cited on page 38).

M Mamtani et al. “Association of copy number variation in the FCGR3B gene with risk of autoimmune diseases”. In: *Genes Immunity* 11.2 (2010), pages 155–160. ISSN: 1476-5470. DOI: 10.1038/gene.2009.71. URL: https://doi.org/10.1038/gene.2009.71 (cited on page 36).

Guillaume Marçais et al. “MUMmer4: A fast and versatile genome alignment system”. In: *PLOS Computational Biology* 14.1 (2018), pages 1–14. DOI: 10.1371/journal.pcbi.1005944. URL: https://doi.org/10.1371/journal.pcbi.1005944 (cited on page 128).

Rose A. Marks et al. “Representation and participation across 20 years of plant genome sequencing”. In: *Nature Plants 2021 7:12* 7.12 (2021), pages 1571–1578. ISSN: 2055-0278. DOI: 10.1038/s41477-021-01031-8. URL: https://www.nature.com/articles/s41477-021-01031-8 (cited on pages 4, 39, 41).

Lyza G. Maron et al. “Aluminum tolerance in maize is associated with higher MATE1 gene copy number”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.13 (2013), pages 5241–5246. ISSN: 00278424. DOI: 10.1073/PNAS.1220766110. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1220766110 (cited on pages 37, 38).

Barbara McClintock. “Mutable loci in maize.” In: *Mutable loci in maize.* (1947) (cited on page 34).

Leah K McHale et al. “Structural Variants in the Soybean Genome Localize to Clusters of Biotic Stress-Response Genes ”. In: *Plant Physiology* 159.4 (2012), pages 1295–1308. ISSN: 0032-0889. DOI: 10.1104/pp.112.194605. URL: https://doi.org/10.1104/pp.112.194605 (cited on page 38).

Duccio Medini et al. “The microbial pan-genome”. In: *Current Opinion in Genetics Development* 15.6 (2005), pages 589–594. ISSN: 0959-437X. DOI: 10.1016/J.GDE.2005.09.006. URL: https://www.sciencedirect.com/science/article/pii/S0959437X05001759 (cited on pages 5, 51, 52).

Rachel S. Meyer and Michael D. Purugganan. “Evolution of crop species: genetics of domestication and diversification”. In: *Nature Reviews Genetics* 14 (2013), page 840. URL: http://dx.doi.org/10.1038/nrg3605 (cited on page 29).

Cecile Monat, Christine Tranchant-Dubreuil, et al. “Comparison of two African rice species through a new pan-genomic approach on massive data”. In: *bioRxiv* (2018). DOI: 10.1101/245431. URL: https://www.biorxiv.org/content/early/2018/01/09/245431 (cited on page 43).

Cécile Monat, Bérengère Pera, et al. “de novo assemblies of three Oryza glaberrima accessions provide first insights about pan-genome of African rices”. In: *Genome Biology and Evolution* (2016), evw253. ISSN: 1759-6653. DOI: 10.1093/gbe/evw253. URL: http://gbe.oxfordjournals.org/lookup/doi/10.1093/gbe/evw253 (cited on pages 8, 88).

Juan D. Montenegro et al. “The pangenome of hexaploid bread wheat”. In: *The Plant Journal* 90.5 (2017), pages 1007–1013. ISSN: 09607412. DOI: 10.1111/tpj.13515. URL: http://doi.wiley.com/10.1111/tpj.13515 (cited on pages 5, 7, 42, 44, 52, 54, 83, 127).

J. Kent Moore and James E. Haber. “Cell Cycle and Genetic Requirements of Two Pathways of Nonhomologous End-Joining Repair of Double-Strand Breaks in Saccharomyces cerevisiae”. In: *Molecular and cellular biology* 16.5 (1996), pages 2164–2173. ISSN: 0270-7306. DOI: 10.1128/MCB.16.5.2164. URL: https://www.tandfonline.com/doi/abs/10.1128/MCB.16.5.2164 (cited on page 147).

Michele Morgante, Emanuele De Paoli, and Slobodanka Radovic. “Transposable elements and the plant pan-genomes”. In: *Current Opinion in Plant Biology* 10.2 (2007), pages 149–155. ISSN: 1369-5266. DOI: 10.1016/J.PBI.2007.02.001 (cited on pages 4, 39).

Ulrich G Mueller et al. “The Evolution of Agriculture in Insects”. In: *Annual Review of Ecology, Evolution, and Systematics* 36.1 (2005), pages 563–595. DOI: 10.1146/annurev.ecolsys.36.102003.152626. URL: https://doi.org/10.1146/annurev.ecolsys.36.102003.152626 (cited on pages 25, 26).

María Muñoz-Amatriaín et al. “Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome”. In: *Genome Biology* 14.6 (2013), R58. ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-6-r58. URL: https://doi.org/10.1186/gb-2013-14-6-r58 (cited on pages 4, 38, 39, 127).

Maria Nattestad and Michael C Schatz. “Assemblytics: a web analytics tool for the detection of variants from an assembly”. In: *Bioinformatics* 32.19 (2016), pages 3021–3023. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw369. URL: https://doi.org/10.1093/bioinformatics/btw369 (cited on page 128).

Hidetaka Nishida et al. “Structural variation in the 5’ upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (Triticum aestivum L.), and their effect on heading time”. In: *Molecular Breeding* 31.1 (2013), pages 27–37. ISSN: 1572-9788. DOI: 10.1007/s11032-012-9765-0. URL: https://doi.org/10.1007/s11032-012-9765-0 (cited on page 38).

Shihui Niu et al. “The Chinese pine genome and methylome unveil key features of conifer evolution”. In: *Cell* 185.1 (2022), 204–217.e14. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2021.12.006 (cited on page 87).

Petr Novák et al. “Repeat-sequence turnover shifts fundamentally in species with large genomes”. In: *Nature Plants* 6.11 (2020), pages 1325–1329. ISSN: 2055-0278. DOI: 10.1038/s41477-020-00785-x. URL: https://doi.org/10.1038/s41477-020-00785-x (cited on page 34).

S. Ohno. “So much "junk" DNA in our genome.” In: *Brookhaven symposia in biology* 23 (1972), pages 366–370. ISSN: 0068-2799 (Print) (cited on page 34).

Kenneth M. Olsen and Michael D. Purugganan. “Molecular Evidence on the Origin and Evolution of Glutinous Rice”. In: *Genetics* 162.2 (2002), pages 941–950. ISSN: 1943-2631. DOI: 10.1093/genetics/162.2.941. URL: https://doi.org/10.1093/genetics/162.2.941 (cited on page 32).

L E Orgel and F H C Crick. “Selfish DNA: the ultimate parasite”. In: *Nature* 284.5757 (1980), pages 604–607. ISSN: 1476-4687. DOI: 10.1038/284604a0. URL: https://doi.org/10.1038/284604a0 (cited on page 34).

Julie Orjuela et al. “An extensive analysis of the African rice genetic diversity through a global genotyping.” In: *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 127.10 (2014), pages 2211–23. ISSN: 1432-2242. DOI: 10.1007/s00122-014-2374-z. URL: http://www.ncbi.nlm.nih.gov/pubmed/25119871 (cited on pages 8, 88).

Lijun Ou et al. “Pan-genome of cultivated pepper (Capsicum) and its use in gene presence–absence variation analyses”. In: *New Phytologist* 220.2 (2018), pages 360–363. DOI: https://doi.org/10.1111/nph.15413. URL: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.15413 (cited on page 43).

Andrew J. Page et al. “Roary: rapid large-scale prokaryote pan genome analysis”. In: *Bioinformatics* 31.22 (2015), page 3691. ISSN: 14602059. DOI: 10.1093/BIOINFORMATICS/BTV421. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817141/ (cited on pages 8, 87).

Jakob Petereit et al. “Pangenomics and Crop Genome Adaptation in a Changing Climate”. In: *Plants* 11.15 (2022). ISSN: 2223-7747. DOI: 10.3390/plants11151949. URL: https://www.mdpi.com/2223-7747/11/15/1949 (cited on pages 10, 105).

Sara Pinosio et al. “Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation.” In: *Molecular biology and evolution* 33.10 (2016), pages 2706–19. ISSN: 1537-1719. DOI: 10.1093/molbev/msw161. URL: http://www.ncbi.nlm.nih.gov/pubmed/27499133%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5026262 (cited on page 43).

Roland Portères. “Historique sur les premiers échantillons d’Oryza glaberrima St. recueillis en Afrique”. In: *Journal d’agriculture traditionnelle et de botanique appliquée* 2.10 (1955), pages 535–537. DOI: 10.3406/jatba.1955.2257. URL: https://www.persee.fr/doc/jatba_0021-7662_1955_num_2_10_2257 (cited on page 45).

Roland Portères. “Berceaux agricoles primaires sur le continent africain”. In: *The Journal of African History* 5.2 (1962), pages 195–210. URL: https://www.jstor.org/stable/179739 (cited on page 46).

Michael D. Purugganan. “What is domestication?” In: *Trends in Ecology Evolution* 37.8 (2022), pages 663–671. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2022.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0169534722000891 (cited on pages 25, 29).

Jianjian Qi, Xin Liu, et al. “A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity”. In: *Nature Genetics* 45 (2013), page 1510. URL: http://dx.doi.org/10.1038/ng.2801 (cited on pages 10, 105).

Xinpeng Qi, Man Wah Li, et al. “Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing”. In: *Nature Communications 2014 5:1* 5.1 (2014), pages 1–11. ISSN: 2041-1723. DOI: 10.1038/ncomms5340. URL: https://www.nature.com/articles/ncomms5340 (cited on page 39).

Lunwen Qian et al. “Deletion of a Stay-Green Gene Associates with Adaptive Selection in Brassica napus”. In: *Molecular Plant* 9.12 (2016), pages 1559–1569. ISSN: 17529867. DOI: 10.1016/j.molp.2016.10.017. URL: http://www.cell.com/article/S1674205216302647/fulltext (cited on page 38).

Qin Qiao et al. “Evolutionary history and pan-genome dynamics of strawberry (Fragaria spp.)” In: *Proceedings of the National Academy of Sciences* 118.45 (2021), e2105431118. DOI: 10.1073/pnas.2105431118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2105431118 (cited on pages 5, 42, 43).

Peng Qin et al. “Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations”. In: *Cell* 184.13 (2021), 3542–3558.e16. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2021.04.046. URL: https://www.sciencedirect.com/science/article/pii/S009286742100581X (cited on pages 9, 43, 101, 129).

Lluís Quintana-Murci and Andrew G Clark. “Population genetic tools for dissecting innate immunity in humans”. In: *Nature Reviews Immunology* 13.4 (2013), pages 280–293. ISSN: 1474-1741. DOI: 10.1038/nri3421. URL: https://doi.org/10.1038/nri3421 (cited on page 37).

Habib Rijzaani et al. “The pangenome of banana highlights differences between genera and genomes”. In: *The Plant Genome* (2021), e20100. ISSN: 1940-3372. DOI: 10.1002/TPG2.20100. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/tpg2.20100 (cited on pages 41, 52).

Anne Rovelet-Lecrux et al. “APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy”. In: *Nature Genetics* 38.1 (), pages 24–26. ISSN: 1546-1718. DOI: 10.1038/ng1718. URL: https://doi.org/10.1038/ng1718 (cited on page 36).

Rachit K Saxena, David Edwards, and Rajeev K Varshney. “Structural variations in plant genomes.” In: *Briefings in functional genomics* 13.4 (2014), pages 296–307. ISSN: 2041-2657. DOI: 10.1093/bfgp/elu016. URL: https://academic.oup.com/bfg/article/13/4/296/2845986?login=false (cited on pages 33, 36).

Michael C Schatz et al. “Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa , document novel gene space of aus and indica”. In: *Genome Biology* 15.11 (2014), page 506. ISSN: 1465-6906. DOI: 10.1186/PREACCEPT-2784872521277375. URL: http://genomebiology.com/2014/15/11/506 (cited on pages 5, 7, 43, 51, 54).

Sarah Schiessl et al. “Targeted deep sequencing of flowering regulators in Brassica napus reveals extensive copy number variation”. In: *Scientific Data* 4.1 (2017), page 170013. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.13. URL: https://doi.org/10.1038/sdata.2017.13 (cited on page 38).

Ted R Schultz et al. “Reciprocal illumination: a comparison of agriculture in humans”. In: *Insect-fungal associations: ecology and evolution. Oxford University Press, New York* (2005), pages 149–190 (cited on pages 25, 26).

Lianguang Shang et al. “A super pan-genomic landscape of rice”. In: *Cell research* 32.10 (2022), pages 878–896. ISSN: 1748-7838. DOI: 10.1038/S41422-022-00685-Z. URL: https://pubmed.ncbi.nlm.nih.gov/35821092/ (cited on pages 43, 54, 123).

Rachel M. Sherman et al. “Assembly of a pan-genome from deep sequencing of 910 humans of African descent”. In: 51.1 (2019), pages 30–35. ISSN: 15461718. DOI: 10.1038/s41588-018-0273-y (cited on pages 41, 87).

Alisa-Naomi Sieber et al. “Copy number variation of CBF-A14 at the Fr-A2 locus determines frost tolerance in winter durum wheat”. In: *Theoretical and Applied Genetics* 129.6 (2016), pages 1087–1097. ISSN: 1432-2242. DOI: 10.1007/s00122-016-2685-3. URL: https://doi.org/10.1007/s00122-016-2685-3 (cited on page 38).

Kristin J Simons et al. “Molecular Characterization of the Major Wheat Domestication Gene Q”. In: *Genetics* 172.1 (2006), pages 547–555. ISSN: 1943-2631. DOI: 10.1534/genetics.105.044727. URL: https://doi.org/10.1534/genetics.105.044727 (cited on page 32).

A B Singleton et al. “$\alpha-$Synuclein Locus Triplication Causes Parkinson’s Disease”. In: *Science* 302.5646 (2003), page 841. DOI: 10.1126/science.1090278. URL: https://www.science.org/doi/abs/10.1126/science.1090278 (cited on page 36).

Jouni Sirén et al. “Pangenomics enables genotyping of known structural variants in 5202 diverse genomes”. In: *Science* 374.6574 (2021). ISSN: 10959203. DOI: 10.1126/SCIENCE.ABG8871 (cited on pages 12, 128, 130).

Jia Ming Song et al. “Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus”. In: *Nature Plants* 6.1 (2020), pages 34–45. ISSN: 20550278. DOI: 10.1038/S41477-019-0577-7 (cited on pages 5, 42, 43).

Nathan M Springer et al. “Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content.” In: *PLoS genetics* 5.11 (2009), e1000734. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000734. URL: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000734 (cited on pages 4, 36, 38, 39, 51, 52, 127).

Anna Stein, Olivier Coriton, et al. “Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in Brassica napus”. In: *Plant Biotechnology Journal* 15.11 (2017), pages 1478–1489. DOI: https://doi.org/10.1111/pbi.12732. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pbi.12732 (cited on page 38).

Joshua C. Stein, Yeisoo Yu, et al. “Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza”. In: *Nature Genetics* (2018), page 1. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0040-0. URL: http://www.nature.com/articles/s41588-018-0040-0 (cited on page 45).

Markus G Stetter et al. “How to make a domesticate”. In: *Current Biology* 27.17 (2017), R896–R900. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2017.06.048. URL: https://www.sciencedirect.com/science/article/pii/S0960982217307856 (cited on pages 26, 29).

Kristian A. Stevens et al. “Sequence of the sugar pine megagenome”. In: *Genetics* 204.4 (2016), pages 1613–1626. ISSN: 19432631. DOI: 10.1534/GENETICS.116.193227. URL: https://academic.oup.com/genetics/article/204/4/1613/6046811 (cited on pages 4, 39).

Anthony Studer et al. “Identification of a functional transposon insertion in the maize domestication gene tb1”. In: *Nature Genetics* 43.11 (2011), pages 1160–1163. ISSN: 1546-1718. DOI: 10.1038/ng.942. URL: https://doi.org/10.1038/ng.942 (cited on page 34).

Hequan Sun, Wen Biao Jiao, et al. “Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar”. In: *Nature Genetics 2022 54:3* 54.3 (2022), pages 342–348. ISSN: 1546-1718. DOI: 10.1038/s41588-022-01015-0. URL: https://www.nature.com/articles/s41588-022-01015-0 (cited on page 87).

Yanqing Sun, Lianguang Shang, et al. “Twenty years of plant genome sequencing: achievements and challenges”. In: *Trends in Plant Science* 27.4 (2022), pages 391–401. ISSN: 1360-1385. DOI: 10.1016/J.TPLANTS.2021.10.006. URL: http://www.cell.com/article/S1360138521002818/fulltext (cited on pages 4, 39, 40).

Tim Sutton et al. “Boron-Toxicity Tolerance in Barley Arising from Efflux Transporter Amplification”. In: *Science* 318.5855 (2007), pages 1446–1449. DOI: 10.1126/science.1146853. URL: https://www.science.org/doi/abs/10.1126/science.1146853 (cited on page 38).

Ruth A Swanson-Wagner et al. “Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor.” In: *Genome research* 20.12 (2010), pages 1689–99. ISSN: 1549-5469. DOI: 10.1101/gr.109165.110. URL: http://genome.cshlp.org/content/20/12/1689.long (cited on pages 4, 38, 39, 51, 52, 127).

Megan T Sweeney et al. “Caught Red-Handed: Rc Encodes a Basic Helix-Loop-Helix Protein Conditioning Red Pericarp in Rice”. In: *The Plant Cell* 18.2 (2006), pages 283–294. ISSN: 1040-4651. DOI: 10.1105/tpc.105.038430. URL: https://doi.org/10.1105/tpc.105.038430 (cited on page 32).

Ken-ichi Tanno and George Willcox. “How Fast Was Wild Wheat Domesticated?” In: *Science* 311.5769 (2006), page 1886. DOI: 10.1126/science.1124635. URL: https://www.science.org/doi/abs/10.1126/science.1124635 (cited on page 29).

Ken-ichi Tanno and George Willcox. “Distinguishing wild and domestic wheat and barley spikelets from early Holocene sites in the Near East”. In: *Vegetation History and Archaeobotany* 21.2 (2012), pages 107–115. ISSN: 1617-6278. DOI: 10.1007/s00334-011-0316-0. URL: https://doi.org/10.1007/s00334-011-0316-0 (cited on page 29).

Yongfu Tao et al. “Extensive variation within the pan-genome of cultivated and wild sorghum”. In: *Nature Plants* 7.6 (2021), pages 766–773. ISSN: 2055-0278. DOI: 10.1038/s41477-021-00925-x. URL: https://www.nature.com/articles/s41477-021-00925-x (cited on pages 43, 127).

Hervé Tettelin et al. “Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome””. In: *Proceedings of the National Academy of Sciences* 102.39 (2005), pages 13950–13955. ISSN: 0027-8424. DOI: 10.1073/PNAS.0506758102. URL: https://www.pnas.org/content/102/39/13950 (cited on pages 4, 5, 39, 51–53).

Davoud Torkamaneh, Marc-André Lemay, and François Belzile. “The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content”. In: *Plant Biotechnology Journal* 19.9 (2021), pages 1852–1862. DOI: https://doi.org/10.1111/pbi.13600. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pbi.13600 (cited on pages 42, 44, 87).

Christine Tranchant-Dubreuil et al. “FrangiPANe, a tool for creating a panreference using left behind reads”. In: *NAR Genomics and Bioinformatics* 5.1 (2023). ISSN: 2631-9268. DOI: 10.1093/nargab/lqad013. URL: https://doi.org/10.1093/nargab/lqad013 (cited on pages 8–10, 88–90, 105, 149–153).

Christine Tranchant-Dubreuil, Mathieu Rouard, and Francois Sabot. “Plant Pangenome: Impacts on Phenotypes and Evolution”. In: (2019), pages 453–478. DOI: 10.1002/9781119312994.apr0664. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119312994.apr0664 (cited on pages 7, 8, 53, 55).

Yusaku Uga et al. “Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions”. In: *Nature Genetics* 45.9 (2013), pages 1097–1102. ISSN: 1546-1718. DOI: 10.1038/ng.2725. URL: https://doi.org/10.1038/ng.2725 (cited on page 38).

Yves Van De Peer, Steven Maere, and Axel Meyer. “The evolutionary significance of ancient genome duplications”. In: *Nature Reviews Genetics 2009 10:10* 10.10 (2009), pages 725–732. ISSN: 1471-0064. DOI: 10.1038/nrg2600. URL: https://www.nature.com/articles/nrg2600 (cited on page 33).

Duncan A Vaughan, H Morishima, and K Kadowaki. “Diversity in the Oryza genus”. In: *Current Opinion in Plant Biology* 6.2 (2003), pages 139–146. ISSN: 1369-5266. DOI: 10.1016/S1369-5266(03)00009-8. URL: https://www.sciencedirect.com/science/article/pii/S1369526603000098 (cited on pages 45, 149).

George Vernikos et al. “Ten years of pan-genome analyses”. In: *Current Opinion in Microbiology* 23 (2015), pages 148–154. ISSN: 1369-5274. DOI: https://doi.org/10.1016/j.mib.2014.11.016. URL: https://www.sciencedirect.com/science/article/pii/S1369527414001830 (cited on pages 5, 51).

Patricio Mendez del Villar and Jean-Martin Bauer. “Le riz en Afrique de l’Ouest : dynamiques, politiques et perspectives”. In: *Cahiers Agricultures* 22.5 (2013), 336–344 (1). ISSN: 1777-5949. DOI: 10.1684/AGR.2013.0657. URL: https://revues.cirad.fr/index.php/cahiers-agricultures/article/view/31039 (cited on pages 46, 47).

Clémentine Vitte et al. “The bright side of transposons in crop evolution”. In: *Briefings in Functional Genomics and Proteomics* 13.4 (2014), pages 276–295. ISSN: 14774062. DOI: 10.1093/bfgp/elu002. URL: https://academic.oup.com/bfg/article/13/4/276/282035 (cited on page 35).

Huai Wang, Tina Nussbaum-Wagler, et al. “The origin of the naked grains of maize”. In: *Nature* 436.7051 (2005), pages 714–719. ISSN: 1476-4687. DOI: 10.1038/nature03863. URL: https://doi.org/10.1038/nature03863 (cited on page 32).

Rong-Lin Wang, Adrian Stec, et al. “The limits of selection during maize domestication”. In: *Nature* 398.6724 (1999), pages 236–239. ISSN: 1476-4687. DOI: 10.1038/18435. URL: https://doi.org/10.1038/18435 (cited on page 32).

Shuo Wang, Yong-Qing Qian, et al. “Graph-based pan-genomes: increased opportunities in plant genomics”. In: *Journal of Experimental Botany* 74.1 (2023), pages 24–39. ISSN: 0022-0957. DOI: 10.1093/JXB/ERAC412. URL: https://academic.oup.com/jxb/article/74/1/24/6762754 (cited on pages 12, 128).

Wensheng Wang, Ramil Mauleon, et al. “Genomic variation in 3,010 diverse accessions of Asian cultivated rice”. In: *Nature* 557.7703 (2018), pages 43–49. ISSN: 14764687. DOI: 10.1038/s41586-018-0063-9 (cited on pages 5, 7, 9, 43, 51, 53, 54, 101).

Yuexing Wang, Guosheng Xiong, et al. “Copy number variation at the GL7 locus contributes to grain size diversity in rice”. In: *Nature Genetics 2015 47:8* 47.8 (2015), pages 944–948. ISSN: 1546-1718. DOI: 10.1038/ng.3346. URL: https://www.nature.com/articles/ng.3346 (cited on pages 39, 45, 83, 127).

Zong-Yang-Y. Wang, Fei-Qin-Q Zheng, et al. “The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene”. In: *The Plant Journal* 7.4 (1995), pages 613–622. ISSN: 1365-313X. DOI: 10.1046/J.1365-313X.1995.7040613.X. URL: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1365-313X.1995.7040613.x (cited on page 32).

Ehud Weiss, Mordechai E Kislev, and Anat Hartmann. “Autonomous Cultivation Before Domestication”. In: *Science* 312.5780 (2006), pages 1608–1610. DOI: 10.1126/science.1127235. URL: https://www.science.org/doi/abs/10.1126/science.1127235 (cited on page 29).

Thomas Wicker et al. “A unified classification system for eukaryotic transposable elements”. In: *Nature Reviews Genetics* 8 (2007), page 973. URL: http://dx.doi.org/10.1038/nrg2165 (cited on pages 34, 148).

Adam S. Wilkins, Richard W. Wrangham, and W. Tecumseh Fitch. “The “Domestication Syndrome” in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics”. In: *Genetics* 197.3 (2014), pages 795–808. ISSN: 1943-2631. DOI: 10.1534/genetics.114.165423. URL: https://doi.org/10.1534/genetics.114.165423 (cited on page 27).

Thilo Winzer et al. “A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine.” In: *Science (New York, N.Y.)* 336.6089 (2012), pages 1704–8. ISSN: 1095-9203. DOI: 10.1126/science.1220757. URL: https://www.science.org/doi/10.1126/science.1220757 (cited on page 36).

Samuel Tareke Woldegiorgis et al. “Identification of Heat-Tolerant Genes in Non-Reference Sequences in Rice by Integrating Pan-Genome, Transcriptomics, and QTLs”. In: *Genes* 13.8 (2022), page 1353. ISSN: 20734425. DOI: 10.3390/GENES13081353. URL: https://www.mdpi.com/2073-4425/13/8/1353 (cited on page 43).

Tobias Würschum, Philipp H G Boeven, et al. “Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat”. In: *BMC Genetics* 16.1 (2015), page 96. ISSN: 1471-2156. DOI: 10.1186/s12863-015-0258-0. URL: https://doi.org/10.1186/s12863-015-0258-0 (cited on page 38).

Tobias Würschum, C Friedrich H Longin, et al. “Copy number variations of CBF genes at the Fr-A2 locus are essential components of winter hardiness in wheat”. In: *The Plant Journal* 89.4 (2017), pages 764–773. ISSN: 0960-7412. DOI: https://doi.org/10.1111/tpj.13424. URL: https://doi.org/10.1111/tpj.13424 (cited on page 38).

Han Xiao et al. “A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit”. In: *Science* 319.5869 (2008), pages 1527–1530. ISSN: 00368075. DOI: 10.1126/SCIENCE.1153040. URL: https://www.science.org/doi/10.1126/science.1153040 (cited on page 39).

Kenong Xu, Xia Xu, et al. “Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice”. In: *Nature* 442.7103 (2006), pages 705–708. ISSN: 0028-0836. DOI: 10.1038/nature04920. URL: http://www.nature.com/articles/nature04920 (cited on pages 5, 52, 83, 127).

Xun Xu, Xin Liu, et al. “Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes”. In: *Nature Biotechnology* 30.1 (2012), pages 105–111. ISSN: 1546-1696. DOI: 10.1038/nbt.2050. URL: https://doi.org/10.1038/nbt.2050 (cited on pages 4, 38, 39, 127).

Haidong Yan et al. “Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet”. In: *Nature Genetics* 55.3 (2023), pages 507–518. ISSN: 1546-1718. DOI: 10.1038/s41588-023-01302-4. URL: https://doi.org/10.1038/s41588-023-01302-4 (cited on page 127).

Sihai Yang, Jing Li, et al. “Rapidly evolving R genes in diverse grass species confer resistance to rice blast disease”. In: *Proceedings of the National Academy of Sciences* 110.46 (2013), pages 18572–18577. DOI: 10.1073/pnas.1318211110. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1318211110 (cited on page 38).

Tao Yang, Rong Liu, et al. “Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics”. In: *Nature Genetics* 54.10 (2022), pages 1553–1563. ISSN: 15461718. DOI: 10.1038/S41588-022-01172-2 (cited on pages 42, 43, 87).

Wen Yao et al. “Exploring the rice dispensable genome using a metagenome-like assembly strategy.” In: *Genome biology* 16 (2015), page 187. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0757-3. URL: http://www.ncbi.nlm.nih.gov/pubmed/26403182%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4583175 (cited on pages 5, 7, 51, 52, 54, 83).

Jingyin Yu, Agnieszka A Golicz, et al. “Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars”. In: *Plant Biotechnology Journal* 17.5 (2019), pages 881–892. DOI: https://doi.org/10.1111/pbi.13022. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pbi.13022 (cited on page 43).

Ping Yu, Cai-Hong Wang, et al. “Genome-wide copy number variations in Oryza sativa L.” In: *BMC Genomics* 14.1 (2013), page 649. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-649. URL: https://doi.org/10.1186/1471-2164-14-649 (cited on page 38).

Fan Zhang, Hongzhang Xue, et al. “Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes.” eng. In: *Genome research* 32.5 (2022), pages 853–863. ISSN: 1549-5469 (Electronic). DOI: 10.1101/gr.276015.121 (cited on page 43).

Xiaohui Zhang, Tongjin Liu, et al. “Pan-genome of Raphanus highlights genetic variation and introgression among domesticated, wild, and weedy radishes”. In: *Molecular Plant* 14.12 (2021), pages 2032–2055. ISSN: 1674-2052. DOI: https://doi.org/10.1016/j.molp.2021.08.005. URL: https://www.sciencedirect.com/science/article/pii/S167420522100318X (cited on page 43).

Qiang Zhao et al. “Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice”. In: *Nature Genetics* (2018), page 1. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0041-z. URL: http://www.nature.com/articles/s41588-018-0041-z (cited on pages 5, 7, 42, 43, 51, 54, 123).

Lei-Ying Zheng et al. “Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor)”. In: *Genome Biology* 12.11 (2011), R114. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-11-r114. URL: https://doi.org/10.1186/gb-2011-12-11-r114 (cited on page 38).

Peng Zhou, Kevin A. T. Silverstein, et al. “Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes”. In: *BMC Genomics* 18.1 (2017), page 261. ISSN: 1471-2164. DOI: 10.1186/s12864-017-3654-1. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3654-1 (cited on page 43).

Yao Zhou, Zhiyang Zhang, et al. “Graph pangenome captures missing heritability and empowers tomato breeding”. In: *Nature* 606.7914 (2022), pages 527–534. ISSN: 0028-0836. DOI: 10.1038/s41586-022-04808-9. URL: https://www.nature.com/articles/s41586-022-04808-9 (cited on page 129).

Zhengkui Zhou, Yu Jiang, et al. “Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean.” In: *Nature biotechnology* 33.4 (2015), pages 408–14. ISSN: 1546-1696. DOI: 10.1038/nbt.3096. URL: http://www.ncbi.nlm.nih.gov/pubmed/25643055 (cited on pages 4, 39, 127).

Appendix

INTRODUCTION

Genetic diversity shaped by evolutionary process

Type	Mechanism	Trigger	SV	Ref.
DNA break error	non-homologous end joining (NHEJ)	misguided fusion of double-strand breaks in DNA	Insertion, deletion, (translocation)	Moore and Haber 1996
Recombination error	non-allelic homologuous recombination (NAHR)	Misalignment in sequences housing highly identical sequences	duplication, deletion, CNV, inversion, translocation	Lupski 1998
Replication error	fork stalling and template switching (FoSTeS)	fork stalling and polymerase switching at a	large rearrangements, inversions,	Hastings et al. 2009; Lee, Carvalho, and Lupski 2007
	microhomology-mediated break-induced replication (MMBIR)	nearby single-stranded DNA	duplications and translocation	Hastings et al. 2009

Table 14.1: Main cellular mechanisms causing structural variations.

African rice

HOW TO BUILD THE AFRICAN RICE PANGENOME ?

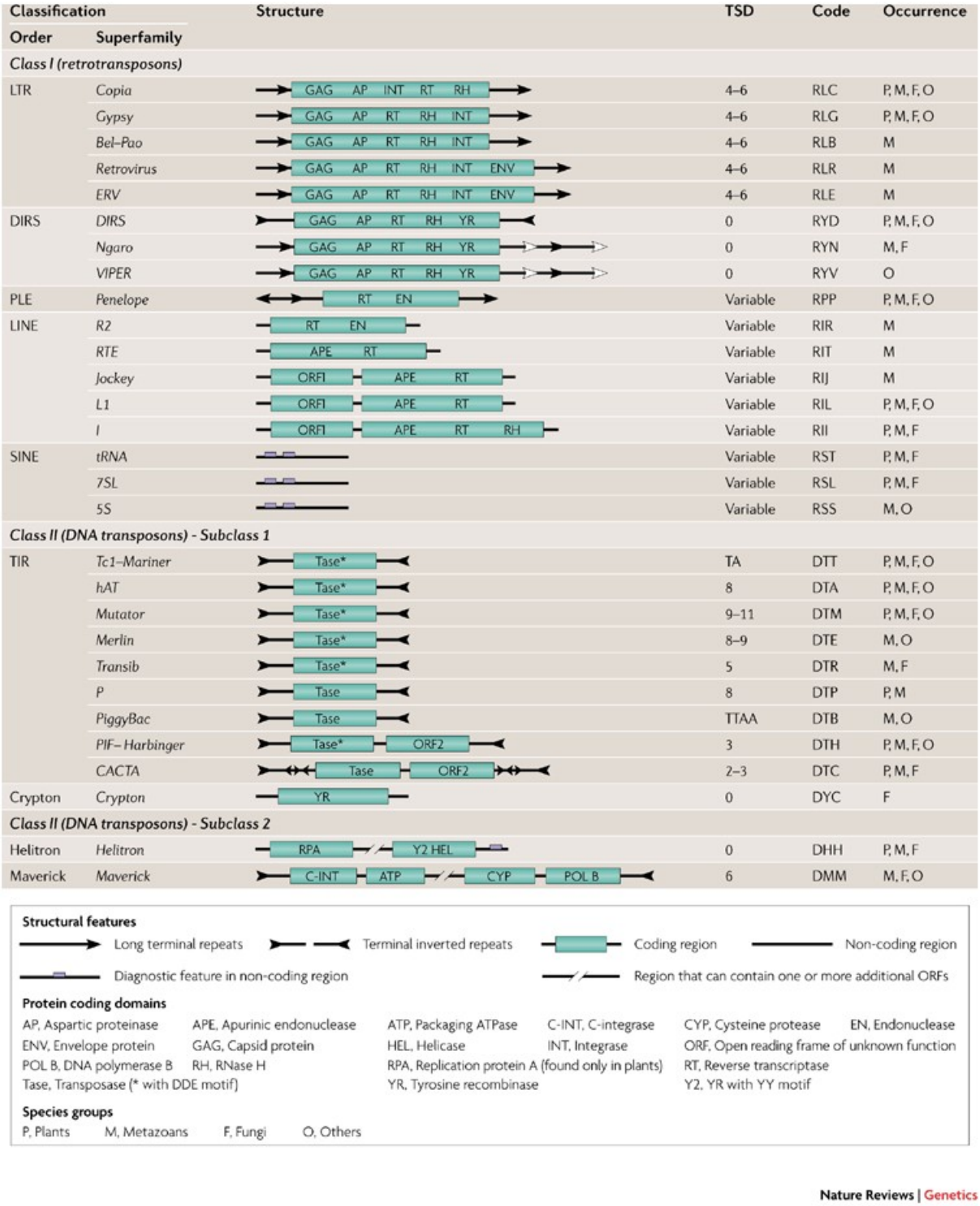

Figure 14.3: Transposable element classification.
Figure from Wicker et al. 2007.

Section Complex *Species*	Chromosome number (DNA content [pg/2C])*	Genome group	Usual habitat
Oryza			
***Oryza sativa* complex**			
Oryza sativa L.	24 (0.91–0.93)	AA	Upland to deepwater; open
O. rufipogon sensu lacto† (syn: *O. nivara* for the annual form *O. rufipogon sensu stricto* for the perennial form)	24 (0.95)	AA	(Annual) Seasonally dry; open (Perennial) Seasonally deepwater and wet year round; open
O. glaberrima Steud.	24 (0.87)	AA	Upland to deepwater; open
O. barthii A. Chev.	24	AA	Seasonally dry; open
O. longistaminata Chev. et Roehr.	24 (0.81)	AA	Seasonally dry to deepwater; open
O. meridionalis Ng	24 (1.02)	AA	Seasonally dry; open
O. glumaepatula Steud.‡	24 (0.99)	AA	Inundated areas that become seasonally dry; open
***O. officinalis* complex**			
O. officinalis Wall ex Watt	24 (1.45)	CC	Seasonally dry; open
O. minuta JS Presl. ex CB Presl.	48 (2.33)	BBCC	Stream sides; semi shade
O. rhizomatis Vaughan	24	CC	Seasonally dry; open
O. eichingeri Peter§	24 (1.47)	CC	Stream sides, forest floor; semi shade
O. malapuzhaensis Krishnaswamy and Chandrasakaran	48	BBCC	Seasonally dry forest pools; shade
O. punctata Kotschy ex Steud.	24 (1.11), 48	BB, BBCC	(Diploid) seasonally dry; open (Tetraploid) forest floor; semi shade
O. latifolia Desv.#	48 (2.32)	CCDD	Seasonally dry; open
O. alta Swallen	48	CCDD	Seasonally inundated; open
O. grandiglumis (Doell.) Prod.	48 (1.99)	CCDD	Seasonally inundated; open
O. australiensis Domin	24 (1.96)	EE¥	Seasonally dry; open
Ridleyanae Tateoka			
O. schlechteri Pilger	48	Unknown**	River banks; open
***O. ridleyi* complex**			
O. ridleyi Hook.	48 (1.31–1.93)	HHJJ	Seasonally inundated forest floor; shade
O. longiglumis Jansen	48	HHJJ	Seasonally inundated forest floor; shade
Granulata Roschev.			
***O. granulata* complex**¶			
O. granulata Nees et Arn ex Watt	24	GG	Forest floor; shade
O. meyeriana (Zoll. et Mor. ex Steud.) Baill.	24	GG	Forest floor; shade
Brachyantha B.R. Lu			
O. brachyantha Chev. Et Roehr.	24 (0.72)	FF	Rock pools; open

Figure 14.4: Oryza species description: their chromosome number, DNA content, genome group and usual habitat.
Table from Vaughan, Morishima, and Kadowaki 2003

.

Software Name	**Version**	**url ref**
Docker		https://docs.docker.com/get-docker/
Jupyter		
Python	3.9.7	http:/www.python.org
biopython		https://biopython.org/
ea-utils	1.01	https://expressionanalysis.github.io/ea-utils/
bwa	0.7.17	https://github.com/lh3/bwa/blob/master/README.md ref
samtools	1.10	http://www.htslib.org/ ref
abyss	2.0	https://github.com/bcgsc/abyss/blob/master/README.md ref
assembly-stats	1.01	https://github.com/sanger-pathogens/assembly-stats ref
CD-HIT	4.8.1	https://github.com/weizhongli/cdhit/blob/master/README ref

Table 14.2: List of main tools required by frangiPANe.
The complete list with python packages is available at https://github.com/tranchant/frangiPANe. Table From Tranchant-Dubreuil et al. 2023

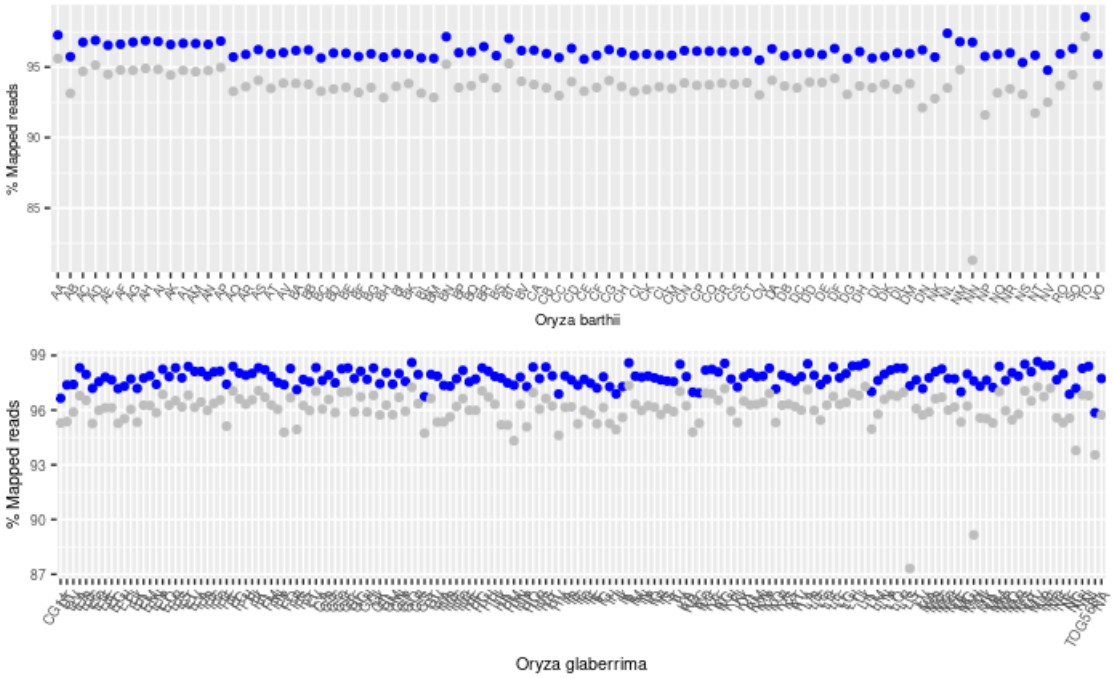

Figure 14.5: Mapping statistics of 248 african rice (164 domesticated and 84 wild relatives).
Each sample is plotted with its respective percentage of mapped reads (blue) and reads correctly mapped in pair (grey) for *Oryza barthii* (top) and *O. glaberrima* (bottom) respectively. Figure From Tranchant-Dubreuil et al. 2023.

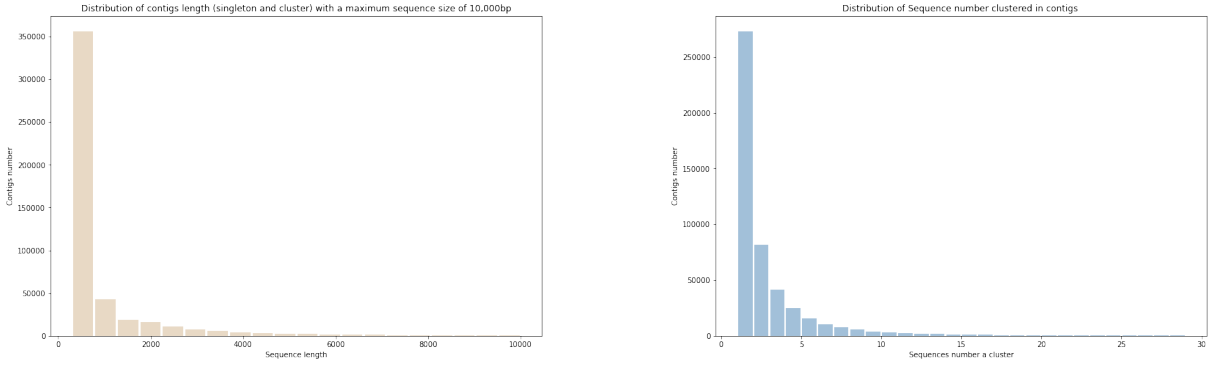

Figure 14.6: Reducing redundancy.
(a) Distribution of the sequence length (bp) of the contigs (singleton and cluster) after removing redundancy with cd-hit (with a maximum length set at 10,000 pb). The average sequence size was 1,060 bp. (b) Distribution of the singleton number and and the sequence number per cluster. Figure From Tranchant-Dubreuil et al. 2023.

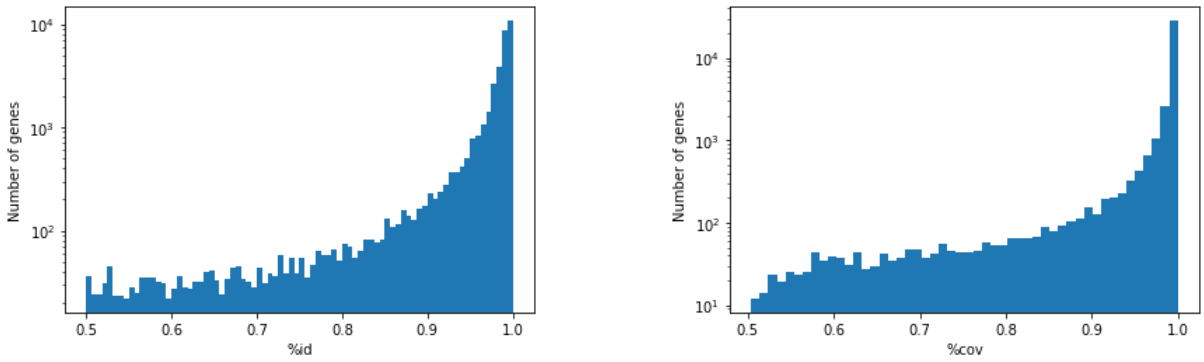

Figure 14.7: Genes mapping.
(a) Distribution of the Nipponbare and the panreference sequence identity. This histogram shows the distribution of exon sequence identity of genes. (b) Distribution of the Nipponbare and the panreference alignment coverage in exon. This histogram shows the distribution of alignment coverage in exons. Figure From Tranchant-Dubreuil et al. 2023.

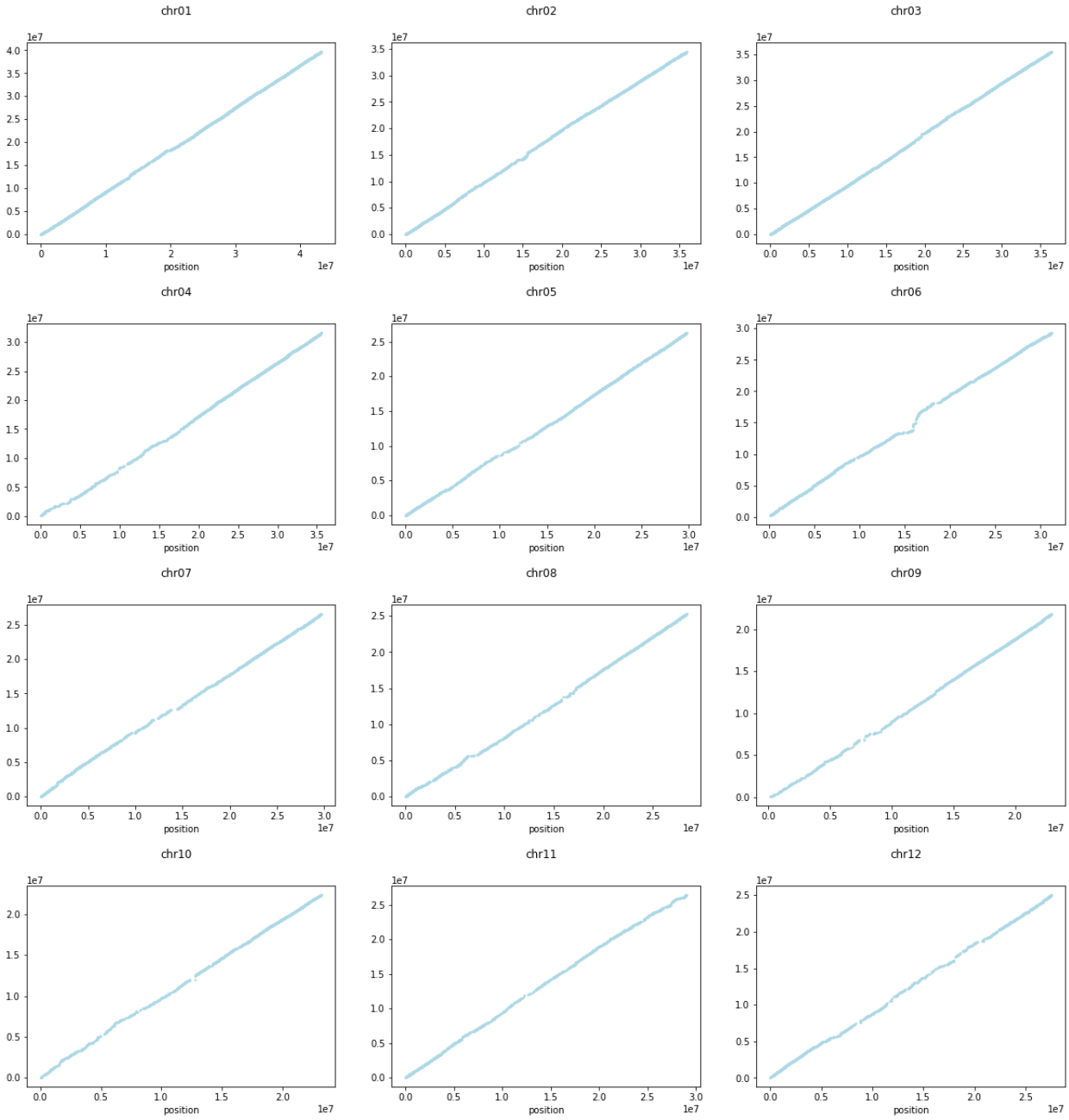

Figure 14.8: The Nipponbare and the CG14 gene order.
Each dot plot shows the position of each gene on the CG14 (x-axis) and the Nipponbare (y-axis) chromosome. Figure From Tranchant-Dubreuil et al. 2023.

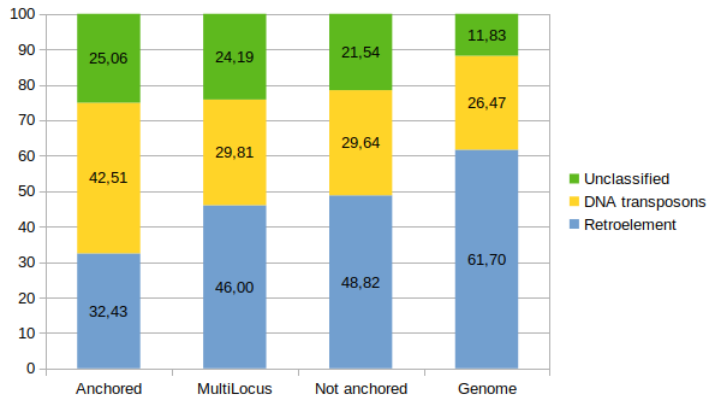

Figure 14.9: Ratio of major transposable element classes (Retrotransposon (or Retroelement), DNA transposon and unclassified element) on the reference genome and all non redundant contigs respectively. The ratio of TE classes is also detailed for contigs that were placed on the reference genome (single or multiple positions) or unplaced. Figure From Tranchant-Dubreuil et al. 2023.

Abstract

Over the last few decades, thanks to dramatic advances in high-throughput sequencing technologies, more and more genomes were assembled. By comparing multiple genomes within the same species, it became increasingly clear that a single reference genome cannot capture all the diversity within a species. Huge variations in the number of genes, and more broadly in structural variations, between individuals were observed. This led to a progressive paradigm shift from the genome towards the pangenome. This latter refers to the set of all sequences (genic or not) present in a given population, including (i) the core genome containing sequences present in all individuals and (ii) the dispensable genome composed of sequences shared by a subset of individuals. The main objectives of this PhD project were to fully explore the genetic diversity within a crop, the African rice, and how the diversity of this species was reshaped during its domestication. The two main questions we wanted to specifically address were which roles do these structural variations play in gene composition and adaptation, and how evolutionary forces have shaped (pan)genome organization and dynamics.

First, we performed a review of the current state-of-the-art on the emerging concept of pangenome, to identify challenges, opportunities and limitations. We presented how this new methodological approach, combined with the high-throughput sequencing technologies, offers unprecedented opportunities to investigate genomic variability, thus providing an alternative way to study genome organization and dynamics.

Next, we developed frangiPANe, a tool for creating a eukaryotic pangenome from multiple, individual short-read-based genome sequencing. Our method was validated using the whole genome resequencing data from 248 African rice accessions, both cultivated and wild, as a proof-of-concept to build the first large panreference for African rice. We identified an average of 8Mb of new sequences per individual, absent from the reference genome, i.e. a total of 513.5 Mb new sequence across all individuals. For a reference genome size of 350Mb, we added 60% more sequences, represented in 484,394 non redundant contigs. Despite the high content of transposable elements, we were able to anchor at an unique position on the reference genome 31.5% of the contigs.

Finally, we explored than pangenome diversity during African rice domestication. We annotated 22,765 genes across all individuals, increasing the number of genes from 40,553 in the reference by 56%. We showed the relative presence of the 484,394 contigs recapitulate the relationship between individuals previously determined using single nucleotide polymorphism, and consequently also reflects the evolutionary history of the cultivated and wild species. We thus used an innovative approach to assess selection in these 483,394 structural variations, and identify 683 candidate genes as under selection. Interestingly, we were able to find selection in the PROG1 gene, a gene selected during African and Asian rice domestication and associated with a major deletion during African rice domestication. We also showed domestication was associated with a loss of genes after the initial bottleneck, a likely consequence of an increase of drift during the domestication process. Altogether, our approaches help to reshape our thinking about the consequences of domestication on diversity, associated both to a loss of diversity and genes.

Résumé

Au cours des dernières décennies, grâce aux progrès spectaculaires des technologies de séquençage à haut débit, de plus en plus de génomes ont été assemblés. En comparant plusieurs génomes au sein d'une même espèce, il est devenu de plus en plus évident qu'un seul génome de référence ne peut capturer toute la diversité présente au sein d'une espèce. D'importantes variations du nombre de gènes, et plus largement des variations structurales, entre les individus d'une même espèce ont été observées. Cela a conduit à un changement progressif de paradigme, du concept de génome à celui de pangénome. Le pangénome désigne l'ensemble des séquences (géniques ou non) présentes au sein d'une population et il se compose du "core genome" regroupant les séquences présentes chez tous les individus et (ii) le "dispensable genome" contenant des séquences uniquement partagées par une partie des individus. Les principaux objectifs de ce projet doctoral étaient d'explorer la diversité génétique au sein d'une céréale, le riz africain, et la manière dont la diversité de cette espèce a été remodelée au cours de sa domestication. Les deux principales questions que nous souhaitions aborder étaient les suivantes : quels rôles jouent ces variations structurales dans la composition en gènes et l'adaptation, et comment les forces évolutives ont façonné l'organisation et la dynamique du (pan)génome.

Tout d'abord, nous avons réalisé un état de l'art des connaissances sur le concept émergent de pangénome, afin d'identifier les défis, les opportunités et les limites. Nous avons présenté comment cette nouvelle approche, combinée aux technologies de séquençage à haut débit, offre des possibilités sans précédent pour étudier la variabilité génomique, en fournissant une approche alternative pour étudier l'organisation et la dynamique des génomes.

Ensuite, nous avons développé frangiPANe, un outil permettant de créer un pangénome d'eucaryote à partir de données de séquençage "short-read" de génomes d'individus. Notre méthode a été validée en utilisant les données de séquençage de 248 génomes de riz africain, cultivés et sauvages, comme preuve de concept pour construire la première panréférence du riz africain. Nous avons identifié, en moyenne, 8 Mb de séquences par individu, absentes du génome de référence, soit au total 513,5 Mb de nouvelles séquences. Pour un génome de référence de 350 Mb, nous avons ajouté 60% de séquences, correspondant à 484 394 contigs non redondants. Malgré un contenu important en éléments transposables dans ces nouvelles séquences, nous avons pu ancrer 31,5% de ces contigs à une position unique sur le génome de référence.

Enfin, nous avons exploré la diversité du pangénome au cours de la domestication du riz africain. Nous avons annoté 22 765 nouveaux gènes, augmentant de 56% le nombre de gènes par rapport aux 40 553 gènes de référence. Nous avons montré que la présence relative des 484 394 contigs récapitule la relation entre les individus établie précédemment à partir de "SNPs" et, par conséquent, reflète également l'histoire évolutive des espèces cultivées et sauvages. Nous avons ensuite utilisé une approche innovante pour évaluer des traces de sélection dans ces 483 394 variations structurales et 683 gènes candidats ont été identifiés comme étant sous sélection. Parmi ces gènes, nous avons détecté le gène PROG1, qui a été sélectionné au cours de la domestication du riz africain et asiatique et qui est associé à une délétion lors de la domestication du riz africain. Nous avons également montré que la domestication était associée à une perte de gènes après le goulot d'étranglement initial, probablement la conséquence d'une augmentation de la dérive génétique au cours du processus de domestication. Dans l'ensemble, nos approches contribuent à enrichir nos connaissances sur les conséquences de la domestication sur la diversité, associée à la fois à une perte de diversité et de gènes.