Contribution des variations structurales de type insertions/délétions à l'adaptation, la variation des caractères et les performances hybrides chez le maïs

Clément Mabire

NNT : 2019SACLS093

Contribution des variations structurales de type insertions / délétions à l'adaptation, la variation des caractères et les performances hybrides chez le maïs

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°581 :
Agriculture, Alimentation, Biologie, Environnement et Santé (ABIES)
Spécialité de doctorat : Sciences agronomiques

Thèse présentée et soutenue à Gif-Sur-Yvette, le 23 avril 2019, par

Clément Mabire

Composition du Jury :

Régine Delourme Directrice de recherche, INRA (IGEPP)	Présidente
Mathilde Causse Directrice de recherche, INRA (GAFL)	Rapporteure
Etienne Paux Directeur de recherche, INRA (GDEC)	Rapporteur
Karine Alix Maitre de conférences, AgroParisTech	Examinatrice
Stéphane Nicolas Chargé de recherche, INRA (GQE Le Moulon)	Encadrant de thèse
Alain Charcosset Directeur de recherche, INRA (GQE Le Moulon)	Directeur de thèse

Thèse de doctorat

Remerciements

Je tiens d'abord à remercier Alain Charcosset, mon directeur de thèse, pour m'avoir accueilli dans son équipe et pour m'avoir permis de réaliser ce projet de thèse. Merci de m'avoir fait profiter de ton expérience, d'avoir pris du temps pour me donner ton avis sur mes travaux et pour toutes ces relectures. Travailler dans ton équipe a été très enrichissant, professionnellement et personnellement, c'était un grand plaisir !

Un grand merci à Stéphane Nicolas qui a été mon encadrant pendant ces trois ans de thèse et quatre ans passés au Moulon. Merci de m'avoir fait confiance pour la réalisation ce projet de thèse. Merci pour ta disponibilité, pour tout ce temps passé à échanger, à confronter nos idées (malgré parfois quelques désaccords) et surtout merci de m'avoir souvent sorti de ma zone de confort. J'ai grâce à toi énormément appris et progressé pendant ces quatre dernières années.

Merci aux membres de mon jury et particulièrement à mes rapporteurs, Mathilde Causse et Etienne Paux, mais aussi à mes examinatrices, Karine Alix et Régine Delourme, qui ont accepté de passer du temps à évaluer mes travaux et venir m'écouter.

Merci à Julie Fievet, pour ton aide, tes conseils, tes relectures et ton soutien pour cette fin de thèse, ainsi qu'à Laurence Moreau et Tristan Mary-Huard pour vos remarques, conseils et/ou relectures avisés. Travailler de près ou de loin avec vous a été un réel plaisir.

Merci également aux membres de mon comité de thèse, les échanges que nous avons eus et vos points de vue ont été importants pour moi et m'ont permis de faire les bons choix (je l'espère) pendant ces trois ans. Merci en particulier à Pierre Dubreuil, d'avoir partagé une partie de ses résultats, nos échanges sur les analyses ont été enrichissants.

Merci à Jorge Duarte pour l'ensemble du travail qu'il a accompli pour le développement de la puce de génotypage et avec qui nous avons beaucoup échangé pour l'écriture du papier. Merci aussi à Romane Guilbaud pour m'avoir épaulé sur la réalisation des dernières analyses de cette thèse.

Merci à Delphine Madur de m'avoir fourni les données de génotypage et de m'avoir éclairé sur les aspects plus techniques. Merci également à Valérie Combes, pour tes délicieux gâteaux qui ont été d'un réconfort absolu ! Merci à Cyril Bauland, que ce soit pour tes lumières sur la sélection du maïs, les cours d'œnologie ou encore pour les escapades en montagne. Merci à tous les collègues de l'équipe que j'ai pu rencontrer depuis le début de cette expérience, Philippe Jamin, Sophie Pin, Héloïse Giraud, Sandra Negro, Romain Barbier, Yohann Benoit, Fabien Laporte, Camille Clipet,… ça a été un plaisir de vous rencontrer et de travailler à vos côtés.

Ces trois ans de thèse ont été très agréables et sont passés très vite, en grande partie grâce à l'ambiance très chaleureuse qui règne au Moulon. Merci à vous tous, pour cette bonne humeur, les pauses café, les longues heures de jardinage, les repas au soleil, le basket (sport national au Moulon, même si mon niveau n'a pas vraiment évolué), les barbecues et soirées crêpes (avec feux d'artifice grandioses),…

Et je garde le meilleur presque pour la fin. Un grand merci à Simon Rio et Adama Seye qui ne m'ont pas lâché depuis mon arrivée au Moulon, avec qui j'ai partagé le même bureau pendant ces quatre ans. Mais aussi Antoine Allier qui est venu se greffer à l'équipe en cours de route et avec qui j'ai enfin pu parler sérieusement de rugby. Votre compagnie, mais aussi votre gentillesse, vos conseils, nos débats, nos discussions, votre humour (je ne ferai pas de commentaire sur les blagues) ont rendu ces années au Moulon inoubliables ! Sans vous, ces quatre dernières années auraient été certainement bien différentes. Je pense qu'aujourd'hui vous êtes plus des amis que des collègues, je vous souhaite vraiment le meilleur pour la suite !

Et enfin, merci à mes proches, à tous ceux qui m'ont encouragé, en particulier à mon père qui m'a depuis toujours soutenu dans mes choix, ce qui m'a permis en partie d'en arriver là, je te dois beaucoup (c'est bon, les études c'est fini, enfin, je crois). Merci à ma sœur, Pauline, qui a toujours été là pour prendre des nouvelles et me soutenir dans les moments plus difficiles. Et enfin merci à Suzanne d'avoir été à mes côtés, pour m'avoir souvent remotivé et pour ta patience ! Me supporter pendant ces derniers mois n'a pas dû être toujours facile.

Contents

Introduction

L'histoire du maïs cultivé

Le maïs a été domestiqué il y a 8 700 ans dans les régions montagneuses d'Amérique centrale près de Mexico à partir de la téosinte (Matsuoka et al., 2002). Il s'est ensuite adapté à une très large gamme de conditions agro-climatiques. Son aire de culture s'est d'abord étendue vers le Nord et le Sud des Amériques puis en Europe (Tenaillon and Charcosset, 2011) (Figure 1) et dans le reste du monde après la découverte du nouveau continent par les Européens (Mir et al., 2013). Cette expansion a entrainé la création de différents groupes génétiques composés d'individus adaptés à des conditions climatiques locales. Par exemple, les maïs dit tropicaux montrent une large gamme de variation de la durée entre le semis et la floraison (Gouesnard et al., 2012) mais leur cycle est généralement trop long pour permettre la production de grains dans les conditions climatiques d'Europe du Nord. Une étape clef dans la diffusion du maïs a donc été l'adaptation aux climats tempérés avec l'ajustement du cycle de la plante à la durée du jour et l'adaptation à des températures basses (Tenaillon and Charcosset, 2011; Romero Navarro et al., 2017; Swarts et al., 2017).

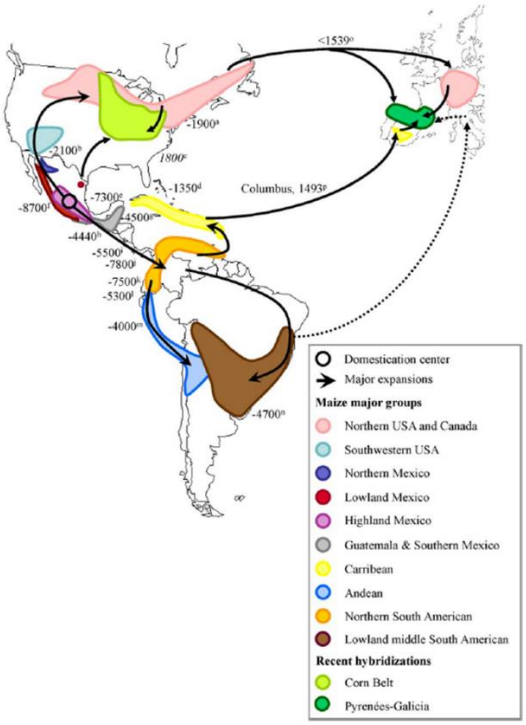


Figure 1 : Expansion du maïs depuis l'Amérique centrale vers le Sud, le Nord et vers l'Europe. D'après Tenaillon et Charcosset (2011).

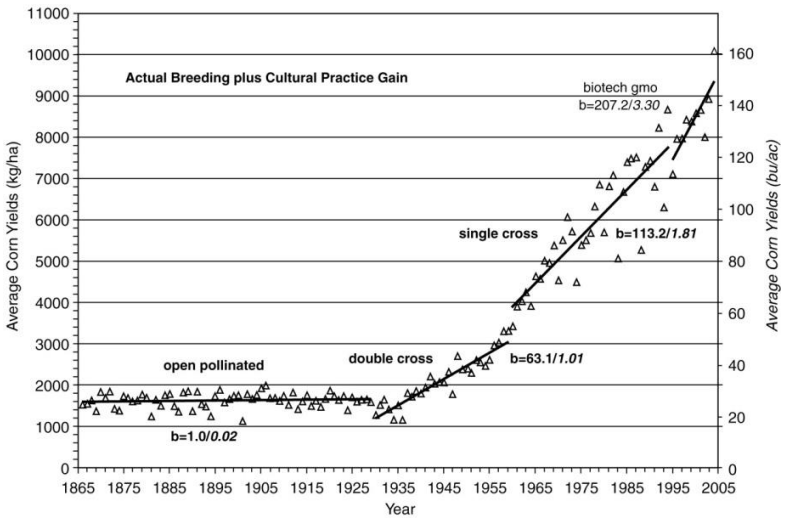


Figure 2 : Evolution des rendements de maïs aux USA depuis 1860 jusqu'à 2005. Le type de matériel végétal est indiqué au-dessus du nuage de points. D'après Duvick et al. (2005).

Le maïs est une espèce allogame qui a longtemps été cultivée sous forme de populations à pollinisation libre, composées d'individus majoritairement hétérozygotes et génétiquement hétérogènes. Ces populations étaient adaptées aux conditions agro-climatiques de leurs régions de production, mais leur rendement n'a pas évolué aux USA entre 1860 et 1930 (Duvick, 2005) (Figure 2). En (1908), East et Shull ont observé chez le maïs que l'autofécondation des individus entrainait un phénomène de diminution de leur vigueur, appelé dépression de consanguinité. Cependant, lorsque deux individus homozygotes et distants génétiquement sont croisés entre eux, l'hybride obtenu présente une performance bien supérieure à celle de ses parents. Cette vigueur hybride a été ensuite définie par Shull (1914) sous le terme d'hétérosis. Ce phénomène a permis de mettre en place de nouveaux schémas de sélection visant à produire des variétés hybrides, en créant tout d'abord des lignées homozygotes. Aux USA, la sélection hybride chez le maïs a permis d'augmenter fortement les rendements à partir des années 30 (Figure 2). Cette transition des populations vers les hybrides s'est faite en deux temps. D'abord en utilisant des hybrides doubles voies (croisement de deux hybrides), puis dans les années 60, des hybrides issus d'un croisement simple entre deux lignées homozygotes ont progressivement remplacé les hybrides doubles voies (Figure 2) du fait de l'accroissement de la productivité des lignées parentales (Figure 3).

La création et la sélection de nouvelles lignées parentales homozygotes plus productives en semences ou en pollen à partir des lignées des cycles de sélection précédent a permis vraisemblablement de progressivement contre sélectionner les allèles délétères et ainsi de purger le fardeau génétique (Barrett and Charlesworth, 1991; Charlesworth and Willis, 2009) ce qui a entraîné une augmentation de leur production. Toutefois, les lignées ont été essentiellement sélectionnées pour leur aptitude à la combinaison, qui est la performance moyenne de leur descendance hybride en croisement avec d'autres lignées (Sprague and Tatum, 1942). Pour la production d'hybride de maïs dans le nord de l'Europe, les lignées sont réparties en deux groupes hétérotiques (corné et denté) et les lignées d'un groupe sont sélectionnées pour maximiser la complémentation avec les lignées du deuxième groupe.

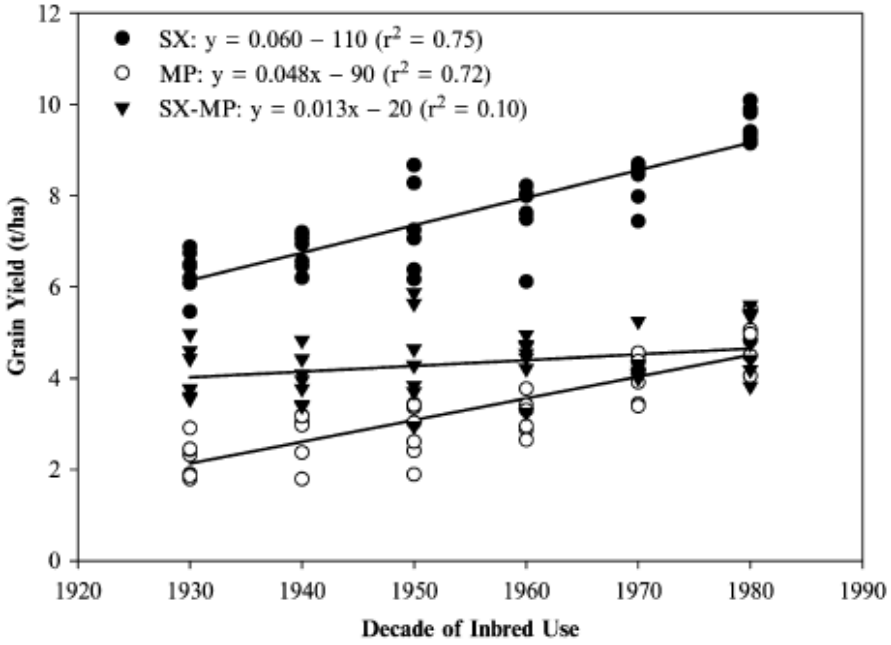


Figure 3 : Evolution du rendement des hybrides F1 (SX) et des lignées parentales (MP) ainsi que de la valeur de l'hétérosis (SX-MP) entre les années 1930 et 1980 aux USA. D'après Duvick et al. (2005)

Développement du génotypage et utilisation en génétique quantitative

Les lignées de maïs constituent une ressource génétique d'une grande richesse pour la sélection et la production d'hybrides mais aussi pour approfondir les connaissances sur le maïs (Gorjanc et al., 2016; Böhm et al., 2017; Brauner et al., 2018). Elles permettent d'étudier l'histoire du maïs (Brandenburg et al., 2017; Swarts et al., 2017), mais aussi l'architecture génétique des caractères (Tian et al., 2011; Bouchet et al., 2013; Peiffer et al., 2014; Pace et al., 2015) et les signatures de sélection (Bouchet et al., 2013; Takuno et al., 2015; Unterseer et al., 2016; van Heerwaarden et al., 2012) notamment grâce au développement rapide et récent des technologies de marquage moléculaire à haut débit.

L'arrivée des premiers marqueurs moléculaires a permis d'étudier le déterminisme génétique des caractères quantitatifs en étudiant le lien entre la variation allélique révélée par ces marqueurs et les régions génomiques responsables de la variation des caractères quantitatifs appelées « quantitative trait loci » (QTL). Cette approche repose sur le déséquilibre de liaison (DL) définit comme une association non aléatoire entre des allèles à des loci différents dans une même population. La recherche de QTL vise à tester statistiquement les différences de phénotype entre les sous-classes de la population, triées selon les allèles au marqueur (Figure 4). Le développement de populations à partir d'un croisement entre deux individus contrastés pour un caractère donné permet de faire ségréger les différents gènes impliqués dans la variation du caractère dans la population. Le développement des analyses de liaison à partir de ces populations biparentales nécessite peu de marqueurs moléculaires car peu d'évènements de recombinaisons ont eu lieu dans la population (Paterson et al., 1988), ce qui entraine cependant une faible résolution autour du QTL. Ces approches ont permis de détecter de nombreux QTL chez de nombreuses

espèces de plantes (Paterson et al., 1988; Lander and Botstein, 1989). Cette approche nécessite de plus le développement de populations biparentales spécifiques et se limite aux caractères contrastés entre les deux parents choisis (Lynch and Walsh, 1998). Afin d'augmenter la diversité dans les populations analysées, différents types de populations ont été développés à partir de plusieurs parents fondateurs comme les populations MAGIC (Cavanagh et al., 2008; Pascual et al., 2015) ou encore les populations NAM (Blanc et al., 2006; Yu et al., 2008; McMullen et al., 2009; Giraud et al., 2017a).

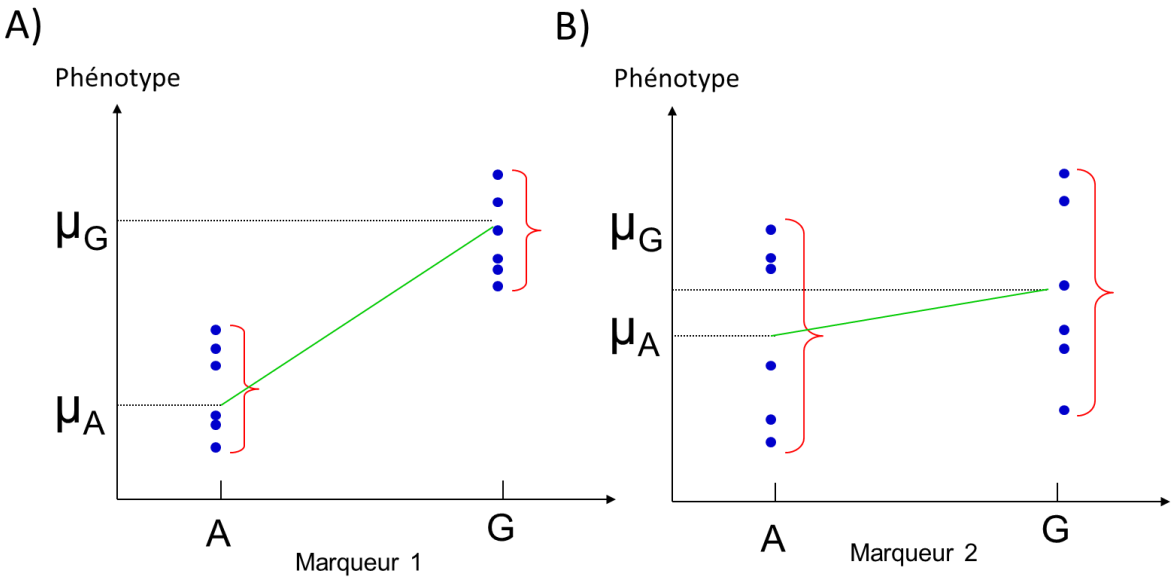


Figure 4 : représentation schématique de la valeur du phénotype en fonction de l'allèle A ou G pour deux marqueurs à deux locus différents. μG et μA sont les moyennes du phénotype pour les individus qui portent l'allèle A et G respectivement. A) le test pour le marqueur 1 est significatif car μG est significativement différente de μA, B) le test n'est pas significatif pour le marqueur 2.

Le séquençage du génome de référence du maïs B73 (Schnable et al., 2009) combiné avec le développement des nouvelles technologies de reséquençage haut débit (NGS) a permis d'accélérer le développement du marquage moléculaire. Cette combinaison a permis d'identifier plusieurs millions de SNP en alignant les séquences des différents génomes reséquencés sur la séquence de B73 (Gore et al., 2009; Chia et al., 2012; Brandenburg et al., 2017). Bien que de moins en moins couteux, le reséquençage demeurait trop couteux à mettre en œuvre sur des populations de très grande taille sur des génomes complexes comme le maïs. A partir de ces SNPs, une première puce de génotypage a ainsi été développée en 2011 à partir de la technologie Illumina® et permet aujourd'hui de génotyper des individus à partir de 50,000 SNP (Ganal et al., 2011). Plus récemment, une seconde puce de génotypage de 600K SNP a été développée à partir de la technologie Affymetrix® Axiom® (Unterseer et al., 2014). Parallèlement, une approche de génotypage par séquençage (GBS) basée sur un reséquençage très faible profondeur a été développée pour génotyper plusieurs centaines de milliers de SNP (Elshire et al., 2011). Ces technologies permettent de génotyper à moindre coût des panels de lignées afin d'étudier leur diversité génétique et d'identifier les régions génomiques d'intérêt, impliquées dans la variation des caractères et/ou dans l'adaptation.

La disponibilité d'outil de génotypage capable de génotyper plusieurs milliers de SNP a permis de génotyper des populations « naturelles » avec une plus grande diversité génétique composées d'individus a priori peu apparentés (Buckler and Thornsberry, 2002; Rafalski, 2002). Les panels de diversité permettent d'identifier plus précisément des QTL car les évènements de recombinaison « historiques » sont nombreux contribuant ainsi à réduire l'étendue du DL (Flint-Garcia et al., 2003). Contrairement aux populations biparentales, les croisements dans ces populations ne sont pas contrôlés et par conséquent la structuration et l'apparentement entre les individus sont inconnus et variables. De plus, le DL ne dépend pas uniquement de la recombinaison mais dépend aussi d'autres facteurs évolutifs (dérive, mutation, démographie). Il peut notamment être causé par l'appartenance d'individus à différents groupes génétiques dont les fréquences alléliques ont divergé du fait de la dérive et/ou la sélection, engendrant du DL entre des locus non liés génétiquement (Flint-Garcia et al., 2003). De même, l'apparentement au sein de la population peut engendrer du DL entre des locus éloignés génétiquement du fait que différentes régions génomiques peuvent dériver d'un même ancêtre commun au sein de groupes d'individus (Mangin et al., 2012). Par conséquent, l'apparentement et la structuration peuvent entrainer de fausses associations entre la variation de fréquence allélique et le phénotype causées par du DL entre des marqueurs non liés génétiquement. Chez le maïs, Camus-Kulandaivelu et al., (2006) ont montré que la structuration de la population pouvait expliquer jusqu'à 40% de la variation de date de floraison. De ce fait, tous les locus différentiés entre les groupes génétiques vont être identifiés comme associés à la floraison, même si ils retracent simplement l'effet de la structuration et ne sont pas associés à un facteur causal de la variation de date de floraison. Or, ce qui intéresse le généticien est la liaison génétique entre les loci impliqués dans la variation du caractère (facteur causal) et les marqueurs utilisés pour le cartographier grâce au DL. L'utilisation d'un modèle mixte intégrant la structuration et l'apparentement entre les individus d'une population étudiée semble aujourd'hui être une approche efficace pour contrôler le taux de fausses associations (Yu et al., 2006). La puissance en génétique d'association dépend de plusieurs paramètres : la taille du panel, l'effet du facteur causal, la fréquence allélique du facteur causal, la différentiation du facteur causal entre les groupes génétiques et le DL entre le marqueur et le facteur causal (si le marqueur n'est pas lui-même le facteur causal) (Huang and Han, 2014; Rincent et al., 2014; Nicolas et al., 2016). Chez le maïs, le génotypage haut débit de large panel de diversité avec plusieurs milliers de SNP a permis d'identifier de nombreux QTL par des approches de génétique d'association, notamment en lien avec la phénologie, l'architecture de la plante ou encore les composantes du rendement, à partir de la puce de génotypage 50K SNP (Li et al., 2013; Bouchet et al., 2013, 2017; Giraud et al., 2017b), 600K (Millet et al., 2016) ou grâce au génotypage GBS (Romay et al., 2013; Peiffer et al., 2014; Pace et al., 2015; Gouesnard et al., 2017).

La disponibilité d'un marquage moléculaire haut débit a aussi permis de développer des approches de « scan » génomique permettant d'identifier des profils spécifiques de diversité au sein du génome, appelés signatures de sélection. Ces profils peuvent être dus à l'augmentation de la fréquence d'un allèle favorable dans une population, qui entraine une réduction de la diversité nucléotidique dans son voisinage, on parle alors de sélection directionnelle (Nielsen, 2005). Ils peuvent également être dus à un phénomène de sélection balancée. Dans ce cas, il y a un avantage dans la population à maintenir un polymorphisme à un locus donné (Nordborg and Innan, 2002). Les signatures de sélection peuvent résulter de la domestication ou encore de l'adaptation à différentes contraintes biotiques ou abiotiques.

De nombreuses signatures de sélection ont été identifiées chez le maïs dans des panels de diversité grâce au génotypage de la puce 50K SNP (van Heerwaarden et al., 2012; Bouchet et al., 2013), de la puce 600K SNP (Unterseer et al., 2016), du GBS (Takuno et al., 2015; Gouesnard et al., 2017) et du reséquençage (Brandenburg et al., 2017; Romero Navarro et al., 2017). Ces travaux ont permis d'identifier et de mieux comprendre les mécanismes génétiques impliqués dans l'adaptation chez le maïs. Par exemple, la découverte de signatures de sélection a permis d'identifier des gènes impliqués dans la tolérance aux stress (van Heerwaarden et al., 2012) ou encore l'adaptation aux besoins des hommes, comme le gène Su1, responsable de la production de sucre chez le maïs doux (Brandenburg et al., 2017), mais aussi des gènes impliqués dans la variation de la précocité de floraison (Hung et al., 2012; Unterseer et al., 2016) comme le gène Vgt1 (Salvi et al., 2007). Bouchet et al., (2013) ont identifié 10 loci sous sélection avec un effet significatif sur la floraison, ainsi que 18 loci associés à la variation de la floraison grâce à une approche de génétique d'association. Bien que les approches de génétique d'association permettent de directement connecter le génotype et le phénotype nous avons vu plus haut qu'il était difficile d'identifier des gènes différentiés au sein des groupes génétiques avec cette approche. Cependant, l'identification de signatures de sélection permet d'identifier ces gènes différentiés, mais il est alors plus difficile de connecter cette information avec le phénotype.

Les variations structurales

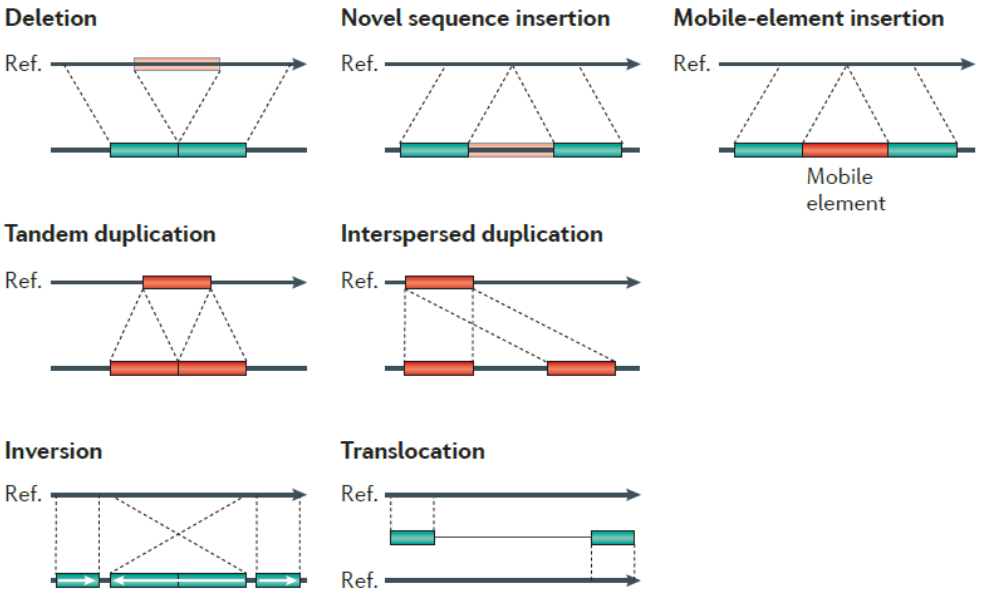


Figure 5 : Représentation schématique de différents types de variations structurales. Pour chaque évènement, un génome « test » est comparé à un génome de référence « Ref. » (Alkan et al., 2011).

Chez le maïs, le développement du génotypage à partir de milliers de SNP a fortement contribué à approfondir les connaissances sur les régions génomiques impliquées dans l'architecture génétique des caractères ainsi que des régions génomiques sous sélection potentiellement impliquées dans l'adaptation à des environnement contrastés ou à des besoins humains spécifiques. En plus du polymorphisme de type

SNP, de nombreuses variations structurales (SV) ont été identifiées chez les plantes (Cao et al., 2011; Saintenac et al., 2011; Causse et al., 2013; Anderson et al., 2014; Hardigan et al., 2016; Varshney et al., 2017; Hurgobin et al., 2018; Owens et al., 2018) et notamment chez le maïs (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010; Liu et al., 2015). L'intérêt porté aux SV chez les plantes est grandissant depuis quelques décennies, bien que ces évènements soient connus depuis longtemps. En 1924, Sturtevant découvrait que la forme de l'œil chez la drosophile était due à une duplication du gène *bar*. Des réarrangements chromosomiques ont été mis en évidence par des approches cytogénétiques bien avant les technologies de séquençage des génomes, notamment chez l'humain (Bobrow et al., 1971 ; Jacobs et al., 1978 ; Jacobs et al., 1992) mais aussi chez les plantes (Rieseberg, 2001; Pires et al., 2004). L'avancée des méthodes de séquençage, a permis de découvrir de très nombreuses variations structurales entre les individus d'une même espèce (Xu et al., 2012; Mace et al., 2013; Pinosio et al., 2016; Varshney et al., 2017; Jiao et al., 2017; Zhou et al., 2017; Sun et al., 2018). Ces modifications de l'ADN de quelques paires de base (bp) à plusieurs centaines de milliers peuvent être de différents types (Figure 5): (i) les inversions, qui entrainent un changement dans l'orientation des séquences à un locus, (ii) les translocations qui entrainent le déplacement d'une séquence d'un locus à un autre, (iii) des variations du nombre de copie entre les individus, ainsi que (vi) les insertions et délétions de séquences (InDel) (Alkan et al., 2011). Les InDel sont des séquences présentes ou absentes à un locus chez un individu « test » par rapport à un individu de référence (Figure 5). Ces séquences peuvent être présentes ailleurs dans le génome de l'individu analysé ou totalement absentes de ce génome, on parle alors dans ce second cas des variations de type présent/absent (presence absence variation, PAV). La longueur minimale d'une variation structurale était historiquement de 1 kb, car il était difficile de découvrir des évènements plus petits (Pinkel et al., 1998). Cependant, le séquençage des génomes permet aujourd'hui d'identifier de nombreuses SV de quelques dizaines de paires de base (Hastings et al., 2009; Alkan et al., 2011).

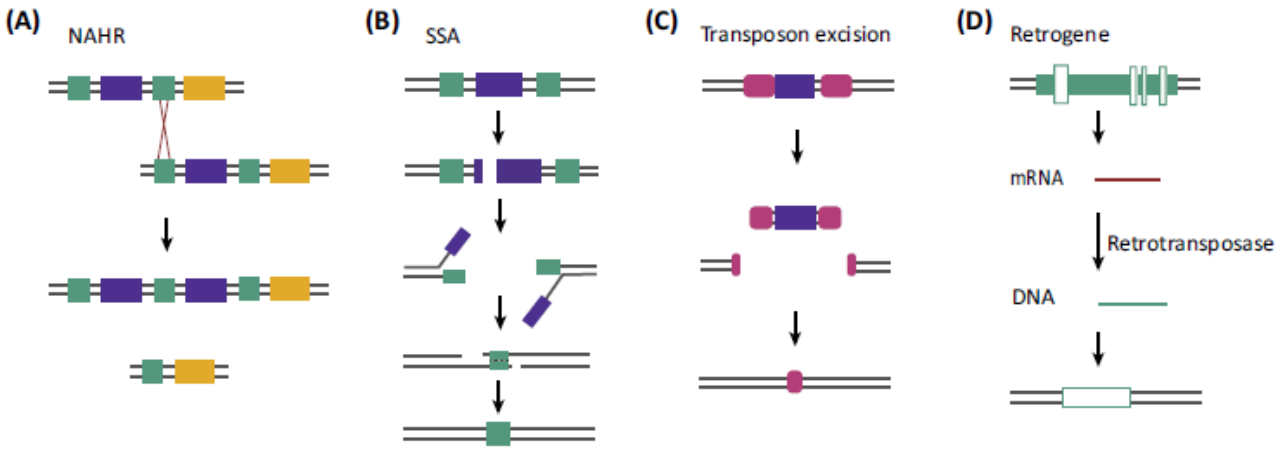


Figure 6 : Représentation schématique des principaux mécanismes de formations des InDel. (A) "Nonallelic homologous recombination" ou crossing-over inégaux, (B) « microhomology-mediated end-joining » (MMEJ), (C) éléments transposables, (D) retro-transposition. D'après Lye et Puruggunan (2019)

Les variations de type présence/absence peuvent avoir plusieurs origines (Figure 6). Premièrement, elles peuvent provenir de crossing-over inégaux entre des régions avec une forte similarité (Hastings et al., 2009) ou de recombinaisons entre régions homologues non-alléliques (Liu et al., 2012; Hurgobin et al., 2018). D'autres mécanismes impliqués dans la réparation de l'ADN peuvent également être à l'origine des InDel comme les « non-homologous end-joining » (NHEJ) ou « microhomology-mediated end-joining » (MMEJ). Les erreurs de réplication d'ADN peuvent être à l'origine des InDel comme les « Fork Stalling and Template Switching » (FoStes) ou les « Microhomology-mediated Break-induced Replications » (MMBIR) (Hastings et al., 2009). Chez l'orge, l'examen des régions flanquantes de 299 InDel, a révélé des séquences signatures qui suggèrent des réparations des cassures double brins par l'insertion d'une séquence de petite taille qui borde exactement le point de cassure ou par l'insertion d'une séquence non-homologue (Muñoz-Amatriaín et al., 2013). Enfin, les éléments transposables (ET) peuvent également dupliquer et/ou déplacer des séquences d'ADN à la manière d'un copier-coller ou couper-coller (McClintock, 1950; Lisch, 2013) et donc être à l'origine de nombreuses variations structurales (Hastings et al., 2009; Conrad et al., 2010). Les ET sont très répandus chez le maïs, couvrant environ 85% du génome (Schnable et al., 2009) et sont en partie responsables des variations de taille des génomes entre les différents individus (Tenaillon et al., 2011). Fu et Dooner (2002) puis Lai et al. (2005) ont fait l'hypothèse que la délétion de la région bz chez le maïs était due à un mouvement de gènes causé par des helitrons, une classe spécifique d'ET. Les espèces anciennement polyploïdes comme le maïs ont connu un phénomène de fractionnement du génome (Langham et al., 2004; Thomas, 2006). Lors de cet évènement, une partie des gènes dupliqués lors de l'événement de polyploïdisation a été supprimée. Chez le maïs, le fractionnement du génome aurait potentiellement contribué aux variations structurales (Schnable et al., 2011; Sun et al., 2018).

Chez le maïs, Fu et Dooner (2002) ont reséquencé par une approche de « Bacterial Artificial Chromosome » (BAC) une portion de 230 Kbp d'une lignée de maïs nord-américaine McC. Après comparaison avec la même portion de génome chez B73 (une autre lignée nord-américaine) ils ont identifié une différence de contenu en gènes et éléments transposables entre les deux lignées. Quatre gènes ont été identifiés comme supprimés de cette région chez B73 par rapport à McC. Malgré la délétion de la séquence entière, 3 de ces 4 gènes restent malgré tout exprimés ce qui suggère que ces gènes sont présents ailleurs dans le génome chez Mo17 et non totalement absents. La disponibilité du génome de référence du maïs en 2009 (Schnable et al., 2009) a permis d'explorer ces variations structurales à l'échelle du génome entier. La méthode de « array comparative genomic hybridization » (aCGH) (Pinkel et al., 1998) permet de comparer l'intensité de fluorescence de plusieurs milliers de sondes d'un individu par rapport à une référence. Cette méthode permet d'identifier à la fois une augmentation ou une diminution du nombre de copies d'une séquence par rapport à une référence, allant jusqu'à l'absence de cette séquence. La comparaison de plusieurs dizaines de lignées de maïs au génome de référence B73 avec une puce CGH a permis d'identifier plusieurs milliers de CNV et PAV (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010). Le reséquençage et l'assemblage de plusieurs génomes d'autres lignées que B73 a permis d'étendre la découverte d'InDel en utilisant comme références d'autres génomes que B73. Lai et al. (2010) ont reséquencé 6 lignées de maïs qu'ils ont comparé au génome de référence et ont identifié 296 gènes présents chez B73 et absents d'au moins une des six lignées ainsi que 570 gènes présents dans

au moins une des 6 lignées mais absents dans B73. Plus récemment, le reséquençage et l'assemblage de plusieurs lignées a permis la découverte de plusieurs milliers d'InDel grâce à des outils de bio-informatique capables de comparer les séquences entre les différents génomes assemblés. 2 714 PAV ont été identifiés entre la lignée élite américaine PH207 et B73 (Hirsch et al., 2016) ainsi que 10 735 PAV entre la lignée fondatrice européenne F2 et B73 (Darracq et al., 2018). Enfin, le reséquençage des lignées américaines W22 et Ki11 et la comparaison avec B73 a permis de découvrir 6 706 PAV (Jiao et al., 2017) et 25 875 PAV entre Mo17 et B73 (Sun et al., 2018). L'étude de l'annotation des gènes impactés par ces InDel montrent un enrichissement en gènes liés à la tolérance aux stress biotiques et abiotiques suggérant que les InDel pourraient être impliquées dans l'adaptation (Swanson-Wagner et al., 2010; Hirsch et al., 2016; Jiao et al., 2017; Darracq et al., 2018).

Des approches gènes candidats ont permis d'identifier des variations structurales qui colocalisent avec des gènes, notamment lié aux résistances aux maladies et tolérance aux stress chez plusieurs espèces, notamment chez l'orge (Muñoz-Amatriaín et al., 2013), le colza (Hurgobin et al., 2018), et le soja (McHale et al., 2012). Un lien entre l'augmentation du nombre de copies d'un gène et le phénotype a été observé chez l'orge notamment pour la tolérance au froid (Francia et al., 2016) et la précocité (Nitcher et al., 2013), la résistance aux nématodes chez le soja (Cook et al., 2012), la résistance au glyphosate chez l'amarante (Gaines et al., 2010) ainsi que la tolérance à l'aluminium chez le maïs (Maron et al., 2013). Ces études suggèrent que les variations du nombre de copies pourraient être impliquées dans l'adaptation à des stress biotiques et abiotiques. Le reséquençage de 292 lignées de pois d'Angole a permis d'identifier 68 CNV et 1 PAV potentiellement sous sélection (Varshney et al., 2017). De plus, la moitié des CNV trouvés chez l'orge sont présents uniquement chez ses ancêtres sauvages suggérant que la domestication de l'orge a affecté la diversité structurale de l'espèce (Muñoz-Amatriaín et al., 2013). Chez le riz, de nombreuses séquences sont présentes uniquement chez les individus sauvages et absentes des lignées cultivées ce qui suggèrent un effet des polymorphismes présent / absent sur la domestication de l'espèce (Monat et al., 2018). Bien que les InDel, soient largement répandues chez le maïs et de nombreuses plantes, peu d'outils existent pour génotyper à haut débit et pour un cout raisonnable des dizaines de milliers d'InDel sur des centaines d'individus, et notamment des séquences non présentes dans le génome de référence. Ainsi, il est aujourd'hui difficile de réaliser des études génétiques sur le génome entier avec des InDel de grande taille, telles que l'indentification de signatures de sélection ou encore la génétique d'association. Pour résoudre ce problème, Chia et al., (2012) puis Lu et al., (2015) ont reséquencé à faible profondeur plusieurs milliers d'individus et découvert des SNP ainsi que des séquences d'ADN qui ne s'alignaient pas sur le génome de référence, estimant alors que ces séquences étaient des InDel. Ils ont réalisé une étude d'association entre les SNP et plusieurs caractères (relatifs à l'architecture de la plante, la résistance aux maladies ou encore la date de floraison) et ont identifié un enrichissement en associations pour les SNP en fort déséquilibre de liaison avec ces InDel. Parallèlement aux InDel, les TE ont été identifiés comme associés aux variations phénotypiques chez les plantes comme le riz, la tabac ou encore le maïs, avec notamment un effet sur la réponse aux stress biotiques et abiotiques (Zerjal et al., 2012; Makarevitch et al., 2015; Grandbastien, 2015). De nombreux auteurs ont de plus fait l'hypothèse que la complémentation des gènes présents/absents entre les lignées parentales des hybrides de maïs pouvait être à l'origine d'une partie de la vigueur hybride (Fu and Dooner, 2002; Beló et al., 2010; Sun et al., 2018).

L'étude de la valeur hybride chez le maïs

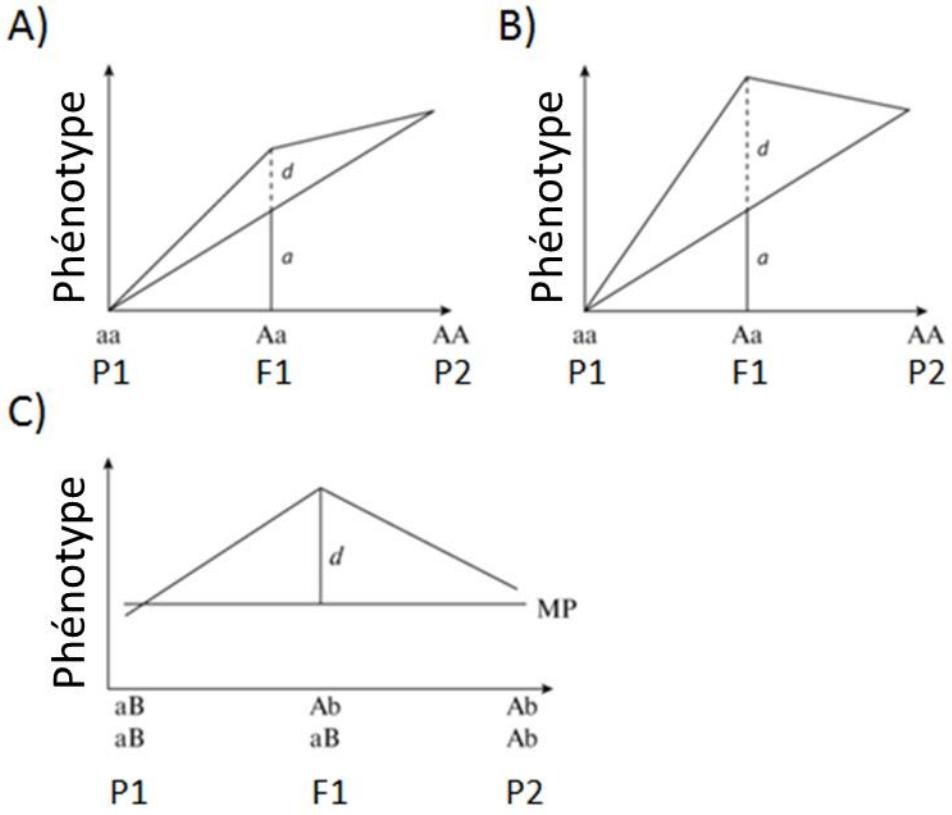


Figure 7 : Mécanismes génétiques responsables de l'hétérosis pour A) la dominance, B) la superdominance et C) la pseudo superdominance. P1 et P2 correspondent au génotype du parent 1 et du parent 2 et F1 le génotype de l'hybride, *a* l'additivité, *d* la dominance et MP la moyenne des parents. D'après Khotyleva et al. (2017)

Malgré de nombreuses études, les bases génétiques de l'hétérosis restent encore en partie à élucider. Plusieurs mécanismes pourraient expliquer l'hétérosis : la dominance, la super dominance, la pseudo-superdominance et l'épistasie mais, leurs contributions respectives au phénomène d'hétérosis restent difficiles à évaluer (Springer and Stupar, 2007). Dans le cas de la dominance (Figure 7 A), la présence d'un allèle a délétère chez un parent est complémentée chez l'hybride par l'allèle dominant A de l'autre parent. Dans ce cas, la valeur de l'hybride (aA) est égale à la valeur de l'homozygote AA. Dans le cas de la super dominance (Figure 7 B), l'état hétérozygote (aA) à un locus entraine une supériorité en comparaison des parents (aa et AA). La superdominance (Figure 7 C) est souvent confondue avec la pseudo-superdominance qui est une situation de dominance complète mais, qui a lieu entre deux gènes fortement liés en répulsion où un gène favorable à un locus est associé à un gène défavorable à l'autre (Stuber et al., 1992). Enfin, le phénomène d'épistasie peut aussi expliquer une part de l'hétérosis due à des interactions entre les gènes à deux ou plusieurs loci (Springer and Stupar, 2007).

Des études ont identifié des QTL impliqués dans l'architecture génétique des performances hybrides. L'intérêt de l'évaluation phénotypique des hybrides par rapport aux lignées est de pouvoir détecter des QTL avec un effet de dominance, de superdominance ou pseudo-superdominance. Dans ce sens, des QTL avec un effet de dominance et superdominance (ou pseudo-superdominance) ont été identifiés pour le rendement en grain (Frascaroli et al., 2009; Larièpe et al., 2012). Des interactions épistatiques significatives ont également été trouvées, en particulier pour la taille de la plante, l'humidité du grain et la floraison (Frascaroli et al., 2009; Larièpe et al., 2012). Enfin, des QTL avec un effet additif ont aussi été découverts, notamment pour des caractères comme la taille, la floraison (Larièpe et al., 2012), mais aussi des caractères de rendement en biomasse (Giraud et al., 2017a).

L'intérêt du sélectionneur est de pouvoir prédire les performances des individus, soit en identifiant des QTL, soit en considérant l'ensemble du génome pour prédire les valeurs phénotypiques en fonction de l'apparentement entre les individus (Bernardo, 1992, 1994). Ces valeurs d'apparentement peuvent être calculées soit à partir du pédigrée des individus, soit des données moléculaires (VanRaden, 2008; Maenhout et al., 2010). En utilisant une seule matrice d'apparentement, la variance génétique due à la dominance peut-être partiellement intégrée (Varona et al., 2018), mais il n'est pas possible de distinguer les effets génétiques additifs et des effets de dominance. Su et al. (2012) ont proposé une méthode pour décomposer les effets génétiques additifs et de dominance. Cette approche a permis d'identifier des effets de dominances significatifs sur le phénotype chez les animaux (Toro and Varona, 2010; Vitezica et al., 2016; Xiang et al., 2016) et chez les plantes (Piaskowski et al., 2018). En plus de l'effet additif, l'intégration de l'effet de dominance dans les modèles a permis d'améliorer les prédictions des performances hybrides chez le maïs (Technow et al., 2012; dos Santos et al., 2016). Outre l'effet polygénique additif et de dominance basé sur les seuls SNP, Lyra et al., (2018) ont ajouté un effet polygénique additif basé évalué le génotypage de quelques centaines de CNV. Leurs résultats montrent que cet ajout ne permet pas d'améliorer significativement les résultats de prédiction. Cependant, ils ont trouvé une corrélation positive entre la taille des hybrides et le nombre de copies présentes dans ces derniers, ce qui suggère un effet quantitatif du nombre de séquences présentes chez les hybrides sur le phénotype.

Objectif de la thèse

Le reséquençage des génomes et leur comparaison montrent que les InDel sont largement répandues chez le maïs et peuvent entrainer la présence et l'absence de gènes. Cependant, la contribution des polymorphismes de type présence/absence à la diversité génétique, l'adaptation, la variation du phénotype des lignées et des hybrides chez le maïs reste encore mal connue. Ce manque de connaissances est en grande partie dû à l'absence d'outil pour génotyper à haut débit des milliers d'insertions et de délétions par rapport au génome de référence de grande taille chez des centaines d'individus de façon fiable.

Dans ce contexte, l'objectif de ma thèse était de déterminer la contribution de plusieurs milliers d'InDel de grande taille à l'adaptation du maïs à des environnements contrastés ainsi qu'à la variation des caractères agronomiques et aux performances hybrides en décomposant les effets additifs et dominants.

Ma thèse s'est alors déroulée en trois parties :

1) Evaluer les propriétés d'une nouvelle puce de génotypage permettant de génotyper précisément et à haut débit la présence et l'absence de milliers d'InDel de toutes tailles, notamment des insertions par rapport au génome de référence, dans des larges panels d'individus (Chapitre 2)
2) Etudier le déséquilibre de liaison entre les InDel et les SNP, ainsi que la contribution des InDel à la diversité génétique, l'adaptation et à la variation des caractères agronomiques à partir d'une population de lignées de maïs par une approche de génétique d'association et de détection de signatures de sélection (Chapitre 3)
3) Etudier la contribution des InDel à la variation des caractères dans un contexte hybride par une approche de génétique d'association et de prédiction génomique (Chapitre 4)

Ces trois parties sont respectivement traitées dans les chapitres 2, 3 et 4. La stratégie expérimentale est présentée dans le chapitre 1. Je discuterai ensuite l'ensemble des résultats obtenus au cours de ma thèse dans la discussion générale.

Chapitre 1

Stratégie expérimentale

Je vais présenter dans ce chapitre la stratégie expérimentale qui m'a permis de répondre aux questions que je me suis posées durant ma thèse :

1) Quelle est la contribution des InDel à l'apparentement et à la structuration des populations chez le maïs ?
2) Quelle est la contribution des InDel à l'adaptation du maïs à des environnements contrastés ?
3) Quelle est la contribution des InDel à la variation des caractères et aux performances hybrides chez le maïs ?

Afin de répondre à ces questions, nous avons phénotypé deux panels de diversité composés de 375 lignées et de 284 hybrides et nous les avons génotypés au moyen de 1 millions de SNP provenant de différentes technologies et avec une nouvelle puce de génotypage haut débit des InDel (Chapitre 2). Le premier panel, composé de 375 lignées de maïs représentant une large gamme de diversité génétique, a été évalué pour 23 caractères en valeur propre. Ce panel a permis de comparer la contribution respective des SNP et des InDel à la diversité génétique, à l'adaptation via l'étude des signatures de sélection et la variation des caractères agronomiques en valeur propre par une approche de génétique d'association (Chapitre 3). Le second est un panel de 284 hybrides qui dérive du croisement de 210 lignées tempérées du premier panel. Ce panel d'hybrides m'a permis de mieux appréhender le rôle des InDel dans les performances hybrides avec des approches de génétique d'association et de prédiction génomique (Chapitre 4).

Etude d'un panel d'association de 375 lignées de maïs

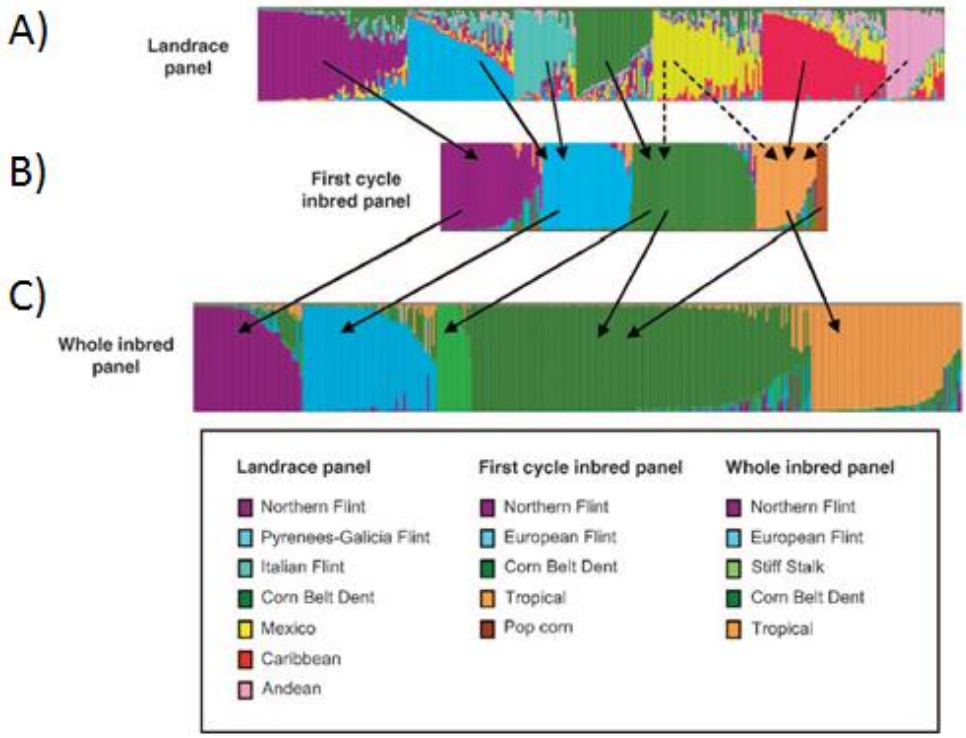


Figure 1.1 : Représentation schématique de la structuration génétique à trois étapes de la construction du panel de lignées : A) populations, B) des lignées dérivées des populations et C) du panel complet. Les différents groupes génétiques sont représentés par des couleurs. Chaque lignée est représentée par une barre verticale, la proportion de chaque couleur dans cette barre indique la proportion d'assignation au groupe génétique correspondant. Les flèches en pointillées indiquent que les groupes ont faiblement contribué (moins de 3 lignées). D'après Camus-Kulandaivelu et al. (2006).

Le premier objectif de ma thèse était d'étudier la contribution des InDel à l'apparentement entre les individus, à la structuration de la population et à l'adaptation et à la variation des caractères. Pour répondre à cet objectif, j'ai utilisé un panel de 375 lignées de maïs qui regroupe une très large gamme de diversité génétique avec des lignées cultivées sur une large gamme de conditions agro-climatiques (figure 1.1). L'expansion du maïs vers le nord a entrainé la création d'un type de maïs avec un cycle de floraison court et adapté à des températures plus froides appelé « Northern Flint » (NF). Ce type est très divergent génétiquement des maïs tropicaux cultivés près de la région d'origine du maïs (Doebley et al., 1986). Le croisement des types tropicaux et NF a permis il y a environ 200 ans de créer un nouveau type, « Corn Belt Dent » (CDB), adapté aux conditions de culture du Midwest américain (Doebley et al., 1988; Brandenburg et al., 2017). L'introduction du maïs en Europe a eu lieu peu de temps après la découverte des Amériques par Christophe Colomb. Deux types auraient alors été introduits en Europe, des NF au nord et des tropicaux au sud. Le croisement de ces différents types et l'adaptation au climat européen ont entrainé la création d'un nouveau type, les « European Flint » (EF) au niveau de la région Pyrénées-Galice (Rebourg et al., 2003; Dubreuil et al., 2006).

Ce panel a été assemblé dans notre laboratoire pour représenter la diversité de l'ensemble de ces groupes génétiques à partir de 153 lignées de maïs directement issues de de 275 populations de maïs auxquelles s'ajoutent 222 lignées issues des programmes de sélection plus avancés (Figure 1.1), notamment des lignées de type « Stiff Stalk » (SS) provenant du croisement des meilleures lignées issues des programmes de sélection américains. L'analyse de ce panel avec d'abord des marqueurs SSR puis des SNPs a permis de mettre en évidence la structuration de ce panel (Camus-Kulandaivelu et al., 2006; Bouchet et al., 2013). Ces études ont retrouvé les cinq groupes génétiques majoritaires présentés précédemment à savoir les EF, NF, CDB, SS et une partie des Tropicaux. La structuration, étudiée à partir de 55 marqueurs SSR a permis de montrer que 40% de la variation de la floraison dans ce panel était due à la structuration avec les lignées des groupes EF et NF les plus précoces et les lignées tropicales les plus tardives (Camus-Kulandaivelu et al., 2006).

Ce panel a été phénotypé pour 24 caractères (phénologie, architecture de la plantes et composantes du rendement) sur trois lieux : Einbeck (Allemagne), Gif sur Yvette et Saint Martin de Hinx (France). Les lignées des deux groupes les plus tardifs ont également été phénotypées à Mauguio, dans le sud de la France, car leur cycle ne leur permettait pas de produire des grains jusqu'à maturité avant l'hiver pour les autres lieux. Les expérimentations françaises ont eu lieu sur deux ou trois ans entre 2002 et 2004 et seulement en 2005 en Allemagne, totalisant 9 environnements (lieux x années). Pour chaque caractère, un phénotypage en moyenne ajusté a été calculé (Bouchet et al., 2013, 2017).

Une première approche gène candidat a été effectuée à partir de ce panel pour tester l'association entre des marqueurs moléculaires et la variation de floraison dans la région du gène *Vgt1* (Ducrocq et al., 2008). Ces travaux ont permis de montrer le rôle du gène *Vgt1* dans l'adaptation du maïs aux régions tempérées. Une seconde approche gène candidat a été réalisée sur ce panel pour tester l'association de 2 gènes (*CyPPDK1* and *Opaque2*) aux caractères de qualité de la graine (Manicacci et al., 2009). Le développement du génotypage haut débit a ensuite permis de génotyper ce panel avec plusieurs dizaines de milliers de SNP grâce à la puce de génotypage 50K SNP Illumina® (Ganal et al., 2011), ce qui a permis de calculer l'étendue du DL dans cette population, ainsi que l'identification de QTL pour des caractères en lien avec la phénologie, l'architecture de la plante et les composantes du rendement (Bouchet et al., 2013, 2017).

Nous avons choisi ce panel car les études précédentes ont montré qu'il est approprié pour réaliser des études de diversité, pour détecter des signatures de sélection et des régions associées aux variations phénotypiques par génétique d'association. Nous avons génotypé ce panel avec 61 492 InDel grâce à notre nouvelle puce de génotypage des InDel présentée en chapitre 2 pour premièrement, évaluer la contribution des InDel à la structuration génétique du panel, ainsi qu'à l'apparentement entre les individus. En plus du génotypage des InDel, j'ai également assemblé le génotypage de 34 964, 483 657 et 486 475 SNP provenant respectivement de la puce SNP 50K Illumina® (Ganal et al., 2011), la puce 600K SNPs Affymetrix® Axiom® (Unterseer et al., 2014) et du génotypage par séquençage (Elshire et al., 2011). En plus de ces SNP, nous avons intégré le génotypage de 6453 petits InDel (de 1 à 4bp) ainsi que 65 243 marqueurs de type « Off Target Variant » (OTV) issus de la puce SNP 600K. Ces OTV sont des SNP pour lesquels la sonde ne s'est pas hybridée chez certains individus à cause soit d'un polymorphisme dans la séquence de la sonde soit de son absence chez certains individus (Didion et al., 2012). L'ensemble de ces

données représente un total de 1 076 792 marqueurs auxquels nous avons ajouté le génotypage du Mite Vgt1 (Salvi et al., 2002, 2007; Ducrocq et al., 2008; Castelletti et al., 2014) ainsi que l'InDel Dwarf8 (Thornsberry et al., 2001; Camus-Kulandaivelu et al., 2006). J'ai ensuite détecté des locus sous sélection entre les groupes et des signatures de sélection grâce à un modèle bayésien implémenté dans le logiciel Bayescan (Alexander et al., 2009). Ce panel a été évalué pour 23 caractères en lien avec la phénologie, l'architecture et les composantes du rendement, ce qui permet de tester l'effet des InDel sur l'ensemble de ces caractères par génétique d'association. Enfin, la comparaison entre les résultats d'analyses entre les InDel et les SNP m'a permis d'évaluer l'intérêt d'intégrer les InDel issues d'une nouvelle puce par rapport aux marqueurs issus des technologies existantes.

Etude d'un panel de 287 hybrides de maïs

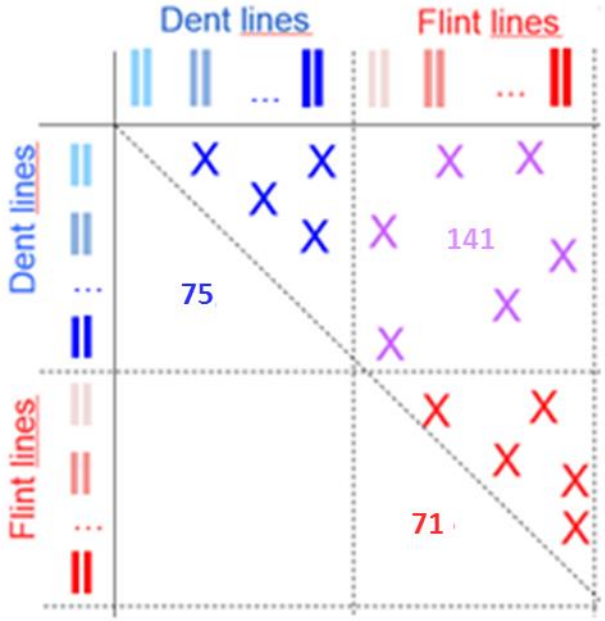


Figure 1.2 : Représentation schématique de la construction du panel de 287 hybrides à partir du croisement de 210 lignées issues des groupes génétiques denté (Dent) et corné (Flint). 75 hybrides sont issus du croisement deux lignées dentées, 71 issus de deux lignées cornées et 141 issus du croisement d'une lignée cornée et une lignée dentée.

Le dernier objectif de ma thèse était d'étudier la contribution des InDel à la variation des performances hybrides. Pour répondre à cet objectif, j'ai utilisé un panel d'hybrides issu d'un plan de croisement diallèle incomplet entre 210 lignées (109 cornées et 101 dentées) issues du panel présenté ci-dessus. Ce panel est composé de 287 hybrides dont 75 hybrides issus du croisement de deux lignées cornées, 141 hybrides issus du croisement d'une lignée cornée et d'une lignée dentée et 71 hybrides issus du croisement de deux lignées dentées (Figure 1.2).

Les hybrides ont été évalués dans six lieux, divisés en deux zones : Nord et Sud. Afin de placer les hybrides dans des conditions favorables pour exprimer au mieux leur potentiel, leur gamme de floraison a été estimée à partir de la date de floraison moyenne de leurs parents. Les 116 hybrides estimés les plus précoces ont été évalués dans la zone Nord tandis que les 110 hybrides estimés les plus tardifs ont été évalués dans la zone Sud. 58 hybrides avec des dates de floraison intermédiaires ont été évalués dans les deux zones. Deux lieux composent la zone nord : Mons (MONS) et Le Moulon (ML) et deux lieux composent la zone sud : Lusignan (LS) et Satolas (SAT). L'ensemble des hybrides nord et sud ont été également évalués dans un 5ème lieu : Saint Martin de Hinx (SMH). Les essais ont été réalisés en 2009 et 2010 ce qui représente un total de 10 environnements (lieux x années). Six caractères ont été mesurés : la hauteur (HEIGHT), les dates de floraison mâle et femelle (FLOM et FLOF), l'humidité des grains (HUM), le poids de mille grains (PMG) et le rendement à 15% d'humidité a été calculé (RDT15) à partir du poids frais récolté.

Les 287 hybrides de ce panel n'ont pas été directement génotypés. Comme les hybrides de ce panel dérivent du croisement de lignées dentées et cornées du panel de lignées présenté ci-dessus, nous avons reconstruit le génotypage des hybrides à partir de celui des lignées parentales pour l'ensemble des marqueurs SNP et InDel. Nous avons donc attribué aux hybrides un génotypage en fonction de la combinaison allélique de leurs parents.

Nous avons pu tester l'effet de l'année sur les résultats de génétique d'association en calculant une moyenne ajustée des performances hybrides pour SMH 2009 et 2010, car tous les hybrides ont été évalués sur ce lieu. Nous avons également calculé une moyenne ajustée des phénotypes pour l'ensemble des environnements. Excepté pour SMH, nous n'avons pas pu réaliser les analyses de génétique d'association et de prédictions génomiques pour chaque lieu car le nombre d'hybrides était trop limité du fait de la séparation nord et sud.

Ce panel d'hybrides nous a permis d'étudier la contribution des InDel aux performances hybrides, mais aussi de tester l'hypothèse de la complémentation des InDel dans les hybrides à travers trois approches. Premièrement, nous avons décomposé la variance de trois caractères (floraison femelle, la hauteur et le rendement), en distinguant notamment les effets génétiques d'additivité et de dominance grâce à l'approche de Vitezica et al. (2016). Ensuite, nous avons réalisé une étude d'association afin de détecter des QTL avec des effets additifs ou dominants. Enfin, nous avons testé l'effet de l'inclusion du génotypage des InDel par rapport à celui des SNP pour prédire les performances hybrides.

Chapter 2

High throughput genotyping of structural variations in a complex plant genome using an original Affymetrix® Axiom® array

Mabire Clément[2*], Duarte Jorge[1*], Aude Darracq[2], Ali Pirani[3], Hélène Rimbert[1,4], Delphine Madur[2], Valérie Combes[2], Clémentine Vitte[2], Sébastien Praud[1], Nathalie Rivière[1], Johann Joets[2], Jean-Philippe Pichon[1], Stéphane D. Nicolas[2]

*Two authors contributed equally to work

Authors Affiliation:

1 Biogemma - Centre de Recherche de Chappes, CS 90126, Chappes 63720, France

2 GQE – Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

3 Thermo Fisher Scientific - 3450 Central Expy, Santa Clara, CA, 95051, USA

4 Present adress : GDEC, INRA, Université Clermont Auvergne, 63000 Clermont-Ferrand, France

Corresponding authors: stephane.nicolas@inra.fr

Key words: Present Absent Variant, Copy Number Variation, Structural Variation, genotyping, array, Zea mays, Genome assembly, Breakpoint, Chromosomal rearrangements

This paper was submitted for publication on February 2019 and is available in BioRxiv: https://doi.org/10.1101/507756

Abstract

Background

Insertion/deletion variants (InDels), and more specifically presence/absence variants (PAVs) are pervasive in maize and have strong functional and phenotypic effect by removing or modifying drastically genes. Genotyping of such variants on large panels remains poorly addressed, while necessary for approaches such as association mapping or genomic selection.

Results

We have developed a new high throughput and cost-effective tool to genotype InDel. We first identified 141,000 InDels by aligning reads from the B73 line against the genome of three temperate maize inbred lines (F2, PH207, and C103) and reciprocally. Next, we designed an Affymetrix® Axiom® array to target these InDels with a combination of probes selected at breakpoint sites (13%) and/or within the InDel sequence either at polymorphic (25%) or non-polymorphic sites (63%) sites. The final array design is constituted of 662,772 probes and targets 105,927 InDels including PAVs, ranging from 35bp to 129kbp. After Affymetrix quality control, we successfully genotyped 89,393 InDels (84%) on 445 maize DNA samples with 479,027 probes (72%). A principal coordinate analysis on dissimilarity estimated from a subset of 57,824 InDels on 362 inbred lines is consistent with the structure obtained using 50K SNP arrays.

Conclusions

We efficiently genotyped thousands of small to large InDels on a large number of individuals using a new Affymetrix® Axiom® array. This powerful tool opens the way to studying the contribution of InDels to trait variation and heterosis in maize. The approach is easily extendable to other species and should contribute to decipher the biological impact of InDels at a larger scale.

Introduction

In the past decade, there has been growing evidence that structural variations (SVs) are pervasive within plant genomes (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010; Cao et al., 2011; Saintenac et al., 2011; Anderson et al., 2014; Saxena et al., 2014; Liu et al., 2015; Owens et al., 2018). Insertions/deletions (InDels) are one class of SVs of particular interest since they lead to the presence or absence of, sometimes large, genomic regions at a given locus, among individuals from the same species. The content of these InDels can either be present elsewhere in the genome, but can also be completely absent from the genome, in which case they are referred to as presence/absence variants (PAVs). Some InDels carry entire genes or affect gene regulatory elements, and are thus likely to have a functional and phenotyping impact (Chia et al., 2012; Mace et al., 2013; Hirsch et al., 2014; Saxena et al., 2014; Lu et al., 2015). Hundreds to thousands of SVs, including PAVs and copy number variations (CNVs), have been discovered in several plant species, including wheat (Montenegro et al., 2017), rice (Zhao et al., 2018), *Arabidopsis thaliana* (Lu et al., 2012), Potato (Hardigan et al., 2016), pigeon peas (Varshney et al., 2017), and Sorghum (Shen et al., 2015). These results support the idea that one single reference genome cannot properly represent the complete gene set of a given species. There has been an increasing interest for building new individual genomes in complement to the reference genome, in order to better describe the genetic diversity within a plant species (Cao et al., 2011; Hirsch et al., 2016; Pinosio et al., 2016; Jiao et al., 2017; Varshney et al., 2017; Zhou et al., 2017; Appels et al., 2018; Sun et al., 2018; Darracq et al., 2018).

In maize, BAC sequence comparison first revealed that gene and transposable element content greatly vary between inbred lines (Fu and Dooner, 2002; Brunner, 2005). Whole genome sequencing of the B73 inbred line then provided the opportunity to explore the extent of SVs across the entire maize genome (Schnable et al., 2009) by designing Comparative Genomic Hybridization (CGH) technology (Pinkel et al., 1998). Several CGH studies found multiple CNVs between the B73 reference genome and other maize inbred lines or teosintes (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010). These studies demonstrated the large extent of SVs among maize inbred lines, including presence/absence variations of low copy sequences such as genes. This was well illustrated by the discovery of a large 2 Mbp presence/absence region between Mo17 and B73 carrying several genes (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010; Hirsch et al., 2016). However, CGH array technology shows several major drawbacks since (i) it does not allow the discovery of sequences that are not present in the reference genome used for designing probes of the arrays, (ii) it has a limited resolution which does not allow detection of InDels smaller than 1kb, and (iii) it is costly and labor-intensive, and therefore not adapted for genotyping several hundreds of individuals.

Methods based on SNP array experiments have been developed to detect CNVs and were shown to be more cost effective and with higher throughput but to reduce breakpoint resolution than CGH arrays (Cooper et al., 2008; Dellinger et al., 2010; Wang et al., 2017). Didion et al. (2012) identified atypical patterns of reduced hybridization intensities that were highly reproducible, so called “off-target variants” (OTVs). OTV patterns could originate either from the absence of the sequence due to a PAV polymorphism, or to a single nucleotide polymorphism within the probe sequence, thus preventing the

correct hybridization of the DNA sample. For instance, 45,974 OTVs were discovered in a maize population using the 600K Affymetrix AXIOM SNP array (Unterseer et al., 2014). While these approaches proved to be useful, there is a strong risk of false positive detection of PAVs using OTV patterns, mainly because these arrays were not designed to target PAVs. In order to reduce this risk of false positive detection of PAVs and more largely CNVs, several methods based either on segmentation or Hidden Markov Chain have been developed to use variation of fluorescent intensity signal of contiguous probes along the genome (Hupe et al., 2004; Olshen et al., 2004; Picard et al., 2005, 2007). These kinds of approaches have been used on 600K Affymetrix® Axiom® SNP array to detect CNV and to explore the contribution of CNV to phenotypic variation (Lyra et al., 2018).

With the emergence of massive parallel sequencing, new methods have been developed to detect structural variations based on the alignment of resequencing reads onto a high-quality reference genome sequence. Among these, three have been mainly used (Alkan et al., 2011): (i) the “read-depth” (RD) method which can only detect copy number variations, (ii) the “read-pair” (RP) method which can detect deletions as well as small insertions (up to the size of the insert), (iii) the “split-read” (SR) method which can also detect deletions and small insertions (up to the size of a read). Chia et al. (2012) used the RD approach to identify CNVs among 104 maize lines and performed association studies for several traits. However, the RD method does not allow the identification of novel sequences and is error prone, especially regarding the size of the discovered CNVs which greatly depends on the size of the sliding window used. The RP method has been implemented in many computational tools like BreakDancer (Chen et al., 2009) and has been widely used. Although it has proven to be highly efficient to detect deletions(Tuzun et al., 2005; Korbel et al., 2007; Kidd et al., 2008), this approach suffers from two limitations: it does not allow precise detection of breakpoints and the size of the insertions which can be detected is directly limited by the library insert size. The SR method, which was first implemented in Pindel (Ye et al., 2009), has the advantage of defining breakpoints at a single base resolution, but again the size of the detectable inserted sequence is limited.

The “assembly” (AS) method is able to detect all types of SVs of any size but is also the most cost- and computation-intensive. It is the only method able to detect large insertions with precise breakpoint definition. However, the assembly of large and complex genomes such as maize remains very expensive and computationally intensive despite recent progress in this area (Hirsch et al., 2016; Jiao et al., 2017; Darracq et al., 2018). There has been in the past some attempts to reduce this complexity by reducing the number of sequences to assemble. For instance, Lai et al., (2010) identified 104 deletions and 570 insertions among 6 maize inbred lines by assembling genomic regions from reads that did not map on the B73 reference genome. The sequences assembled by this approach were enriched in erroneous reads or reads coming from external contamination and they were too short to be anchored to the reference genome B73. Hirsch et al. (2014) identified several putatively expressed genes that were not present within B73 reference genome by assembling and comparing the transcriptome of hundreds of inbred lines. This new approach was limited to the transcribed part of the genome and suffered from a high level of false positives. More recently, Lu et al. (2015) used genotyping by sequencing approaches on 14,129 inbred lines to identify 1.1 million short and unique sequences (GBS tags) that (i) did not align on the B73 reference genome or were aligned but outside of a 10Mbp windows around their mapped position, or (ii)

were mapped at the same location by joint linkage mapping in NAM populations using co-segregation with SNP and logistic regression between InDel and SNP in an association panel. The main drawback of this approach is the high percentage of missing data due to the low depth of sequencing which requires imputation before being able to make genetic analysis. Recent whole genome sequence assembly of PH207 (Hirsch et al., 2016), and F2 (Darracq et al., 2018) have allowed the identification of thousands of large InDel and PAV sequences. For instance, 2,500 genes were found either present or absent in PH207 and B73 genomes and 10,735 PAV sequences larger than 1kb were discovered between F2 and B73, including 417 novel genes in F2. These discovery approaches have been limited to a few individuals due to sequencing costs and computational challenges, so they have not been adapted for characterization of SVs on large maize panels. Darracq et al. (2018) developed an interesting approach for the genotyping of PAVs from mapping of low depth (5-20X) resequencing datasets. This method is based on the comparison of reads aligning to the region found in F2 and in the line of interest. While this method is potentially adapted to genotype PAVs on any set of line with low resequencing data, it has been so far used for PAV genotyping on a low (<30) number of maize lines. Moreover, it is restricted to the analysis of PAVs, and is not adapted for genotyping other types of SVs.

To our knowledge, no high-throughput genotyping approach has been developed for genotyping large numbers of InDels, including PAVs, on a large set of individuals. In this study, we present an approach which is both (i) comprehensive, as it includes the discovery and localization of deletions as well as insertions regarding the B73 reference genome at the base pair level and (ii) high throughput, as it allows to genotype thousands of InDels on hundreds of individuals. Our strategy takes advantage of next generation sequencing (NGS) technologies and recent advances in assembly of complex genomes. It also benefits from the high efficiency of SNP arrays like the high-throughput Affymetrix AXIOM technology. In this paper, we detail how we discovered thousands of small to large InDels, including PAVs, from three maize inbred lines (F2, PH207 and C103) as compared to the B73 reference genome. We then describe how we designed and selected 600,000 probes to create a new Maize Affymetrix AXIOM array to genotype these InDels. Finally, we describe how we successfully used this array to genotype an association panel of 362 maize inbred lines.

Results

InDels and PAVs discovery

To design a comprehensive InDel genotyping array, we first needed to discover a set of InDels which would be representative of the maize temperate germplasm. We already had access to sequence data for the European flint line F2 and we benefited from a first set of 42,330 F2-specific sequences larger than 150pb, and totaling 16Mb. This dataset was constituted from the *de novo* assembly of F2 paired-end that failed (at least for one read of the pair) to align onto the B73 AGPv2 sequence and which were totally devoid of coverage by B73 reads ("Reference guided assembly" in Figure S2.2 so called "no map" approach). We also took advantage of the work done by Darracq et al., 2018 to add another 10,044 F2-insertions (size >1 kb, total size of 88Mb) with less than 70% of their length covered by B73 reads.

To complement these two datasets of F2/B73 deletions and insertions, we generated Illumina paired-end and mate-pair sequences from two other key founders of temperate maize breeding programs: PH207 and C103. We then used these F2, PH207 and C103 sequence data to detect, not only PAVs this time, but all InDels, at base pair resolution between these three lines and B73. This methodology allowed us to have access both to their sequences and their breakpoints allowing to genotype such InDels in several individuals (See material and methods for more details).

We first aligned F2, PH207 and C103 sequences against the B73 reference genome sequence in order to detect deletions. Here, the term "deletion" does not reflect any underlying biological process of DNA excision but refers to a sequence of at least 100bp present in the B73 genome at one locus and absent in another line at the same locus. Deletions were detected for the three lines simultaneously using the "genotyping" option of Pindel (Ye et al., 2009), generating a set of 26,368 non-redundant deletions with precise identification of their breakpoints (Figure S2.1 A). The number of deletions found for each line was similar, respectively 12,165, 11,922 and 13,432 for F2, PH207 and C103. 67% of the deletions found were unique to one line, 24% were shared by two lines and 9% by three lines. These results confirm the good complementarity of the lines chosen in this study.

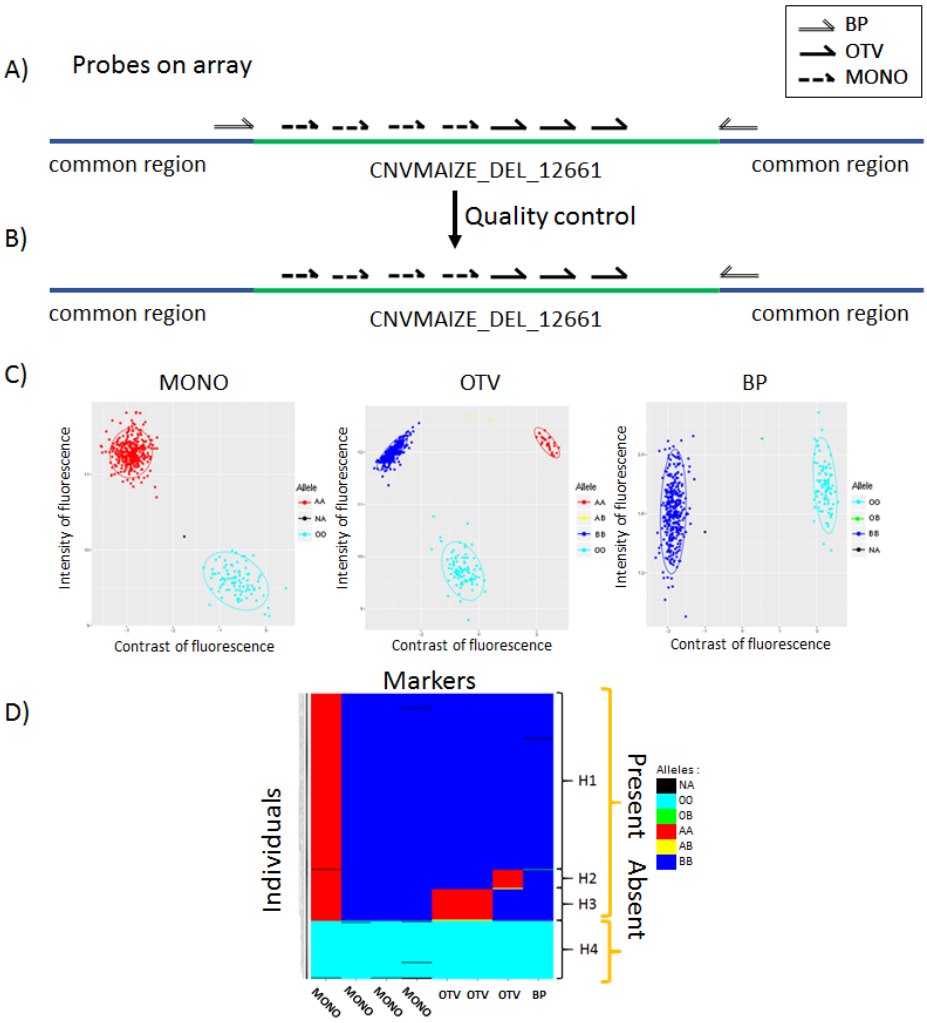


Figure 2.1: Genotyping of InDel CNVMAIZE_DEL_12661 using three probe types on 445 individuals. A) Schematic distribution of the 9 probes along the sequence of InDel CNVMAIZE_DEL_12661 (green line) and the bordering sequence common between all individuals (blue line) genotyped by the array. Double, dotted, and full arrows represented the probes designing on the forward and reverse flanking sequences of the breakpoint sites (BP), at not polymorphic (MONO) and polymorphic sites (OTV) within internal sequence of InDel. B) Schematic distribution of the 8 probes passing Affymetrix® quality control and called by Affymetrix® pipeline C) Clustering produced by Affymetrix® algorithm for an OTV, MONO and BP probe from InDel based on both fluorescence contrast (X axis) and intensity (Y axis) of the 445 inbred lines. Red, blue and yellow dots indicated the presence of the sequence (genotype "present") either homozygous for allele A (AA), or allele B (BB) or heterozygous (AB), respectively. Cyan and green indicated that the sequence were absent in the individual (OO), or only in one copy of the sequence, e.g hemizygous for presence/absence (OB or OA). Black dots indicated individuals for which no genotype could be assigned (Missing data) D) Haplotypes displayed by the genotyping using 8 probes (column) on the 445 inbred lines (row). Colors corresponded to the genotype of individuals produced by clustering in C)

Next, we generated a draft genome assembly for each of these lines, which were used as template for alignment of B73 reads to detect insertions regarding B73 reference genomes. As for deletions, here the term "insertion" does not reflect any underlying biological process of DNA integration but defines a sequence larger than 100bp that is present in one line at a given locus, and absent from B73 at the same locus. These three draft assemblies cover less than one third of the expected maize genome size but include a large portion of low copy sequences, including genes, as shown by BUSCO results (Table 2.1). Detection of insertions was processed this time separately for each inbred line, and generated 28,221 insertions for F2, 27,904 insertions for C103 and 26,795 insertions for PH207, with their precise breakpoints. The number of insertions is similar between lines, but significantly greater than this obtained for deletions. Among these insertions, 26,691 cases could be uniquely anchored at base pair resolution onto the B73 reference genome sequence (Figure S2.1b). Again, a majority of insertions were unique to one line (72%) confirming the complementarity of the material chosen.

Finally, the results from the different approaches were merged into a non-redundant set of 141,325 InDel sequences (see material and methods) comprising 52,175 deletions and 89,150 insertions. These regions were then used for the design of genotyping probes.

Table 2.1: F2, PH207 and C103 de novo assembly metrics. For each assembled genome are detailed: the number of scaffold sequences which were assembled, the length of the shortest scaffold, the length of the longest scaffold, the average size, the N50 of the assembly, the total number of bases included in the assembly, the percentage of Ns present in the assembly and finally the BUSCO statistics including the percentage of complete (C), fragmented (F) and missing (M) BUSCO genes from a total of 1440 BUSCO groups searched for maize.

Maize line	Number of scaffolds	Min size	Max size	Average size	N50	Total (Mb)	% of Ns	Complete BUSCOs (C)	Fragmented BUSCOs (F)	Missing BUSCOs (M)
F2	76,563	892	112,956	16,900	14,042	646.3	9.48%	89.30%	4.90%	5.80%
PH207	81,688	884	2,024,489	29,557	16,860	797.5	8.90%	91.80%	2.70%	5.50%
C103	84,990	886	120,582	19,305	16,146	793	8.21%	90.60%	4.20%	5.20%

Design of the genotyping array

Genotyping strategy

Large InDels can be efficiently genotyped with a SNP array using a combination of two types of probes: (i) "external" probes which target breakpoints using the two flanking sequences of a given InDel (BP probes) and (ii) "internal" probes which target presence/absence regions (PARs) within the internal sequence of InDels on polymorphic (OTV probes) or monomorphic sites (MONO probes). We define PARs as small portions of DNA sequence of at least 35bp that were observed present or absent at the genome level when comparing two individuals. They are thus suitable for the design of presence/absence genotyping probes. Ideally, each InDel should be called by two BP probes on either side and by multiple internal probes regularly distributed along the internal sequence of the InDel (Figure 2.1 A). However, in practice, this combination of different probes is not always possible. For instance, precise breakpoints were not described for all PAVs from our "No-map" approach (Darracq et al., 2018), and PARs for internal probes were not always found in our InDels.

Probes design

On one hand, BP probes, which should behave like classical SNP probes where one allele corresponds to the presence and the other to the absence of the InDel. They are useful to explore the conservation of the localization of large insertion/deletion events across multiple individuals, even when no internal probe can be designed due to the absence of PARs (Figure S2.6). Among the 141,325 selected variants, 86,406 InDels (22,420 deletions and 63,986 insertions as compared to the B73 reference genome sequence) had breakpoints defined at base pair resolution and were suitable for BP probe design. Four different breakpoint types were identified according to the presence of micro-homology and/or shorter non homologous sequence (Muñoz-Amatriaín et al., 2013) in place of a complete deleted sequence (Figure S2.3): (type I) 3,397 cases with sharp breakpoints; (type II) 45,987 cases with a micro-homology sequence (8.6 bp on average and no more than 237 bp) which was present in one copy in the reference sequence and duplicated at both extremities of the novel inserted sequence; (type III) 36,893 cases harboring insertion of a short non-homologous fragment (42.2 bp on average and up to 892 bp) in place of a large deleted sequence; and (type IV) 156 cases with a combination of type II and type III breakpoints. Following Affymetrix® recommendations, 19,010 InDels with type II breakpoints having a micro-homology sequence longer than 5bp were excluded from the design process. In the end, 67,396 InDels, representing 48% of all available InDel variants, were submitted to the Affymetrix® design pipeline. Two probes, one on forward (FW) and one on reverse (REV) strand, were designed for each breakpoint. These probes were classified as *not possible* (18%), *not recommended* (33%), *neutral* (15%) and *recommended* (35%) by this automated pipeline (see Methods for details), leaving 33,430 InDels (51%) that could be targeted by at least one *recommended* probe.

On the other hand, internal probes, which should behave like an “off-target” variant (Didion et al., 2012) where the hybridization of the probe stands for the presence and the absence of hybridization for the absence of the InDel, are useful to explore the genetic diversity within InDel sequences (Figure 2.1 D). They will also be particularly interesting to target InDels for which no breakpoint could be identified (such as PAVs from the “no map” approach).

For the design of OTV probes, we benefited from the availability of SNPs which had been previously identified from the alignment of resequencing data from a core collection of 25 temperate maize inbred lines against the B73-F2 maize pan-genome from Darracq et al. (2018). As a consequence, OTV probes have only been designed for deletions positioned on B73 reference genome and F2 insertions coming from Darracq et al. (2018). Among these, the context sequences of 436,162 SNPs, corresponding to 21,390 InDels, were extracted and submitted to Affymetrix® design pipeline. Again, two probes, one on forward (FW) and one on reverse (REV) strand, were designed for each SNP. Finally, a total of 872,324 OTV probes could be designed and scored as *not possible* (0.05%), *not recommended* (71%), *neutral* (14%) and *recommended* (16%), leaving 17,589 InDels (82%) which could be targeted by at least one *recommended* probe.

For the design of BP and OTV probes we could rely on Affymetrix® design pipeline to identify probes localized in PARs and thus suitable for the Affymetrix® AXIOM® technology. For the design of MONO probes, we first had to identify PARs within 141,325 InDels cumulating 133Mbp of sequence. We used sequence masking methods to exclude repeats based on similarity to known maize repeats or on occurrence of 17-mers found within the sequencing datasets we had for B73, F2, PH207 and C103 (more details in methods). By doing so, we identified 122,972 PARs, representing a cumulated size of 27Mbp, corresponding to 20.3% of the initial size and allowing the possibility to design MONO probes for 79,987 InDels (56.5%). These PAR sequences were successfully used for the design of 25,735,797 MONO probes, among which 59% were scored as *recommended* and allowed to target 62,875 InDels (79%).

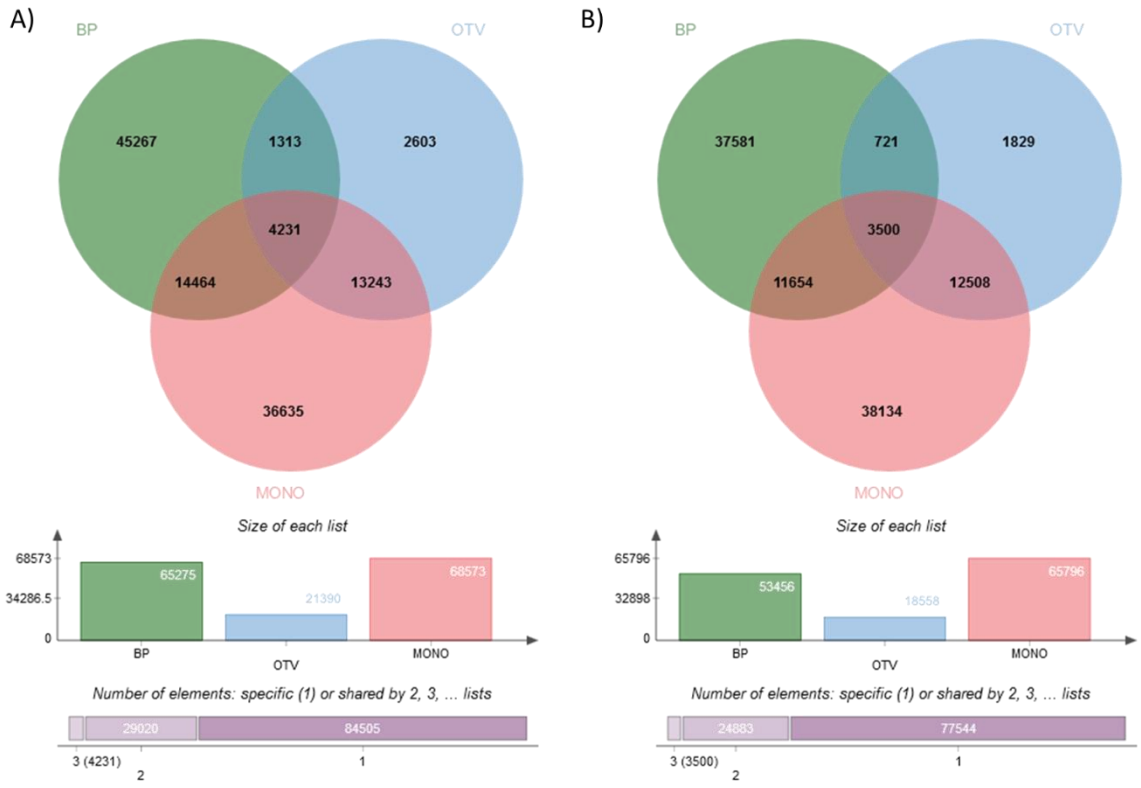


Figure 2.2: Number of InDels that could be (A) targeted by each type of probes designed and (B) selected to be included in the final array.

With this combined approach, we designed a total of 26,715,361 probes targeting 117,756 InDels, which represent a cumulated length of 250 Mbp including 27 Mbp of PARs (Table 2.2). Among these InDels, 97,748 (83%) can only be targeted with either internal or external probes, but not both (Figure 2.2 A). These results support our overall strategy which includes the discovery of InDels with precise breakpoints in a preliminary step, and the use of complementary internal/external probes for the genotyping of large InDels.

Table 2.2: Number of probes before and after selection for array design and passing the Affymetrix® quality control. At each step, are detailed the number (and percentage) of each probe type and the corresponding number (and percentage) of targeted InDels. Note that a same InDel could be genotyped by several probe types which conducted to a sum of percentage superior to 1 in InDel columns.

	Before selection		On array		Called by Affymetrix® pipeline	
	Probes	**InDels**	**Probes**	**InDels**	**Probes**	**InDels**
BP_Type1	6,648	3,324	4,691	2,751	2,092	1,482
BP_Type2	51,770	25,885	38,790	22,662	20,540	14,407
BP_Type3	71,820	35,910	41,272	27,897	23,631	18,485
BP_Type4	312	156	241	146	119	93
OTV	872,324	21,390	163,278	18,558	96,867	15,064
MONO	25,735,797	68,573	414,500	65,796	335,778	63,597
ALL	26,738,671	117,756	662,772	105,927	479,027	89,393

Array design

We used the Affymetrix® recommendations to select the 700,000 probes to be included in the final array, plus some other criteria depending on the probe type. Nevertheless, because of their added value, we decided to keep all BP probes as soon as they had less than 3 hits on the B73 reference genome sequence. This first selection consumed 84,994 probes targeting 53,456 InDels, among which 70% could only be targeted by BP probes. Concerning OTV and MONO probes, we first selected *neutral* and *recommended* probes having no hit at all (for insertions), and only one hit (for deletions), against the B73 reference genome sequence. We then considered their density with the objective to maximize the number of InDels that could be surveyed, as well as to have an even distribution of probes along targeted InDel sequences (see Methods for more details). We then performed a second selection among *not recommended* OTV and MONO probes for 4,541 InDels that were still not targeted. After filtering some duplicated probes, we built a final array design containing 662,772 probes targeting 105,927 InDels that represent a cumulated length of 232 Mbp, including 25.9 Mbp of PARs.

Description of the array content

The final array design allows to genotype InDels with various sizes, ranging from 37 bp to 129.7 kbp, with a median of 501 bp (Figure S2.4). They are covered by 1 to 482 probes with a median of 3 probes per InDel (Figure S2.5). But the number of probes does not always reflect the length of the InDels, as the proportion of PARs within InDels is highly variable. Indeed, while 8,040 InDels (ranging from 37 bp to 2,409 bp with a median of 163 bp) were completely covered by PARs and could thus be considered as a proper PAVs, 34,372 InDels (ranging from 101 to 129,700 bp with a median of 320 bp) were not covered by any PAR at all (Figure 2.3). In fact, the number of internal probes were more strongly correlated to the size of the PARs (r2 = 0.79) rather than to the size of the InDels (r2 = 0.16) (Figure S2.6).

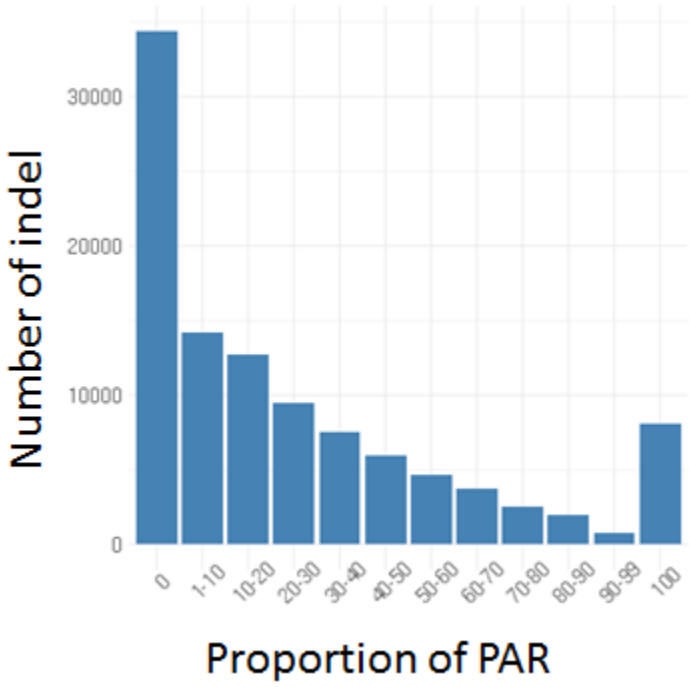


Figure 2.3: Distribution of the number of InDels genotyped by the array according to the proportion of presence/absence regions (Specific fraction) identified in their internal sequence.

As expected, the probe selection process did not impact the overall distribution of probe types among targeted InDels as 35% of them can exclusively be genotyped by BP probes, whereas 50% can only be genotyped thanks to the use of internal probes, among which 73% are only targeted by the use of the original MONO probes (Figure 2.2b). Indeed, a large number of InDels did not contain PARs and cannot be genotyped with 35bp internal probes but only with BP probes whereas some others InDels contains PARs but not have BP due to InDel discovery approach ("No map").

Among the 43,117 InDels that could be anchored onto the B73 reference genome sequence and which were included in the array design, 13,737 were located inside a gene, 57 close to a gene (less than 1 kb away), 1,311 inside a pseudogene and 2,212 inside a transposable element. From the localization of these InDels, evaluated InDels and probe density across each chromosome. We observed a higher density in chromosome arms than in peri-centromeric regions (Figure S2.7). We also identified clusters of InDels with large specific sequence at the beginning of chromosome 6 (10-20Mbp) or at the end of chromosome 5 (~190Mbp).

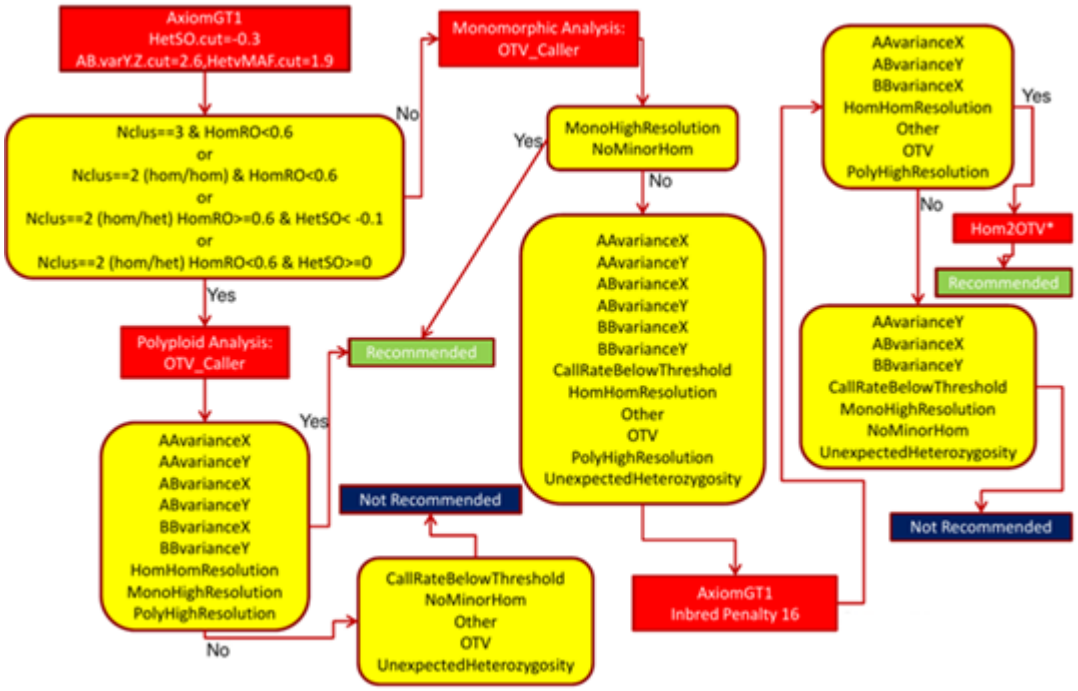


Figure 2.4: Dedicated Affymetrix® pipeline used for calling InDel polymorphisms from the fluorescent intensity variation of MONO probes. Each probe was classified into different categories according to the number of clusters, the call rate and quality metrics of the clustering based on the position, variance and separation of different cluster. In order to retrieve the best clustering for each probe, successive step of clustering using different clustering algorithms (Red square, Axiom GT1, OTV caller, Hom2OTV) or/and with different parameters. According to their classification at each step (yellow square) and threshold used for quality metrics, probes could be classified as recommended (green square), not recommended (blue square) or to be submitted to another step. A new pipeline and an algorithm (Hom2OTV) must be specifically developed for calling InDel genotype of MONO probes since we expected only 2 clusters (absence / presence) that varied exclusively for fluorescent intensity rather than for fluorescent intensity ratio between two labelled nucleotides. At the end, all probes were classified into 14 categories either as recommended or not recommended depending on threshold.

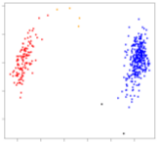
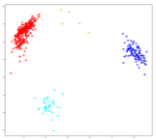
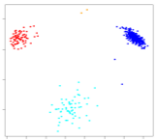
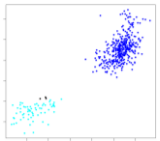
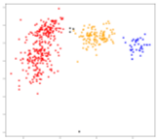
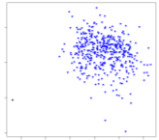
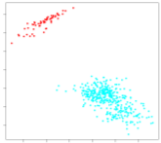
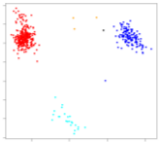
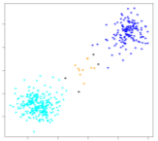
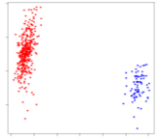
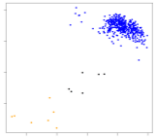
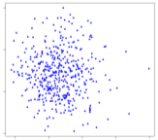
Assessing array quality by genotyping 105,927 InDels on 480 maize DNA samples

InDels calling using dedicated Affymetrix pipelines

We genotyped 480 maize DNA samples including 440 inbred lines, 24 highly recombinant inbred lines and 16 F1 hybrids. Dedicated Affymetrix pipelines were implemented for each of the probe types to call genotype of the InDels based on fluorescent intensity and contrast variation of the probes. It included two algorithms already developed by Affymetrix (Didion et al., 2012) for BP and OTV probes (Figure S2.8 A et B) and a third one which was newly developed for the calling of presence/absence alleles using MONO probes (Figure 2.4). 35 DNA samples including all F1 hybrids, did not pass Affymetrix quality control due to their low call rate (<0.9) and were eliminated. Out of 662,772 probes, 479,027probes representing 89,393 InDels (84%) passed Affymetrix quality control and were called on 445 DNA samples. Respectively 55%, 59% and 81% of BP, OTV and MONO probes were converted into recommended markers after clustering by Affymetrix pipelines (Table S2.1, S2.2, and S2.3). Thanks to the 3 probe types and redundancy, 84% of InDels could be called with an average of 5.4 probes per InDel.

To evaluate the genotyping capacity of the probes, we first compared the clustering of inbred lines expected for three probe types (BP, OTV, and MONO) with the observed clustering of inbred lines based on fluorescence intensity and contrast of 445 inbred lines genotyped with the array. For BP probes, we expected at least two clusters corresponding to the individuals homozygous either for presence (“AA” or “BB”) or absence (“OO”). A third cluster could be observed when individuals were heterozygous individuals for presence/absence (“OA” or “OB” hemizygous) (Figure 2.1 C). For OTV probes, we expected at least 3 different clusters: two clusters corresponding to the individuals homozygous for allele A or B of SNP (“AA”, “BB”), and a third “off-target” cluster for the individuals homozygous for absence (“OO”). A fourth cluster could be observed when some individuals were heterozygous at the within-InDel SNPs (AB). For MONO probes, we expected only two clusters corresponding to the individuals for which the sequence was present (“AA” or “BB”) or absent (“OO “, ”AA” or “BB”) (Figure 2.1 C). The observed clustering by the three dedicated pipelines was consistent with the expected clustering for 43% of BP, 83% of OTV and 63% of MONO probes (Table 2.3).

Table 2.3: Comparison between the clustering expected for BP, MONO and OTV probe type and the clustering produced by Affymetrix® pipelines based on the fluorescent intensity and contrast of 445 inbred lines for 479,027 probes. Clustering example: typical example of clustering based on the fluorescent intensity (y-axis) and contrast (x-axis). Colors indicate the assignation of the individuals to different clusters identified by pipeline. Description: brief characteristic of each classification based on the clustering of individuals (homoz.= homozygote, het=heterozygous, OT= off-target)

		Classification based on the clustering produced by Affymetrix® pipelines and genotyping assignment					
Probes		**BP**	**OTV**				
BP	**Nbr (%)**	20,370 (43.9%)	26,012 (56.1%)				
	Clustering example						
	Description	Two homoz. clusters	Two homoz. and one OT clusters				
		OTV	**MONO**	**SNP**	**monomorphic**		
OTV	**Nbr (%)**	78,799 (81.3%)	502 (0.5%)	17,562 (18.1%)	4 (0.0%)		
	Clustering example						
	Description	Two homoz. and one OT clusters.	One homoz. and one OT clusters.	Two homoz. clusters.	One cluster		
		MONO	**OTV**	**Unexpected MONO 1**	**SNP**	**Unexpected MONO 2**	**monomorphic**
MONO	**Nbr (%)**	212,434 (63,3%)	15,690 (4,7%)	68,562 (20,4%)	1,981 (0.6%)	9,525 (2.8%)	27,586 (8.29%)
	Clustering example						
		One homoz. and one OT clusters	Two homoz. and one OT clusters.	One homoz., one OT and one het. clusters.	Two homoz. clusters.	One homoz. and one het. clusters.	One cluster

We observed also some unexpected clustering. For 57% of BP probes, we observed an additional off-target cluster (OTV in Table 2.3). This indicates that some BP probes did not hybridize properly in some inbred lines, which can either be due to the presence of polymorphism within flanking sequences of the targeted InDels or to the existence of more complex rearrangements removing the breakpoints. To explore these two hypotheses, we took advantage of the availability of forward (FW) and reverse (REV) probes for 12,150 InDels to determine whether the clustering between FW and REV BP probes from the same InDel was similar or different. While 12% of these InDels had their FW and REV BP probes classified identically either as OTV, 35% had their FW and REV probes classified differently (one as BP and the other as OTV).

Regarding MONO probes, 25% displayed additional cluster(s) when sequence was present suggesting the presence of a single nucleotide polymorphisms at this position. Among these, we were able to distinguish two types of clustering (Table 2.3). 4.7% of MONO probes exhibited a clustering similar to those observed for OTV probes suggesting that these MONO probes revealed really by chance a single nucleotide polymorphism. In contrast, 20.4% of MONO probes displayed an unexpected clustering pattern for inbred lines with the presence of a heterozygous cluster but absence of a second homozygous cluster for SNP (Figure S2.12 B). In the end, 2.8% of MONO probes displayed an additional heterozygous cluster for SNP when sequence is present but no “off target” cluster corresponding to individuals for which sequence are absent (Figure S2.12 D)

For 18% of OTV (Figure S2.12 A) and 8.3% of MONO probes, clustering displayed no “off target” cluster for absence suggesting no presence/absence polymorphism at this position (Table 2.3). Note that some BP were also classified as monomorphic for presence/absence but were filtered out by the BP pipeline (MonoHighResolution in Table S2.1).

Finally, 422,369 probes were able to call both presence and absence alleles, which allowed us to successfully genotype a total of 86,648 InDels (82% of InDels targeted by the array) on 445 inbred lines.

Evaluation of genotyping reproducibility and quality

Consistency of genotyping among the four inbred lines used for InDels discovery

We used the 479,027 probes passing Affymetrix quality controls to evaluate the quality of Presence/Absence genotyping by comparing the genotyping results from our array with those generated from sequencing data from the 4 lines used for the discovery of InDels (B73, F2, PH207, and C103). Respectively 97.5%, 92.7% and 90.3% of the BP, OTV and MONO probes predicted a genotyping result consistent with this obtained with BLAST. We observed a strong asymmetry for concordance rate depending on whether we expect the locus to be present or absent from sequencing data (94.9% vs 86.2% for allele present and absent, respectively). Interestingly, we observed no asymmetry for BP probes and a strong asymmetry for OTV and MONO probes for concordance rate (Table 2.4). The four inbred lines showed very similar concordance rates, F2 being the most concordant (97.9%). The median consistency rate of probes within InDels remained relatively high and stable, around 90%, independently of the number of probes per InDel (Figure S2.9).

Table 2.4: Consistency rate between expected and observed genotype for 4 individuals used to discover InDel, according to the three type of probes and the two different genotypes expected: presence (P) or absence (A) of the sequence.

Probe types	Expected genotyping	B73	F2	C103	PH207	All individuals
BP	**A**	0.96	0.93	0.94	0.94	0.94
	P	0.96	0.95	0.95	0.95	0.95
OTV	**A**	0.85	0.89	0.80	0.78	0.83
	P	0.93	0.97	0.96	0.96	0.96
MONO	**A**	0.77	0.81	0.82	0.81	0.80
	P	0.90	0.98	0.94	0.94	0.95
All probe types	**A**	0.80	0.85	0.83	0.82	0.82
	P	0.92	0.97	0.94	0.94	0.95
	A & P	**0.85**	**0.94**	**0.89**	**0.89**	**0.89**

Consistency among probes from the same InDel

To estimate the consistency of different probes for typing a given InDel, we analyzed genotyping results for 48,486 InDels genotyped with at least two probes in a collection of 24 temperate inbred lines. Among these 24 lines, there are the four lines used to discover PAVs and the twenty used to discover SNPs within specific regions of InDels (Darracq et al., 2018). For each InDel and each inbred line, we calculated the frequency of presence call over all probes. Frequencies of 1 (presence) and 0 (absence) indicated that all probes displayed consistent genotyping for the corresponding inbred line. Overall, 78% of these InDels genotyping displayed an average allelic frequency for the presence allele of 1 or 0 meaning that all probes had consistent genotyping results for calling the allele at both present and absent states, respectively (Figure 2.5). A total of 12,308 InDels (25%) displayed only two states across the 24 inbred lines, corresponding to the presence or the absence of the sequence, while for 75% at least one inbred line had at least one inconsistent probe conducting to the presence of more than two haplotypes across 24 inbred

lines. Some contradictory calls were repeatedly found across the 24 samples (Figure S2.10), thus suggesting that some between-probe inconsistencies could have biological origins rather than being calling errors.

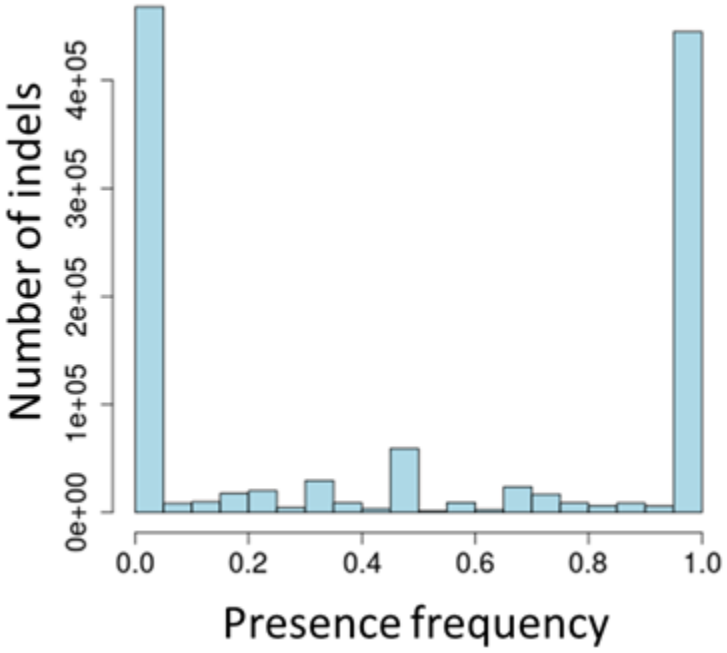


Figure 2.5: Distribution of the average allelic frequencies of the presence across different probes within 48,486 InDels with at least two probes genotyped for 24 inbred lines.

To investigate the consistency between the forward (FW) and reverse (REV) BP probes, we compared the genotyping results of 8,116 InDels having both FW and REV BP probes called on our 24 inbred lines. 33% of these InDels have a consistent calling between their FW and REVs probes for all inbred lines. The proportion of InDels displaying an inconsistent calling between the FW and REV probes for 24 lines varied according to the breakpoint type and their classification (Figure S2.11). We observed also more similar calling when both FW and REV probes had similar classification (BP-BP or OTV-OTV) than when they had different classification status (BP-OTV) (Figure S2.11 A). Altogether, these results suggest that some calling inconsistencies could come from polymorphisms in the flanking sequence while some other could be due to local rearrangements in the lines under genotyping as compared to the lines used for InDels discovery.

Assessing array quality to call highly hemizygous individuals using BP

In order to evaluate our ability to identify individuals displaying hemizygous genotype (heterozygous for presence / absence of the sequence), we rescued for BP probes the genotyping of DNA samples for 12 F1 hybrid eliminated by Affymetrix quality control due to their low call rate. This low call rate came mainly from inability of current Affymetrix algorithms to identify hemizygous cluster for OTV and MONO probes and therefore to assign a genotype to hemizygous individuals. As a consequence, it

strongly increases missing data for F1 hybrids only for OTV and MONO probes. We selected 20,370 BP probes classified as expected by the design (Table 2.3) to compare them with those expected from their 9 parental lines. 89% of observed homozygous and 94% of observed hemizygous alleles were consistent with expected genotyping results of F1.

Reproducibility

We evaluated the reproducibility of genotyping by comparing the genotyping results of 13 different inbred lines that were replicated in the experiment (Table S2.4). Note that these are not perfect biological replicates as they represent the same variety but come either from different seed lots or from different accessions. These replicates exhibited a genotyping difference varying from 0.6% to 5.2%. This is similar to the amount of inconsistencies obtained on the same material using a 50K SNP array (Ganal et al., 2011) suggesting that InDel genotyping inconsistencies for replicates come mostly from seed lot divergences rather than genotyping errors (Table S2.4).

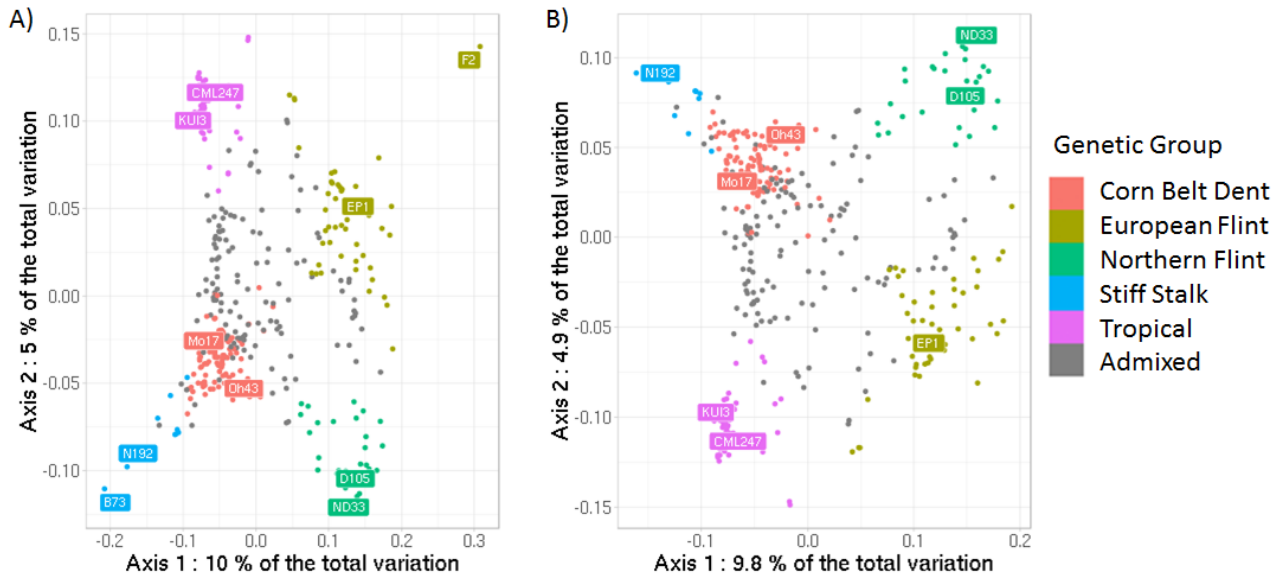


Figure 2.6: Principal coordinate analysis on the genetic distance (1-IBS) between inbred lines from an association panel estimated by 57,824 InDels. A) 362 maize inbred lines were represented B) 360 maize inbred lines were represented excluding B73 and F2 that are used for discovering InDels. Colors represented the assignation of the inbred lines to the 5 genetic groups defined by admixture using Panzea SNPs from 50K Illumina array when the probability of assignation to a group (membership) were superior to 60%. Inbred lines that are not assigned to a group (membership<60%) were considered admixed. Common name of two maize accessions typical of each genetic group were indicated.

Application: Diversity analysis of 362 maize inbred lines panel

In order to evaluate the interest of this new array to analyze the contribution of InDels to the genetic diversity, we analyzed 57,824 polymorphic InDels among a subset of 362 out 445 inbred lines, representing a large genetic diversity and previously studied (Camus-Kulandaivelu et al., 2006; Bouchet et al., 2013). To give same weight to each InDel in the diversity analysis, we selected one single probe per InDel based on the probe genotyping quality (see methods).

We first used these InDels to calculate the genetic distance between inbred lines and to perform Principal Coordinate Analysis (PCoA) (Figure 2.6 A). To compare our InDel-based results to this of previously characterized SNPs, we displayed on this PcoA the genetic structuration of these 362 inbred lines as obtained from the Panzea 50K SNP array (Bouchet et al., 2013). The first axis showed good discrimination of European Flint from Corn Belt Dent and Stiff Stalk lines, while the second axis discriminated European Flint and Northern Flint lines. Overall, the clustering of individuals based on genetic distance estimated with InDels (1-IBS) by PCoA was consistent with the genetic structuration obtained from SNPs. We observed that B73 and F2, that were used to discover the majority of InDels, deviated from other inbred lines. We thus performed a second PCoA excluding B73 and F2 (Figure 2.6 B). The two PCoAs gave similar patterns.

Discussion

An original high throughput approach for genotyping InDels

The comparison of whole genome sequence assemblies is in theory the best approach to identify precisely and exhaustively structural variations between two individuals (Hirsch et al., 2016; Jiao et al., 2017; Darracq et al., 2018). But even though great progress has been made recently in this area, whole genome assembly is still too costly, time consuming and computationally intensive to be applied to hundreds of individuals considering the complexity of maize genome (Darracq et al., 2018; Gabur et al., 2018). Other whole genome approaches based on sequencing and alignment of reads, and using "read-depth", "read-pair" and "split-read" identification methods (Tuzun et al., 2005; Korbel et al., 2007; Kidd et al., 2008; Chen et al., 2009; Ye et al., 2009) were mostly limited to the identification of deletions (i.e. sequences absent compared to a reference genome). Liu et al., (2015) partially addressed the lack of insertions (i.e. novel sequences compared to a reference genome) in previous studies by the identification 1,973,746 InDels. Although, among these a majority were very small (85% smaller than 11bp) and the use of PCR markers to genotype them was time-demanding, labor-intensive and costly at large scale level. In this paper we describe an original approach combining the accuracy of the detection of insertions and deletions using high coverage sequence data and multiple reference genome assemblies, along with the high-throughput and accuracy of SNP arrays. We further show that using this approach, we were able to design and use an innovative array which allowed for the first time to genotype accurately thousands of small to large insertion/deletion variants, including PAVs, on hundreds of maize individuals. We used different methods to compile 52,175 deletions and 89,150 insertions between three newly sequenced maize inbred lines (F2, PH207 and C103) and the maize B73 AGPv2 reference genome, among which 75% were included in our array. Contrary to older studies, we did not focus solely on PAVs, but we also included in our array many insertion and deletion events, even if they contained non-unique sequences, by targeting their breakpoints.

By designing probes directly on InDel breakpoints for both insertions and deletions, our approach overcomes some of the limitations of CGH or SNP array-based studies. To our knowledge none of the previous studies which have used an array technology for genotyping InDels have specifically targeted such a high number of insertion/deletion breakpoints. Unterseer et al., (2014) genotyped 6,759 small deletions which were discovered by aligning reads of 30 inbred lines against B73 genome, but it included no insertions. However, CGH and SNP arrays did not usually design probes to target breakpoints and detected InDels by analyzing the variation of fluorescent intensity signals of ordered probes (Cooper et al., 2008; Dellinger et al., 2010 Wang et al., 2017). Consequently, these technologies targeted exclusively low copy regions of the genome excluding InDels containing repeats such as TEs as soon as their breakpoints were not included in design (Springer et al., 2009; Beló et al., 2010; Lyra et al., 2018). This is a strong drawback for maize and many other crops since a large part of their sequence is composed of transposable elements (Feschotte et al., 2002; Schnable et al., 2009) that may be highly variable between individuals (Morgante et al., 2007; Liu et al., 2015; Sun et al., 2018) and may impact phenotypes (Salvi et al., 2002, 2007; Ducrocq et al., 2008). Using BP probes allow to target Present/Absent Variation whose

sequence were unique and not present elsewhere in the genome as well as transposable elements whose internal sequence could be present/absent at one locus but present elsewhere in the genome. Another advantage to genotype breakpoints is that we are almost certain to genotype the same mutational event across all individuals of the population because it is highly unlikely that two independent mutational events can lead to the same breakpoint. On the contrary, when we detected classically InDels using CGH or SNP array, it is much harder to identify common InDels among a population of individuals as we don't know precisely the breakpoint at base pair level. Genotyping breakpoint is also very cheap since only one or two probes by InDel are required. InDel size is therefore no longer a limitation for genotyping using breakpoints on the contrary to SNP and CGH arrays which have limited resolution when they used fluorescent intensity variation (Alkan et al., 2011). The genotyping of breakpoints by sequencing is possible with a tool like Pindel (Ye et al., 2009) which has a genotyping mode, but at a much greater cost and with lower call rate compared to the use of an SNP array. Finally, breakpoint probes are codominant markers and allow to accurately genotype hemizygous individuals (Heterozygous for presence/absence) since their genotyping are based on fluorescent contrast rather than fluorescent intensity variation which are known to be noisier as for MONO and OTV probes (Alkan et al., 2011).

Although the use of BP probes is clearly the simplest way to genotype InDels using an SNP array, breakpoints are not always available (no map approach discovery) or "designable" with 35bp probes, for instance when sequence of microhomology at breakpoint site were larger than 5bp. In order to genotype the 52,471 InDels without breakpoints and explore the genetic diversity within InDels, we also designed 577,778 internal probes both on monomorphic and polymorphic sites on PARs for both insertions and deletions. To genotype PARs in InDel internal sequences using SNPs, we took advantage of the already available Affymetrix algorithms to call Off-Target Variants (OTVs) which can detect variation of fluorescent intensity signals for a single probe (Didion et al., 2012) (Figure 2.1 C). This approach was used by Unterseer et al. (2014) who was able to detect 45,974 OTVs on a set of maize inbred lines using a 600K SNP array. Nevertheless, the array was designed in a classical way to target SNPs and there was no prior evidence that the probes called as OTVs would belong to real InDels like in our approach. Additionally, detecting SNP in insertion required to assemble a pangenome combining common and specific sequence from different individuals in order to retrieve SNP by aligning reads from sequenced lines. In our case, the sole use of OTV probes would have conducted to the elimination of a lot of InDels since 87,372 InDels including 74,648 insertions had no known SNPs within their internal sequence. In order to avoid this ascertainment bias due to prior knowledge of the presence of SNPs, we designed 414,500 MONO probes on putative monomorphic sites within PARs of InDel sequences. It permitted to genotype 38,134 supplementary InDels that could be targeted neither by OTV or BP probes. This new type of probes required the development of a new algorithm in order to cluster individuals according to their fluorescent intensity variation only, to be able to assign a genotype to each individual (Figure 2.4). A limitation of current Affymetrix algorithms to genotype InDels using OTV and MONO probes is that they are currently unable to genotype hemizygous individuals. While it was not a strong issue for maize inbred lines (or individuals from autogamous species) that are mostly homozygous, it is a strong issue for individuals from allogamous species that are highly heterozygous. By improving the current Affymetrix® algorithms, it should be possible to identify hemizygous cluster according to fluorescence intensity for OTV and MONO probes. We observed indeed some clusters that seem badly interpreted as heterozygote for SNP although they

correspond more probably to hemizygous individuals for OTV and MONO probes (Figure S2.12B, see below for more detailed discussion). Alternatively, other algorithms/software based on fluorescent intensity variation of either a single probe or several ordered probes exists and could be used to detect copy number variations and therefore hemizygote in individuals (Hupe et al., 2004; Olshen et al., 2004; Picard et al., 2005; Marioni et al., 2006; Picard et al., 2007; Stjernqvist et al., 2007).

Reliability of genotyping / calling results

Our approach provides a reliable and reproducible genotyping strategy for InDels since (i) 91.5% of alleles called from probes are consistent with expected genotype from the resequencing data available for the 3 lines (F2, PH207, C103), (ii) 78% of InDels genotyping had internal calls totally consistent between each other exhibiting either absence or presence for an inbred line, and (iii) the genotyping results were highly reproducible (94.8-99.4%) between biological replicates.

We observed a higher inconsistency between observed and expected calls for genotype “absent” than for genotype “present” with MONO and OTV probes, but not with BP probes (Table 2.4). This asymmetry between present and absent for consistency suggests a greater number of false positives in absent than present. We found that 20,574 InDels were in fact totally monomorphic and present across all lines suggesting they represented false-positive InDels coming certainly from regions which were not assembled in our draft genomes. Indeed, the probes targeting sequence regions present in one line but not assembled in their draft genome assembly, were falsely expected absent but they correctly hybridized with DNA and were called “present” on the array. This explains why the number of false positives was higher for B73, as all B73 absence genotypes were defined in comparison to draft assemblies, whereas for the other 3 lines absence genotypes were defined in comparison with the gold standard B73 genome sequence. The fact that we obtained a better result on OTV probes coming from F2 can be explained because we used only SNPs discovered on the B73-F2 pan-genome and not on other genomes. On the contrary, the fact that BP probes had similar consistencies for genotype “absent” and “present” could be explained because the BP probes were designed exclusively on B73 reference genome whatever we genotype insertions or deletions. One possible improvement of our approach to reduce the number of false-positive absences would be to not only align B73 reads onto each draft genome assembly but to align reads from each sequenced genome on each other and against itself. This would have several benefits: (i) it would allow to discover even more InDels and of better quality since each putative deletion discovered in one sample could potentially benefit from supporting reads from another sample, (ii) this would also simplify the identification of InDels common to more than on genotype, and last but not least (iii) it would help to identify and eliminate false-positive deletions by the alignment of each sample on its own draft assembly.

Nevertheless, the use of incomplete draft genomes does not explain all discrepancies between expected and observed genotypes. First, these genotyping errors could also be due to a wrong clustering leading to assign incorrectly genotype “present” instead of “absent” for a subset of individuals. It was well exemplified by some MONO probes classified as SNP although the clustering pattern looks like a MONO clustering with a strong difference of fluorescence intensity between two clusters. It suggests strongly for the cluster displaying the lowest fluorescent intensity a wrong assignment of homozygous genotype for

one of two SNP alleles (presence of sequence) instead of the assignment of the homozygous absence of the sequence (Figure S2.12 C). Similarly, the more detailed inspection of the clustering of MONO probes displaying unexpected cluster pattern (Table 2.4, figure S2.12 D) and OTV probes classified as SNP (Table 2.4, figure S2.12 A) suggests for some probes a wrong assignment of genotype for the cluster displaying the lowest fluorescent intensity since the clustering looks like MONO and OTV clustering. Second, the genome divergence within probe sequence for some inbred lines could conduct to group those individuals in an OTV cluster and therefore lead this time to the assignment of an absent allele even though the sequence is present for these lines.

Surprisingly, 4.7% of MONO probes displayed a classical OTV clustering suggesting that an unknown SNP was targeted by these probes by chance. These 15,690 new OTVs are very interesting since they were discovered by chance on a large set of 445 inbred lines. We could therefore expect that these OTVs have no ascertainment bias which can be very useful for analyzing genetic diversity within InDels carrying PARs regions. On the contrary, 20.4% of MONO probes displayed an unexpected clustering with one off-target cluster corresponding to absence of the sequence, one cluster corresponding to heterozygous inbred lines for SNP but only one homozygous cluster (Unexpected MONO 1 in Table 2.4). Considering these "unexpected MONO 1" as true SNP would conduct to a density of SNP (1 SNP every 5 bp) which are not compatible with level of diversity observed in maize in different previous studies (Gore et al., 2009; Brandenburg et al., 2017). Deeper investigation of these MONO probes clustering showed for some probes that the cluster of heterozygous inbred lines displayed intermediate position for both intensity and contrast between two homozygous clusters for presence and absence of the sequence, respectively (Figure S2.12 B). It suggested strongly that these clusters of inbred lines assigned as heterozygous were in fact inbred lines carried only one copy of the sequence (hemizygous genotype). An alternative hypothesis to explain this unexpected pattern is the presence of divergent duplicated sequence leading to the presence of an artefactual heterozygous cluster for SNP corresponding to the presence of two paralogous sequences rather than one copy. This result suggests therefore that there is probably room to improve Affymetrix algorithms in order to better identify additional clusters corresponding to the presence of hemizygous individuals for both MONO and OTV probes and therefore improve the quality of the genotyping of InDels when using a SNP array.

These potential clustering errors as well as the bad design of some probes previously mentioned can explain that only 27% of InDels displayed consistent genotype for presence/absence between all probes from same InDels across the 24 inbred lines. Interestingly, some InDels showed reproducible inconsistent genotypes for presence/absence across their probes in several inbred lines (Figure S2.10). It suggested that this pattern could not be due to random errors but could have instead a biological origin with possibly rearrangements having occurred several times within the same genomic region in some inbred lines. Following this hypothesis, Gu et al. (2008) observed two different types of rearrangement which could explain our observations: (i) rearrangements with an unique breakpoint in population and therefore common size between individuals conducting to two haplotypes in a population (ii) rearrangement with non-unique breakpoints scattered in a genomic region which conducted to several haplotypes. This hypothesis is also supported in our experiment by the 56% of BP probes classified as OTVs indicating that FW or/and REV flanking sequence did not well hybridize in all lines.

The development of a statistical approach to merge either *a posteriori* the calling results of independent clustering of individual probes or *a priori* the fluorescent intensity signal of successive probes within an InDel could be interesting in order to improve the robustness of InDel genotyping. This would have the advantage to limit the effect of genotyping errors due to a bad clustering and to reduce the noise in fluorescent intensity signals. It would also help to identify true different haplotypes representative of the complexity of a region in a population.

Finally, 72% of probes were converted into markers, a number which is comparable to other maize Affymetrix Axiom SNP arrays in comparison to 74.9% in Unterseer et al., (2014). Out of these, only 88% were really polymorphic for presence/absence. This conversion rate is not so bad considering that Affymetrix Axiom array analysis pipelines, which have been optimized for the detection of bi-allelic SNPs, are more sensitive to variations in fluorescent contrast (x-axis) compared to variations in fluorescent intensity (y-axis) which are known to be more noisy (Alkan et al., 2011; Didion et al., 2012). Moreover, we did not always follow Affymetrix recommendations as we did not filtered out probes with a bad design score.

To conclude, we developed a high-throughput and cost-effective InDels genotyping array based on the InDels discovered by sequencing on four inbred lines. It could be highly valuable to use more lines for the initial InDel discovery step since our four inbred lines do not well represent the whole genomic diversity of maize, notably tropical lines. As a consequence, it could lead to ascertainment bias by reinforcing the differentiation of inbred lines genetically close to the four inbred lines used to discover InDels (Clark et al., 2005; Ganal et al., 2011; Gouesnard et al., 2017) as we observed in our diversity analysis for lines close to B73 and F2. Several new maize genome assemblies are now available in the public domain and more and more could become available in the future. Our approach could easily be applied to these new genome assemblies to discover new InDels on a larger set of inbred lines representative of maize diversity with the aim to design a new InDels array. Although our arrays were not yet designed to genotype duplications and inversions, our approach could be easily extended to genotype breakpoints of inversions but required further development of pipeline for genotyping duplications using internal probes.

Material and Methods

InDel and PAV discovery

Three maize inbred lines, which are key founders of maize breeding program and originated from three different heterotic groups, have been selected for in depth sequencing and InDel discovery: the European Flint line F2 and two American dent lines, PH207 (Iodent) and C103 (Lancaster). DNA for genotyping were extracted from leaves following a NaBisulfite method modified from Tai and Tanksley (1990) and Dellaporta et al. (1983). For each inbred line, paired-end and mate-pair whole genome shotgun libraries were sequenced on Illumina HiSeq 2000 platforms (Table S2.6). A data set of B73 paired end reads (35x) was downloaded from the Sequence Read Archive (accession SRR404240).

For deletion discovery step, F2, PH207 and C103 paired end reads were aligned against B73 AGPv2 genome sequence using novoalign version 3.01.01 (http://www.novocraft.com) (default parameters). Samtools (Li et al., 2009) version 0.1.18 was used to coordinate sort and retain reads with a mapping quality of at least Q30. Duplicated reads were eliminated using MarkDuplicate from the picardtools suite (http://broadinstitute.github.io/picard) version 1.48. Pindel (Ye et al., 2009) version 0.2.5a2 was ran in parallel on each chromosome to perform multi-genotype calling of deletions. Raw formatted results were converted to VCF (Variant Calling Format) standard format using the script Pindel2vcf. BreakDancer (Chen et al., 2009) was used in complement to Pindel but only for F2. Deletions shorter than 100bp were discarded. Deletions spanning a B73 assembly gap or located in regions prone to mis-assemblies such as telomeric, knob and centromeric regions, were also excluded from further analysis using IntersectBed BEDTools (Quinlan and Hall, 2010) version 2.16.1.

For whole genome sequence reconstruction of F2, PH207 and C103 inbred lines, paired-end and mate-pair reads were used together and assembled using ALLPATHs-LG (Gnerre et al., 2011) version R41008. For F2, the script CacheToAllPathsInputs.pl was used to cache the data to use for assembly: 100% of the non-overlapping 230bp insert paired end data set, 100% of the overlapping 170bp insert paired end data set, 30% of the non-overlapping 370bp insert paired end data set, and 100% of the 2.4kb insert mate pair data set. Indeed, only overlapping paired end reads are used by ALLPATHs-LG for building contigs, but the supplementary non-overlapping paired end reads for F2 was used for error correction. RunAllPathsLG was then run for all three genotypes using these optional parameters. For each assembly, the coverage of the gene space was evaluated using BUSCO (Waterhouse et al., 2018) version 3.0.2 using genome mode and maize species (-m geno -sp maize).

B73 paired end reads were successively aligned to ALLPATHs-LG F2, PH207 and C103 genome sequence assemblies. The same tools and parameters used to call deletions against B73 genome were applied to detect B73 deletions against F2, PH207 and C103 genome sequences. For commodity, these B73 deletions will be reciprocally called insertions of F2, PH207 and C103 compared to B73 reference. Again, only insertions smaller than 100bp were discarded, but not the ones spanning assembly gaps as they were real assembly gaps (with approximate size inferred from paired reads average distance) and not “unsized” gaps like in B73 genome. When possible, insertions were anchored onto B73 AGPv2 genome

sequence using a dedicated pipeline combining Megablast version 2.2.19 (AltschuP et al., 1990) and Age version 0.4 (https://github.com/abyzovlab/AGE). Again, insertions that could be anchored on B73 reference and were overlapping regions prone to mis-assemblies such as telomeric, knob and centromeric regions, were also excluded from further analysis using IntersectBed.

F2 specific sequences coming either from the no-map approach (Figure S2.2) or from the work of Darracq et al. (2018) were included as such, without any further filtering.

The multiple references and approaches used during the InDels discovery step led to a set of InDels with various levels of redundancy. Some "intra-tool" redundancy was found (*eg.* multiple calls found by one tool within the same genotype at highly polymorphic loci). These "ambiguous" calls were systematically identified using the Bedtools suite version 2.16.1 (Quinlan and Hall, 2010) and eliminated. Moreover, for F2 deletions, some "inter-approach" redundancy was also expected and eliminated using intersectBed utility also from the Bedtools suite. When redundancy was found, Pindel calls were preferred to BreakDancer ones because they had precise breakpoints and contained also the calls for PH207 and C103. The same filter was applied to all insertions that could be anchored to the B73 genome sequence. Furthermore, for non-anchored InDels, in order to avoid too much redundancy in internal genotyping probes design, RepeatMasker (http://www.repeatmasker.org) was used to mask redundant regions by similarity using an iterative approach. First, "ALLPATHs-LG assembly" F2 insertions were masked with "ABySS assembly" F2 insertions (at least 95% of identity) to generate a non-redundant set of F2 insertions. Then C103 insertions were masked with F2 insertions (at 90% of identity), PH207 insertions were masked with C103 and F2 insertions (90%), and finally F2 No-Map specific sequences were masked with PH207, C103 and F2 insertions (90%).

Design of Affymetrix Axiom array

Preparation of sequences for probes for design

To identify presence/absence regions (PARs) within InDel sequences more suitable for the design of "off-target" probes, we used the genometools Tallymer utility (Gremme et al., 2013) version 1.5.6 to create two indexes for B73, F2, PH207 and C103: one from their genome assemblies (17-mers with a minimal occurrence of 1) and one from a 5x genome equivalent subset of their raw sequenced data (17-mers with a minimal occurrence of 5). Then B73 genome was iteratively annotated with the script tallymer2gff3.plx (options used: -k 17 -min 35 -occ 1|5 depending on the index) to identify regions not covered by F2, PH207 and C103 kmers. Reciprocally, the two F2 draft genomes, PH207 and C103 ALLPATHs-LG draft genomes were ran through the same procedure to identify regions not covered by B73 kmers. The gff files generated by this process were then used in combination with gff files of repeats annotated with RepeatMasker to define PARs of a minimum size of 35bp for each type of InDel and each draft genome.

BP preparation

Breakpoints could be targeted by probes (Figure 2.1 A) providing that the nucleotide flanking the breakpoint at the beginning of the deleted sequence were different from the nucleotide right after the end of deleted sequence (and reciprocally on the reverse strand). Type I and type III breakpoints without

micro-homology sequence can be submitted to Affymetrix' straightforward design procedure whereas type II breakpoints have to go through an iterative design process, shifting the sequence by one base on each attempt until reaching a discriminative position. This iterative process stops after 5bp.

Probes scoring

All potential probes were evaluated in an in-silico analysis to predict their microarray performance. A p-convert value, which arises from a random forest model intended to predict the probability that the SNP will convert on the array, was determined for all probes. The model considers factors including probe sequence, binding energies, and the expected degree of non-specific binding and hybridization to multiple genomic regions. This degree of non-specific binding is estimated calculating 16-mer hit counts, which is the number of times all 16 bp sequences in the 30 bp flanking region from either side of the SNP have a matched sequence in the genome. These scores were generated both for forward and reverse probes. A probeset is recommended if p-convert>=0.6 and there are no expected polymorphisms in the flanking region. A probeset is neutral if p-convert>=0.4, the number of expected polymorphisms in the flanking region is less than 3, and the polymorphisms are further than 21 bp of the variant of interest. Probesets not falling into these two categories are scored as *not recommended*. Probesets that cannot be designed are scored as *not possible*.

Probes selection

Concerning OTV and MONO probes, we applied three successive filtering steps. First, we selected only probes classified as recommended and neutral based their scoring, with no more than one hit on B73 reference genome for deletion probes and no hit at all for insertion probes were selected. After this step, 204,213 OTV probes and 18,884,827 MONO probes remained. Secondly, only probes with more than 70% in PARs were kept. An additional filtering step was implemented specifically for MONO probes to optimize probes distribution along the targeted PARs. To optimize probes distribution along the targeted PARs, these ones were cut in 75bp windows using windowmaker (Bedtools) and the MONO probe with the highest p-convert value was selected for each window. In case there were InDels with less than 4 MONO probes selected using 75bp windows, these probes were eliminated, and a second iteration was made using this time 50bp windows, followed by a last iteration with 25bp windows. This gave at this point a total of 616,286 probes including BP and OTV probes targeting 108,703 InDels (90% of InDels selected for design).

We completed the design by rescuing 6,219 OTV and 3,441 MONO probes from InDels or PARs still not targeted by any probes, bringing the total number of probes selected to 625,946 to target 109,292 InDels. At the last step, duplicated probeset were removed based on their sequence by Affymetrix during the chip design procedure, leaving 662,772 probeset (105,927 InDels) corresponding to 1,404,570 different probes to be arrayed on the array.

Genotyping of 105k InDels on 480 maize DNA samples

Plant Material for genotyping

662,772 probes selected in the array were used to genotype 480 diverse DNA samples including 440 inbred lines, 24 highly recombinant inbred lines and 16 F1 hybrids. Both F1 hybrids (obtained by crossing inbred lines) and their parental inbred lines were genotyped on the array but seed lots used to produce F1 hybrids and those used to extract DNA for genotyping were different. Among these 480 DNAs, 13 inbred lines were genotyped using two different DNAs from two different seed lots and was used to evaluate the reproducibility of the genotyping (Table S2.4). DNA samples of one F1 hybrid were also genotyped 6 times.

DNA for genotyping were extracted from leaves following a NaBisulfite method modified from Tai and Tanksley (1990) and Dellaporta et al. (1983).

Variant calling using Affymetrix algorithm

DNA samples from 480 individuals were hybridized to array using the Affymetrix system. The genotyping, sample QC, and marker filtering was performed according to the Axiom Best Practice genotyping analysis workflow. Genotype calls and classifications were generated from the hybridization signals in the form of CEL files using the Affymetrix Power Tools (APT) and the SNPolisher package for R according to the Axiom Genotyping Solution Data Analysis Guide.

The APT results were then post-processed using SNPolisher, which is an R package specifically designed by Affymetrix. Markers metrics were generated using the *Ps_Metrics* function. The markers QC metrics were used to classify probesets into 14 categories (Figure S2.13) using the *Ps_Classification* and *Ps_Classification_Supplemental* functions with all default setting for diploid, except for an empirically determined, more stringent heterozygous variance filter (AB.varY.Z.cut=2.6). Example of clusters from each classification were visualized using the *Ps_Visualization* function (Figure S2.13).

Each type of probe had a dedicated algorithm (Figure 2.4 and Figure S2.8) to call genotyping according to expected behavior from the probe design. Variant were preferentially selected as recommended if they were exhibiting stable category assignments with clearly separated clusters. Each variant was ranked into a category (Figure S2.13) at each step of the algorithm.

Algorithms used to convert BP and OTV were similar, as BP and OTV behaved like classical SNP. For initial genotype calling, a priori cluster positions were used since no information about expected position was available. A first analysis was performed according to Affymetrix recommendations. Secondly, level of inbreeding was considering for a posteriori cluster definition because of the high amount of inbred lines in the panel. This parameter took values from 0 for fully heterozygous to 16 for completely homozygous samples. For OTV and BP algorithms, an inbred penalty of 4 (lower penalty for inbred species) was applied to try to re-labelled probes that fall into categories: CallRateBelowThreshold (CRBT), HomHomResolution (HHR), NoMinorHom (NMH), Other and UnexpectedHeterozygosity after the first cluster analysis. Markers that were classified as OTV may also be considered recommended after *OTV_caller* function has been used to re-label the genotype calls. The SNPolisher *OTV_Caller* function

performed post-processing analysis to identify miscalled AB clustering and identify which samples should be in the OTV cluster and which samples should remain in the AA, AB, or BB clusters. Samples in the OTV cluster were re-labelled as OTV. Finally, the recommended markers list is created by combining the list of markers that are classified into the recommended categories (PolyHighResolution (PHR), MonoHighResolution (MHR), and OTV).

BP and OTV probes exhibited only two clusters (AA or BB and OTV) should fall into monomorphic classification and classify as not recommended. A new MONO algorithm was developed (Figure 2.4) because we expected for this probes fluorescence pattern no polymorphism in the present sequence (Figure 2.1 C). Contrary to BP and OTV algorithm, *OTV_caller* was used before inbred penalty for MONO probes analysis. To classify monomorphic sequence genotyping, the *OTV_Caller* function was called and as we expected monomorphic genotyping, only MHR and NMH were considered as recommended. Other monomorphic probes are then analyzed with an inbred penalty of 16 (highest level) to re-labelled probes considering maximum level of heterozygosity. Finally, a new function called *Hom2OTV* was used to classified probes exhibiting two homozygous clusters but with a different position in the Y-axis (high and low position). This function tried to decide if the difference of contrast represent actually one homozygous and one OTV cluster as we expect (respectively presence and absence of the corresponding probe sequences).There are no parameters in this function. The lower intensity homozygous cluster is recalled as OTV.

Evaluation of genotyping quality

We compared the genotyping for 479,027 probes from InDel array with expected genotyping from resequencing of 4 inbred lines used to discover InDels: B73, F2, PH207 and C103. Expected genotyping was built from alignment of probes sequences on reference genome B73 and de novo assembly of 3 inbred lines (F2, PH207 and C103) with Blast software. Sequences were considered present in lines when the probes were aligned with less than 5% of mismatch and absent when not.

Genotyping consistency for B73, F2, PH207 and C103 was calculated between expected and observed genotyping for “presence” and “absence” (Table 2.4). For this purpose, Affymetrix genotyping was converted into two genotypes, present and absent and hemizygote from BP were considered as missing data. Consistency of Presence/Absence genotypes between resequencing and array genotyping was analyzed for four individuals (B73, F2, PH207, C103) according to probe types (BP, OTV, MONO): Number of similar genotypes between observed and expected/number of genotypes observed. Note that the seed lot used for B73 and F2 genotyping is different from this used for InDel discovery, while it is the same one for inbred lines PH207 and C103.

In order to evaluate the consistency of probes genotyping within InDels (Figure 2.5), we used 24 inbred lines including 20 inbred lines from a core collection (Darracq et al., 2018) and the 4 inbred lines used for InDel discovery. From 479,027 probes, we selected 294,650 polymorphic probes and totally consistent between sequencing and array genotyping in order to limit the genotyping errors due either to array or sequencing. These probes allowed us to genotype 72,555 InDels. We selected 48,486 polymorphic InDels that are genotyped with at least two probes (corresponding to 270,581 probes), and calculated the frequency of presence allele for each InDel and inbred lines.

To evaluate quality of genotyping for hybrids, we predicted the genotype of hybrids based on the genotyping of 2 parental lines for 20,370 BP probes without OTV cluster. This expected genotype for hybrids was then compared with the observed genotyping from array of the corresponding hybrid. With following formula (Number of similar alleles (homozygous or hemizygous) between expected and observed)/(number of expected alleles (homozygous or hemizygous)).

To evaluate the reproducibility of the 479,027 probes of the array (Table S2.4), we compared genotyping of 13 duplicated inbred lines (A554, A632, A654, B73, C103, CO255, D105, EP1, F2, F252, KUI3, Oh43, and W117) originated from different seed sources. The genotyping of these 13 duplicated lines were also compared using 43,982 SNPs from the Illumina 50K SNP array.

Diversity analysis

We performed diversity analysis on 362 inbred lines from an association panel representing a wide range of diversity (Camus-Kulandaivelu et al., 2006; Bouchet et al., 2013) using genotyping from our InDels Affymetrix Axiom. We compared these results with diversity analysis performed on same lines using genotyping of Illumina 50K SNP array (Ganal et al., 2011). Genotyping of InDels were treated as bi-allelic 0/2 for “present” and “absent” respectively.

To perform diversity analysis, we first selected 237,629 probes among the 479,027 probes for which (i) the clustering observed were consistent with expected one (Table 2.3) and (ii) for which genotyping produced by our array for 4 lines used for discovered InDels were totally consistent with genotyping based on the alignment of probes on genome assemblies using BLAST software. We filtered out 219,068 probes based on their genotyping quality (missing data rate below 20%, heterozygous rate below 15% and minor allele frequency above 5%). In the end, we selected a single probe by InDel considering both genotyping and Affymetrix quality leading to a set of 57,824 probes genotyping 57,824 InDel to analyze diversity in 362 inbred lines.

We estimated two kinship matrices between 362 lines using “identity by state” estimators (IBS) based on 57,824 InDels (Figure 2.6). Kinship matrix was estimated with the “ibd” function in R package GenABEL (Aulchenko et al., 2007). Genetic structuration was estimated using only 28,143 panzea SNP using admixture software (Alexander et al., 2009). We selected Admixture results corresponding to five genetic groups (Q=5) since it corresponded to the number of genetics group defined in previous studies using panzea SNP from Illumina 50K (Bouchet et al., 2013). Lines were assigned to one genetic group providing that the probability of assignment to the groups were superior to 0.6 whereas lines below this threshold were considered “admixed”. In order to compare genetic structuration based on InDels and SNPs, we performed Principal Coordinate Analysis on genetic distance between lines with (362 lines) and without F2 and B73 (360 lines) based on their dissimilarity (1-IBS) using InDels. Each line were plotted on two first plan of PcoA and colored according to assignment to 5 genetics groups (Figure 2.6).

Data Access

The array content is available at https://doi.org/10.15454/DWB4UT

Acknowledgements

This work was supported by the project CNV-MAIZE (ANR-10-GENM-003) and the project Investement for the future AMAIZING ANR-10-BTBR-01 (ANR-PIA AMAIZING) and France Agrimer. PhD student C. Mabire is jointly funded by the program CNV4sel in the framework of metaprogram Selgen and by the Plant Biology and Breeding department of the French National Institute for Agricultural Research (INRA). We are very grateful to Patrick Schnable to provide a subset of Presence/Absent Variants coming from their RNAseq and Sequence capture approach and for his helpful discussion. We are also very grateful to Alain Charcosset for his helpful discussion and comments on the manuscript.

Disclosure declaration

Ali Pirani is an employee of Affymetrix.

Authors' contributions

SDN designed and supervised the study and conducted CNVMaize project

CM, JD and SDN drafted the manuscript, CV and JJ corrected the manuscript;

NR, SDN, JPP and SP conceived the array, AP, SDN, JJ and JD designed the array;

AP develop calling Affymetrix pipelines and did the call of InDel;

JPP, JJ and CV contributed to the sequencing;

JD, AD, HR and JJ performed the InDel discovery, JD and JJ build genome assemblies, JJ and AD discovered SNP within InDels;

CM evaluated the quality of genotyping and conducted genetic diversity analysis;

DM and VC did DNA extraction and prepared the samples for arrays genotyping;

Supplementary figures

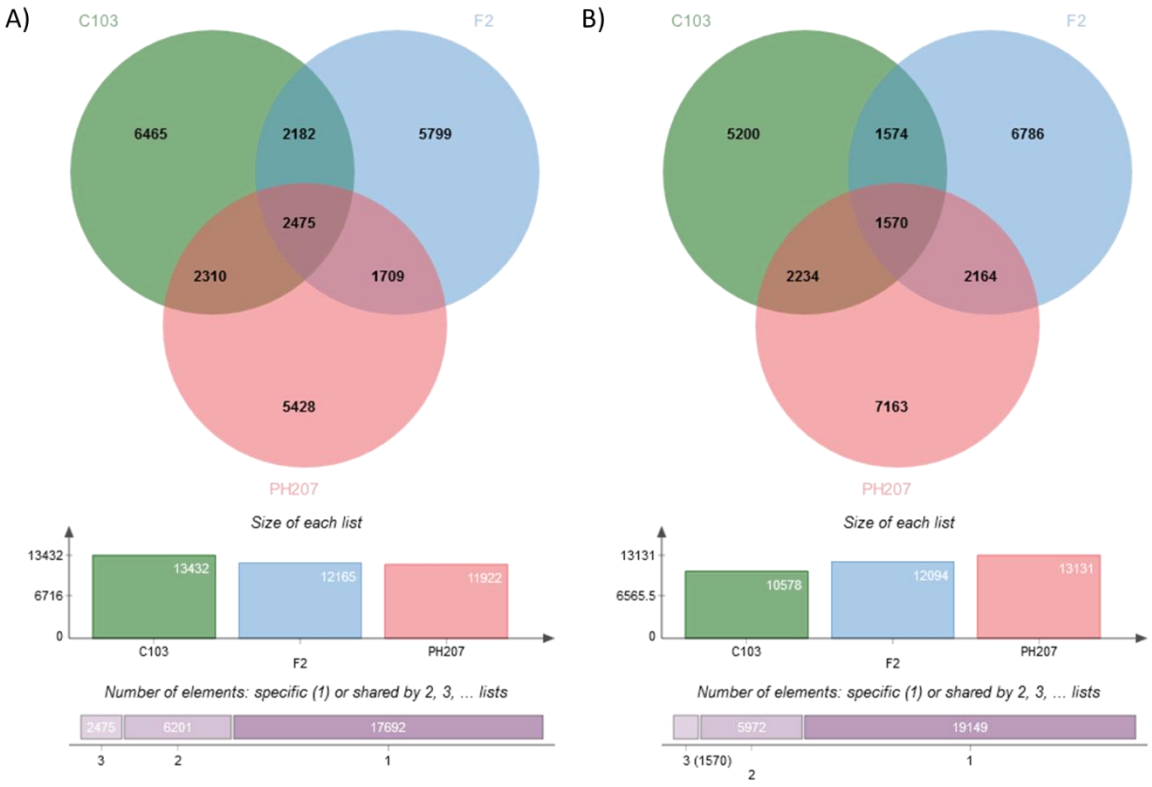


Figure S2.1: Number and complementarity of deletions (A) and insertions (B) regarding B73 reference genome discovered between F2, PH207 and C103 inbred lines and B73.

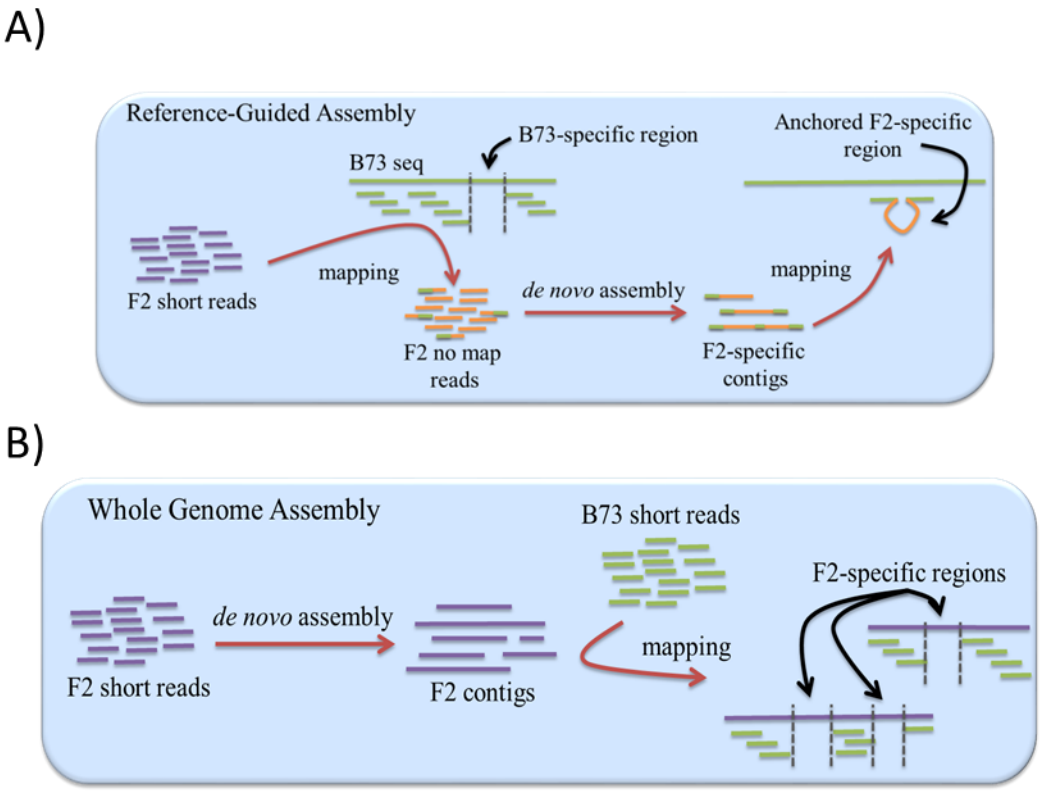


Figure S2: Description of two approaches used to discover InDel using resequencing data of DNA: A) reference guided assembly ("no map" approach) used only on F2 and B) whole genome assembly used on F2 PH207 and C103

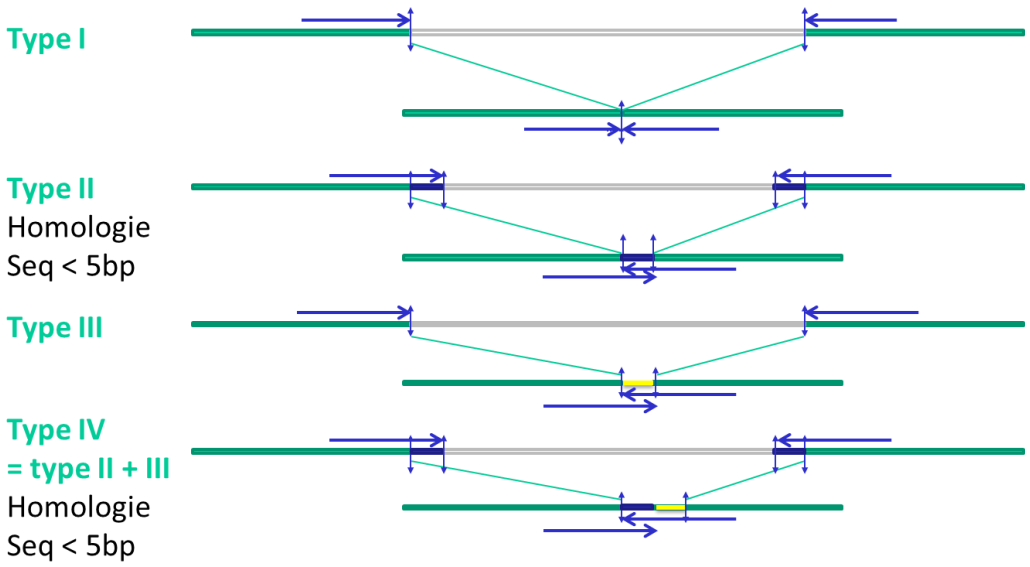


Figure S2.3: Schematic representation of four different breakpoint types identified by PINDEL at InDel breakpoints according to the presence of micro-homology sequence or not in place of the deleted sequence. Green, gray, blue and yellow color on the Horizontal line represented the common sequence between lines, the internal sequence of InDels, the sequence of microhomology and the shorter sequence replacing the longer sequence initially present, respectively. Horizontal blue arrows represented the forward and reverse BP probes designed at breakpoint site (vertical blue lines).

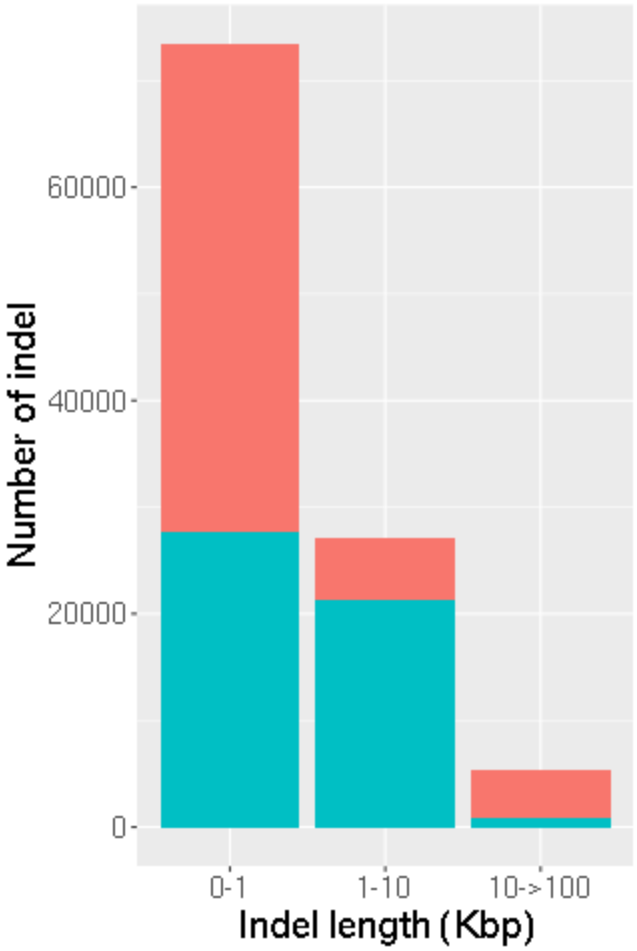


Figure S2.4: Distribution of 105,927 InDels genotyped by the array according to their length (kbp). Red Color indicates the proportion of InDels with (red) or without (blue) presence/absence regions for the 3 classes of InDel length.

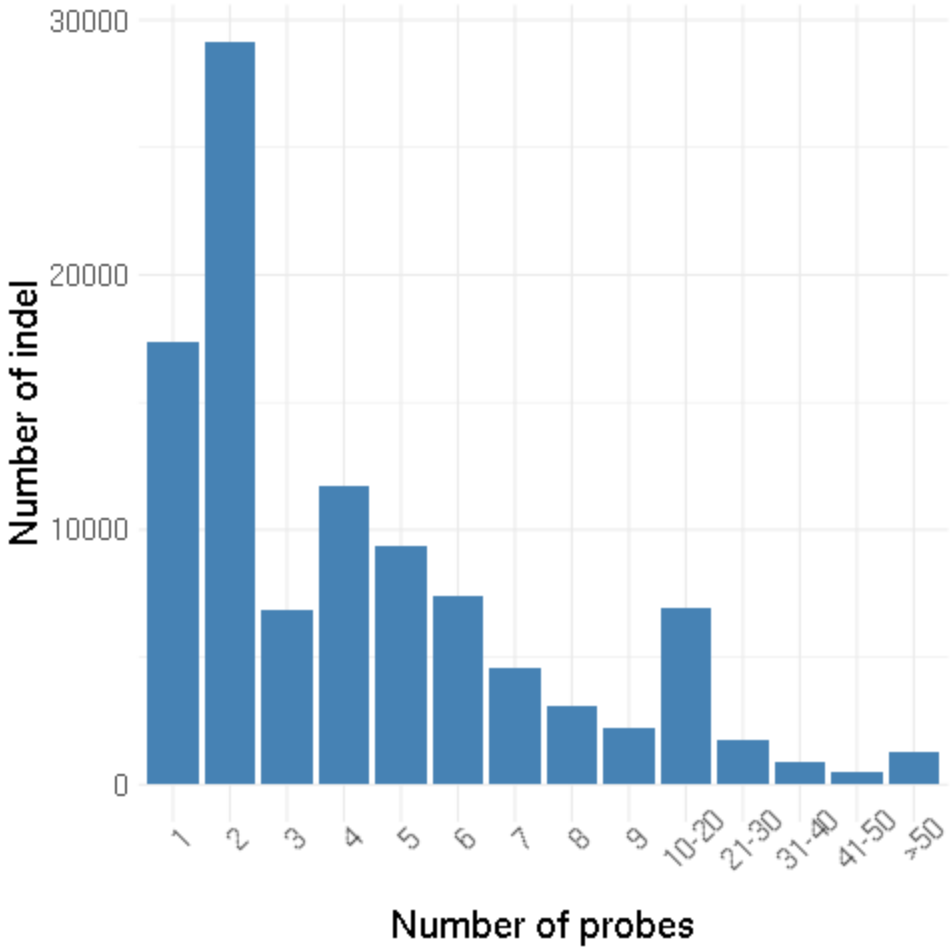


Figure S2.5: Distribution of probe number per InDel for 105,927 InDels genotyped with the array. InDels with at least 50 probes were represented in the same category.

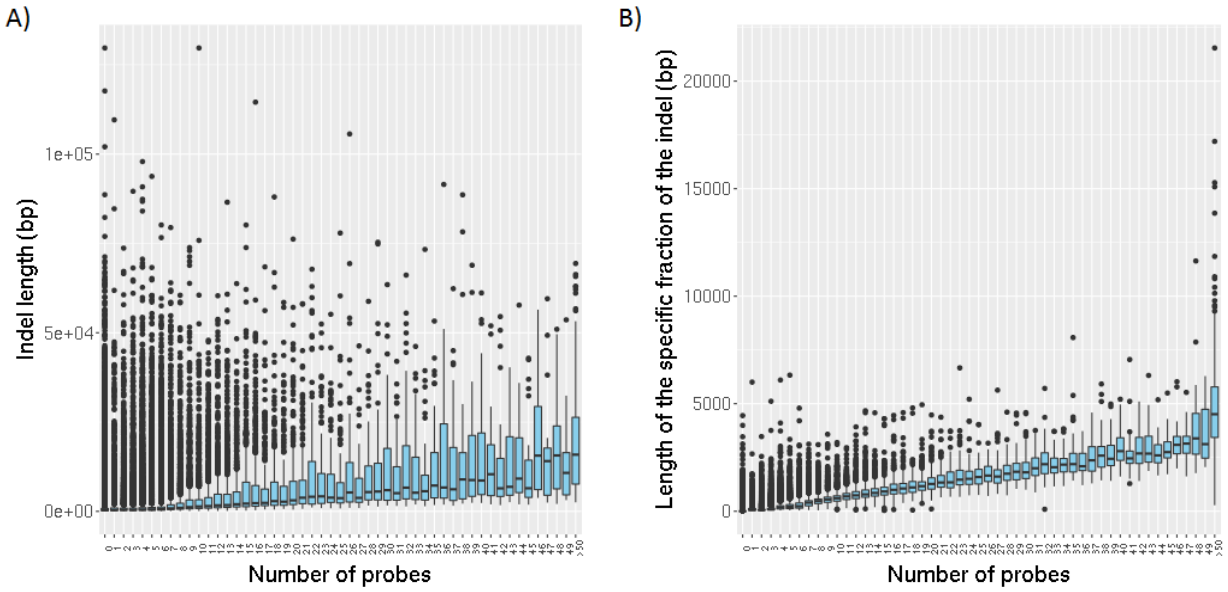


Figure S2.6: Relationship between probes number genotyping the InDels and A) the InDels length B) cumulated length of specific sequence (PARs) within InDels.

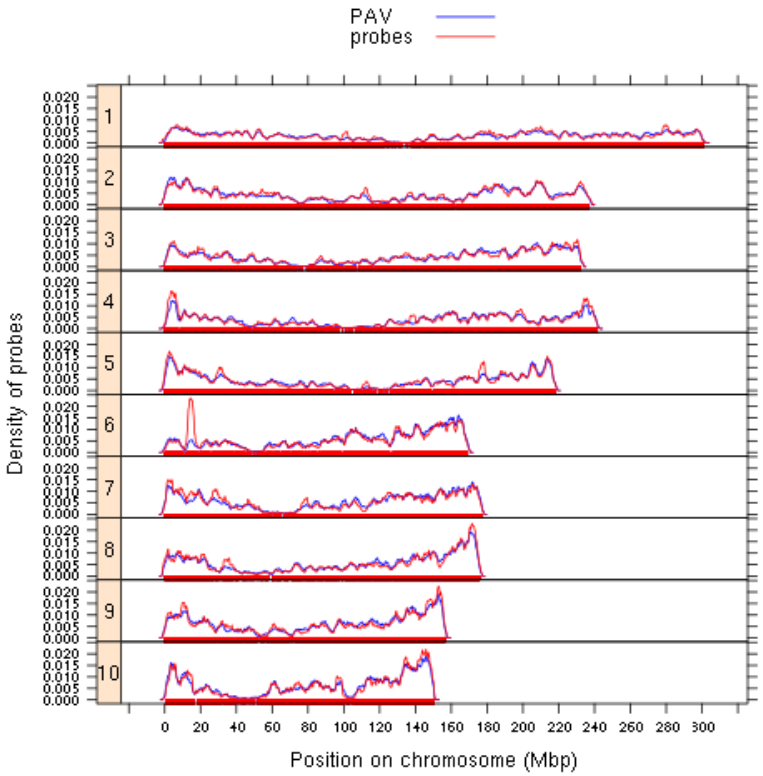


Figure S2.7: Variation of probes and InDels density across the 10 maize chromosomes. Red and blue lines represented the density of 237,257 probes and 43,117 InDels anchored in the maize genome, respectively.

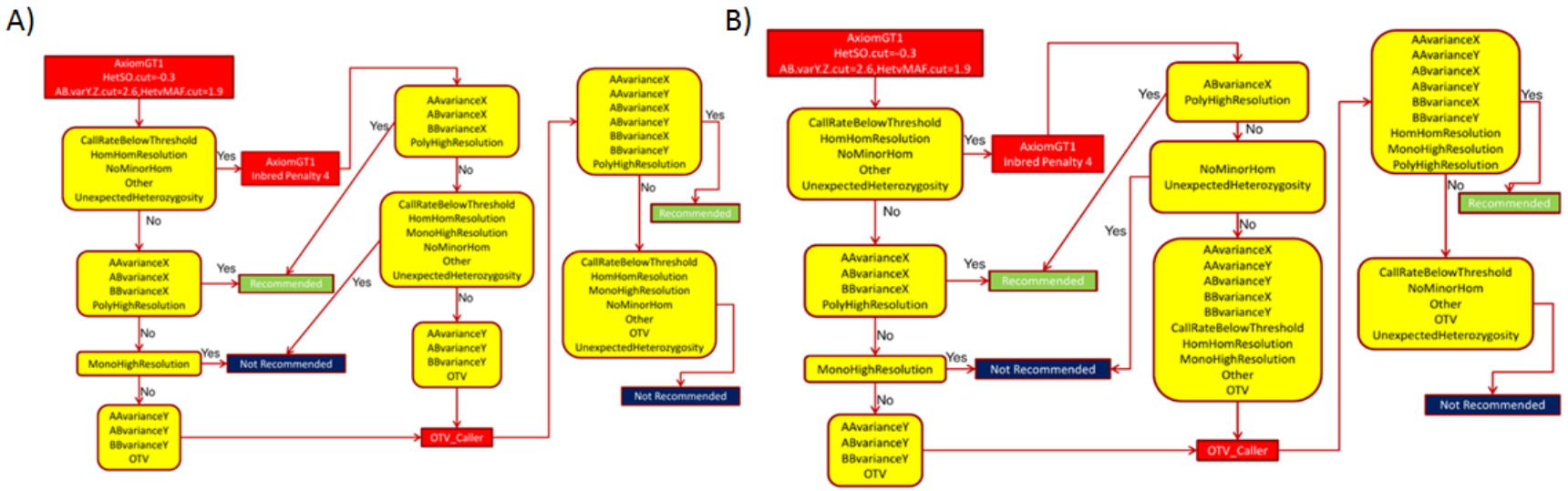


Figure S2.8: Two dedicated Affymetrix pipelines used for calling InDel polymorphisms from the fluorescent intensity variation of BP probes (A), and OTV probes (B). Each probe was classified into different categories according to the number of clusters, the call rate and quality metrics of the clustering based on the position, variance and separation of different cluster. In order to retrieve the best clustering for each probe, successive step of clustering using different clustering algorithms (Red square, Axiom GT1, OTV caller, Hom2OTV) or/and with different parameters. According to their classification at each step (yellow square) and threshold used for quality metrics, probes could be classified as recommended (green square), not recommended (blue sqare) or to be submitted to another step. According to probes origin (BP, OTV), algorithms and parameters of these different steps of clustering varied. For instance, BP and OTV did not differ for the first step of clustering but differed for the following step since more categories were called with “OTV caller” for OTV probes. Except this small switch for some categories, BP and OTV probes were called globally in the same way (A, B). At the end, all probes were classified into 14 categories either as recommended or not recommended depending on threshold.

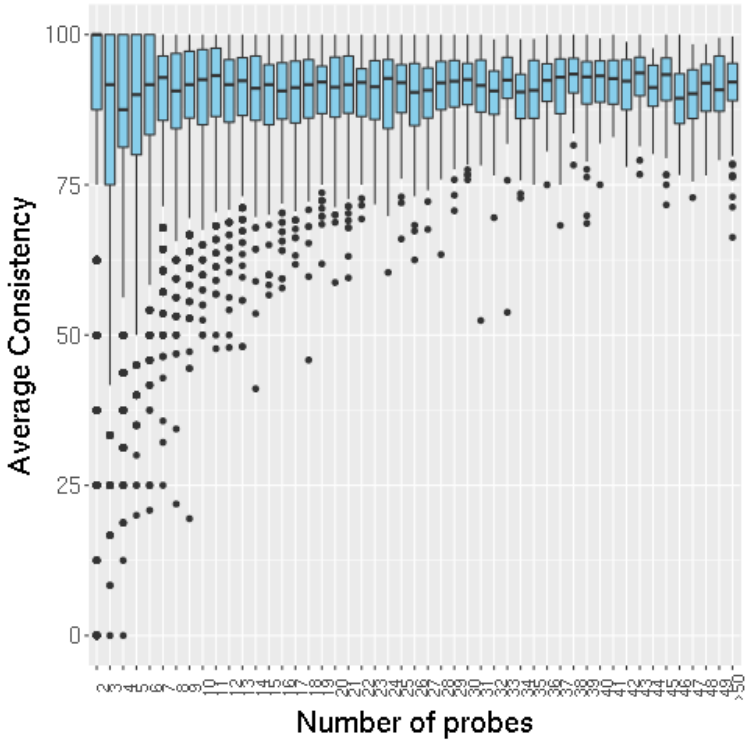


Figure S2.9: Variation of the distribution of the average consistency rate (%) of InDel between expected and observed genotyping of probes according to number of probes within InDel. Only InDels genotyped with at least two probes were considered and InDels with more than 50 probes were classified in one category (>50).

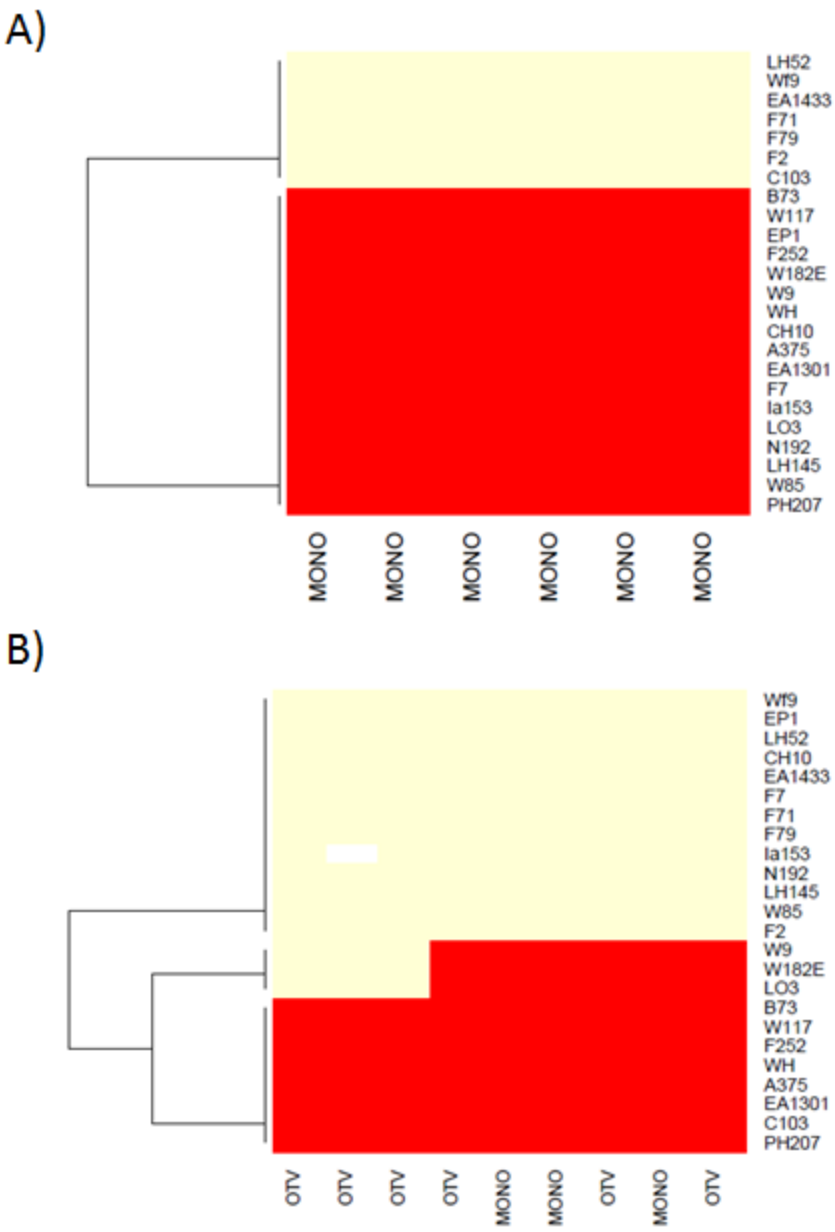


Figure S2.10: Haplotype of two InDels genotyped with multiple probes (in column) for 24 individuals (in rows). Inbred lines were ordered by hierarchical clustering according to their similarity based on the genotyping of individual probes A) InDel CNVMAIZE_INS_105228 genotyped with 6 probes displayed two haplotypes indicating either that sequence was totally present (yellow) or absent (red) in 24 individuals. B) InDel CNVMAIZE_INS_2516 genotyped with 8 probes displayed 3 different haplotypes according to probes genotyping indicating that sequence was only partially absent in some individuals. Yellow and red indicated that the sequence of the probes was either present or absent, respectively

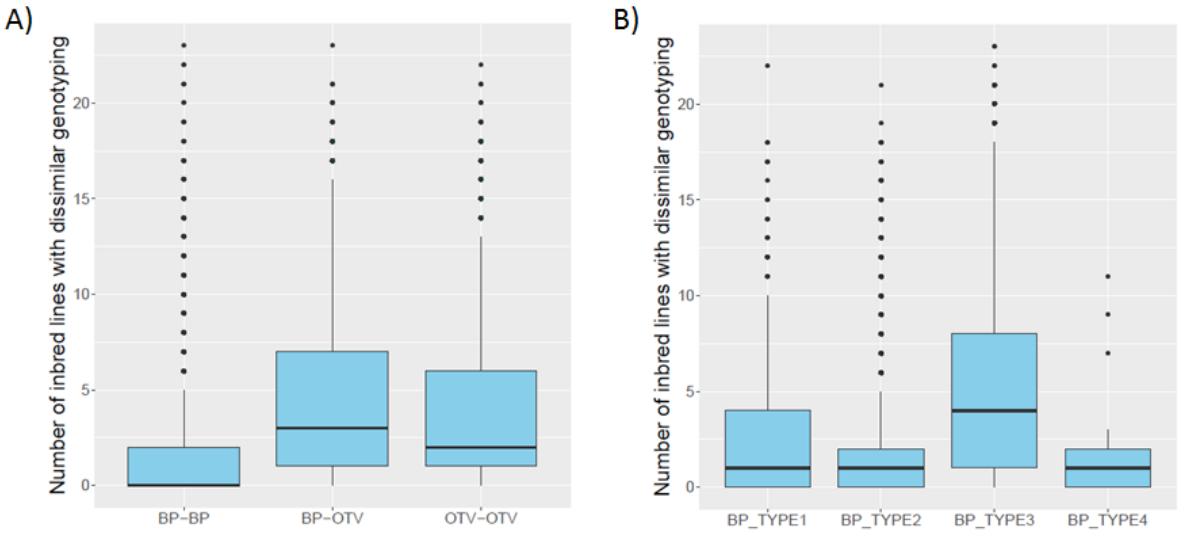
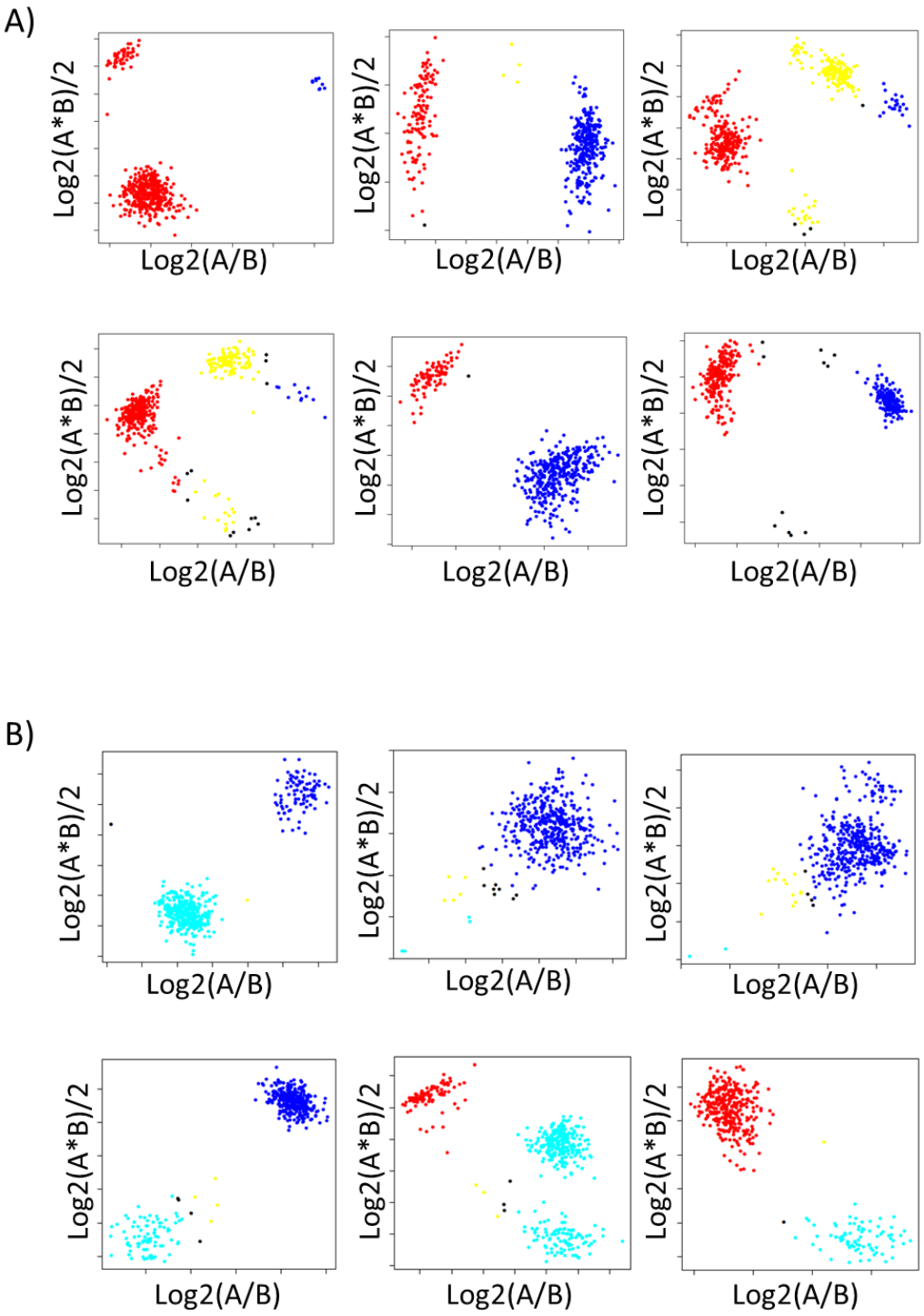


Figure S2.11: Comparison of genotyping between forward and reverse BPs for 8,116 InDels and 24 individuals. (A) Effect of the BP classification on the distribution of the individual number displaying dissimilar genotyping between forward and reverse BP. (B), Effect of BP types on the distribution of the individual number displaying dissimilar genotyping between forward and reverse BP


A)
Log2(A*B)/2
Log2(A/B)
B)
Log2(A*B)/2
Log2(A/B)

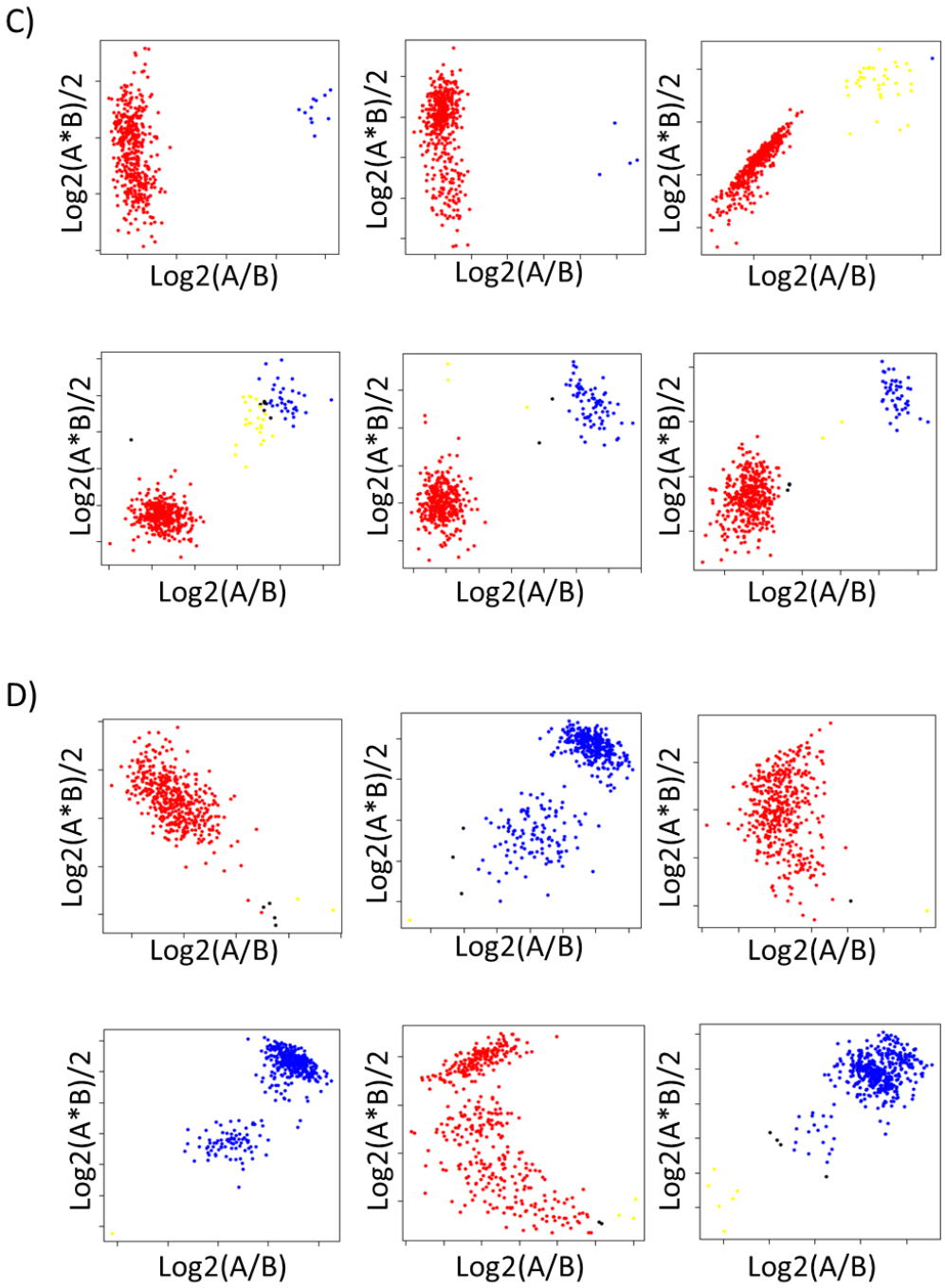


Figure S2.12: Example of clustering for 6 randomly probes among A) 17,562 OTVs classified as SNP, B) 68,562 MONOs probes with an unexpected heterozygous cluster, C) 1,981 MONOs classified as SNP, and D) 9,525 MONO probes with an unexpected heterozygous cluster but without cluster for absence of the sequence. Blue, red and yellow dots indicated that the sequence was present and were homozygous for allele A (AA) allele B (BB) or heterozygous (AB). Cyan dots indicated that the probes did not hybridize and that the sequence of probes was therefore absent.

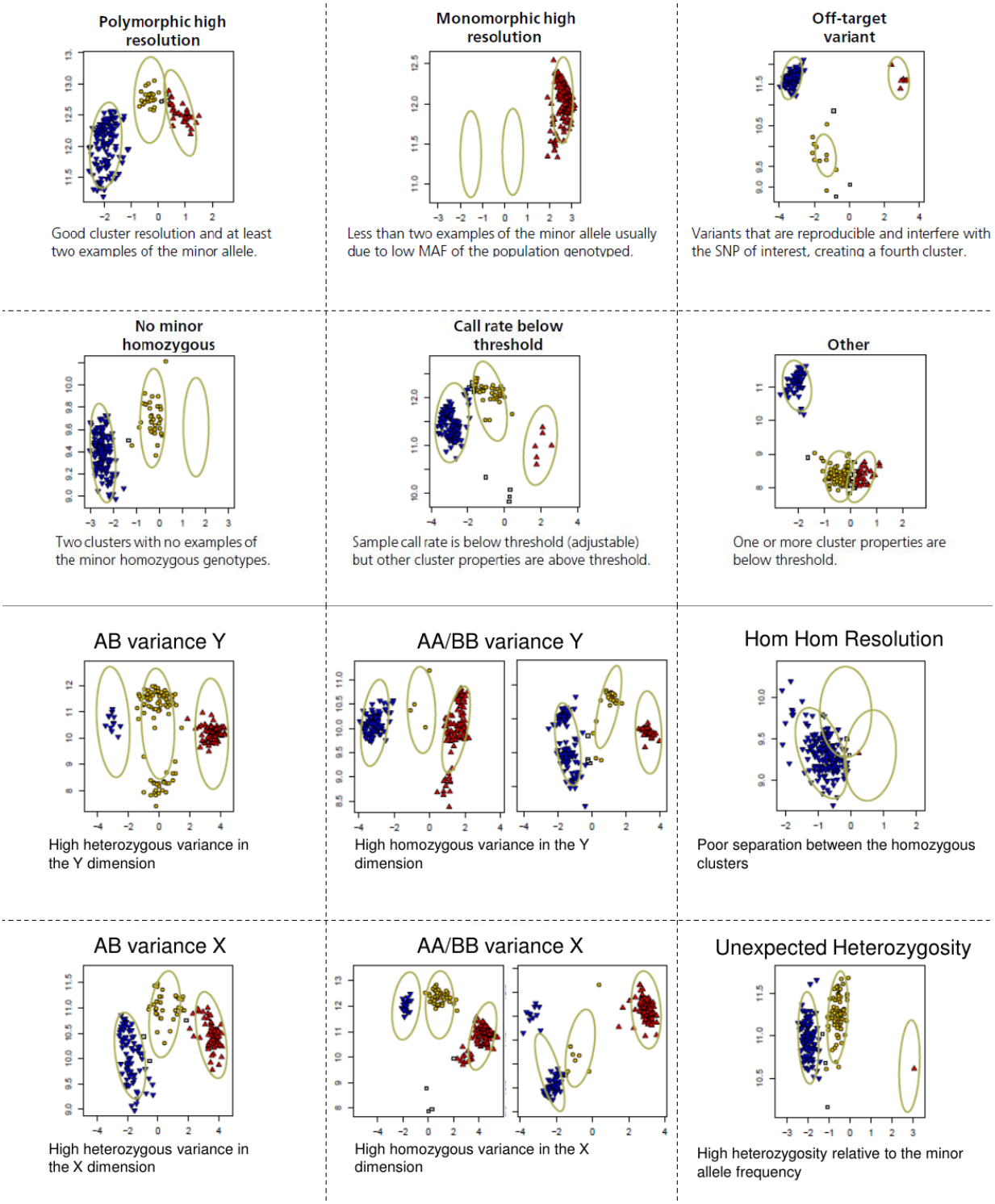


Figure S2.13: Example of clustering based on probes fluorescence (intensity in y-axis and contrast in x-axis), for 14 different classifications of probes assigned by Affymetrix algorithms based on cluster number, separation of cluster and variance of fluorescent intensity and contrast within clusters and call rate. For each classification, particular characteristics is described under the figure.

Supplementary tables

Table S2.1: Classification by Affymetrix pipeline of 84,994 BP probes based on cluster number, separation, variance and call rate. A) probes recommended for genotyping by the pipelines B) Probes not recommended for genotyping. "Markers" represented the number of probes for each classification (Affy Classification) at a particular step of the algorithm (Pipeline step). For each classification, we shown the number of InDel (Nbr_InDel) targeted by corresponding probes.

A)

		Recommended		
Affy Classification	**PipelineStep**	**Nbr. Markers**	**MarkersRate**	**Nbr_InDel**
AAvarianceX	AxiomGT1	123	0,27%	123
AAvarianceX	Inbred	139	0,30%	139
AAvarianceX	OTV	242	0,52%	241
AAvarianceY	OTV	760	1,64%	752
ABvarianceX	AxiomGT1	36	0,08%	36
ABvarianceX	Inbred	2	0,00%	2
ABvarianceX	OTV	25	0,05%	25
ABvarianceY	OTV	223	0,48%	221
BBvarianceX	AxiomGT1	122	0,26%	122
BBvarianceX	Inbred	122	0,26%	122
BBvarianceX	OTV	254	0,55%	253
BBvarianceY	OTV	759	1,64%	749
PolyHighResolution	AxiomGT1	14932	32,19%	12981
PolyHighResolution	Inbred	4653	10,03%	4484
PolyHighResolution	OTV	23990	51,72%	20061
	Total	46382		34232

B)

		Not Recommended		
Affy Classification	**PipelineStep**	**Nbr. Markers**	**MarkersRate**	**Nbr_InDel**
CallRateBelowThreshold	inbred	2915	7,55%	2845
HomHomResolution	inbred	2228	5,77%	2185
MonoHighResolution	inbred	4972	12,88%	4676
NoMinorHom	inbred	2583	6,69%	2457
Other	inbred	20533	53,18%	18302
UnexpectedHeterozygosity	inbred	21	0,05%	21
CallRateBelowThreshold	OTV	4345	11,25%	4256
HomHomResolution	OTV	2	0,01%	2
MonoHighResolution	OTV	112	0,29%	111
NoMinorHom	OTV	304	0,79%	300
Other	OTV	378	0,98%	373
OTV	OTV	211	0,55%	210
UnexpectedHeterozygosity	OTV	8	0,02%	8
	Total	38612		31049

Table S2.2: Classification by Affymetrix pipeline of 163,278 OTV probes algorithm based on cluster number, separation, variance and call rate. A) Probes recommended by the pipeline for genotyping B) Probes not recommended probes for genotyping. "Markers" represented the number of probes for each classification (Affy Classification) at a particular step of the algorithm (Pipeline step). For each classification, we shown the number of InDel (Nbr_InDel) targeted by corresponding probes.

A)

Recommended				
Affy Classification	**PipelineStep**	**Nbr. Markers**	**MarkersRate**	**Nbr_InDel**
AAvarianceX	AxiomGT1	102	0,11%	102
AAvarianceX	OTV	784	0,81%	666
AAvarianceY	OTV	1291	1,33%	1100
ABvarianceX	AxiomGT1	20	0,02%	19
ABvarianceX	Inbred	1	0,00%	1
ABvarianceX	OTV	145	0,15%	139
ABvarianceY	OTV	658	0,68%	570
BBvarianceX	AxiomGT1	127	0,13%	124
BBvarianceX	OTV	1030	1,06%	903
BBvarianceY	OTV	1352	1,40%	1128
HomHomResolution	OTV	71	0,07%	71
MonoHighResolution	OTV	506	0,52%	467
PolyHighResolution	AxiomGT1	11485	11,86%	5179
PolyHighResolution	Inbred	4913	5,07%	3285
PolyHighResolution	OTV	74382	76,79%	13097
	Total	96867		15064

B)

Not Recommended				
Affy Classification	**PipelineStep**	**Nbr. Markers**	**MarkersRate**	**Nbr_InDel**
NoMinorHom	inbred	5746	8,65%	3705
UnexpectedHeterozygosity	inbred	32	0,05%	31
AAvarianceY	OTV	1	0,00%	1
CallRateBelowThreshold	OTV	31066	46,78%	10471
MonoHighResolution	OTV	522	0,79%	486
NoMinorHom	OTV	2758	4,15%	2218
Other	OTV	25086	37,77%	9297
OTV	OTV	686	1,03%	637
PolyHighResolution	OTV	2	0,00%	2
UnexpectedHeterozygosity	OTV	512	0,77%	470
	Total	66411		14293

Table S2.3: Classification by Affymetrix pipeline of 414,500 MONO probes based on cluster number, separation and variance and call rate. A) Probes recommended for genotyping by the pipeline B) Probes not recommended for genotyping. "Markers" represented the number of probes for each classification (Affy Classification) at a particular step of the algorithm (Pipeline step). For each classification, we shown the number of InDel (Nbr_InDel) targeted by corresponding probes.

A)

Recommended				
Affy Classification	**PipelineStep**	**Nbr. Markers**	**MarkersRate**	**Nbr_InDel**
AAvarianceX	Inbred	11	0,00%	11
AAvarianceX	Poly	309	0,09%	305
AAvarianceY	Poly	384	0,11%	382
ABvarianceX	Poly	84	0,03%	84
ABvarianceY	Poly	256	0,08%	253
BBvarianceX	Inbred	3	0,00%	3
BBvarianceX	Poly	156	0,05%	156
BBvarianceY	Poly	878	0,26%	856
HomHomResolution	Inbred	3867	1,15%	3622
HomHomResolution	Poly	4	0,00%	4
MonoHighResolution	Mono	19771	5,89%	13300
MonoHighResolution	Poly	27165	8,09%	14222
NoMinorHom	Mono	25820	7,69%	16716
Other	Inbred	213759	63,66%	55225
OTV	Inbred	11317	3,37%	8933
PolyHighResolution	Inbred	16394	4,88%	11036
PolyHighResolution	Poly	15600	4,65%	12531
	Total	335778		63597

B)

Not Recommended				
Affy Classification	**PipelineStep**	**Nbr. Markers**	**MarkersRate**	**Nbr_InDel**
CallRateBelowThreshold	Poly	7401	9,40%	6466
NoMinorHom	Poly	3232	4,11%	2905
Other	Poly	17850	22,67%	13539
OTV	Poly	3194	4,06%	2912
UnexpectedHeterozygosity	Poly	23	0,03%	23
AAvarianceX	Inbred	106	0,13%	106
AAvarianceY	Inbred	144	0,18%	143
ABvarianceX	Inbred	1	0,00%	1
ABvarianceY	Inbred	18	0,02%	18
BBvarianceX	Inbred	32	0,04%	32
BBvarianceY	Inbred	138	0,18%	138
CallRateBelowThreshold	Inbred	2717	3,45%	2583
HomHomResolution	Inbred	5380	6,83%	4940
MonoHighResolution	Inbred	22627	28,74%	15509
NoMinorHom	Inbred	3404	4,32%	3160
Other	Inbred	8095	10,28%	7126
OTV	Inbred	786	1,00%	776
PolyHighResolution	Inbred	3574	4,54%	3328
	Total	78722		38307

Table S2.4: Comparison of the reproducibility of InDels and SNP genotyping between 13 maize varieties replicated on 50K Illumina SNP and Affymetrix Axiom InDel arrays. Percentage of difference represented the proportion of genotype different between replicates from a same variety. Note that DNA samples from replicated varieties originated from different seed sources.

	% of difference in genotyping	
Variety	**50K SNP array**	**InDel array**
A554	4.4%	4.5%
A632	1.7%	2.1%
A654	1.6%	1.4%
B73	0.0%	5.2%
C103	0.2%	0.6%
CO255	1.5%	0.9%
D105	1.7%	0.7%
EP1	1.7%	2.0%
F2	1.6%	1.7%
F252	3.1%	0.9%
KUI3	6.0%	5.2%
Oh43	0.3%	2.6%
W117	1.6%	0.8%

Table S2.6: Newly sequenced data used for insertion/deletion discovery and whole genome sequence assemblies.

Inbred lines	Type	Distance R1-R2	Nb. reads	% used	scov	Nb pairs	pcov
F2	PE	-36 (+/- 27)	516,844,994	18.2	8.5	174,374,819	15.7
	PE	26 (+/- 38)	97,286,460	32.9	5.7	27,367,318	11.8
	PE	181 (+/- 55)	219,355,916	22.3	8.6	47,582,696	33.3
		Total PE	833,487,370	21	31.3	174,374,819	76.3
	MP	2153 (+/- 262)	565,822,128	17.2	16.4	18,850,205	108.4
C103	PE	-37 (+/- 15)	823,506,904	31.9	36.4	246,941,948	61.9
	MP	2898 (+/- 150)	652,864,236	9.6	8.6	11,167,486	69.7
PH207	PE	-26 (+/- 16)	754,615,040	25.5	27.9	217,832,024	59.1
	MP	3053 (+/- 177)	631,282,744	8.5	7.3	10,592,560	69

Chapter 3

High throughput genotyping of large insertions and deletions revealed genomic regions absent from the reference genome with important contribution to maize adaptation.

Clément Mabire[1], Delphine Madur[1], Valérie Combes[1], Sophie Bouchet[1,2], Jorge Duarte[3], Alain Charcosset[1], Stéphane Nicolas[1]

Authors Affiliation:

1 GQE – Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

2 Current adress : GDEC, INRA, Université Clermont Auvergne, 63000 Clermont-Ferrand, France

3 Biogemma - Centre de Recherche de Chappes, CS 90126, Chappes 63720, France

Corresponding authors: stephane.nicolas@inra.fr

Key words: InDels, association study, genetic diversity, signature of selection, SNP, trait variation.

Abstract

Large insertions and deletions (InDels) are pervasive in maize and could have strong functional and phenotypic effect by removing or modifying drastically genes. Moreover, the contribution of these InDels to adaptation and phenotypic variations in comparison to largely used SNPs are not well known.

We genotyped a 362 maize inbred lines panel representing a broad range of diversity, with 61,492 InDels from 37bp to 129Kbp as well as one million of SNPs. Relatedness between inbred lines and the genetic structuration estimated with SNPs and with InDels were similar. By calculating the linkage disequilibrium between markers, we found 51% of 21,360 anchored InDels not in high linkage disequilibrium (LD>0.8) with SNPs. We found a significant enrichment of InDels under selection compared to SNPs and 28 regions under selection were targeted only with InDels. We finally identified 13 QTLs by GWAS only with InDels. The highest number of InDels associated was found for flowering, a highly adaptive trait. Two InDels were associated with flowering close to Vgt3 on chromosome 3 were also under selection.

Our results strongly suggested that InDels were probably more suitable to identify regions under selection involved in maize adaptation. This study reinforces the idea that one reference genome is not sufficient to identify all genomic regions involved in adaptation and phenotypic variations.

Introduction

In the past years, there has been growing evidence that structural variations (SVs) are pervasive within plant genomes (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010; Saintenac et al., 2011; Cao et al., 2011; Xu et al., 2012; Lu et al., 2015; Pinosio et al., 2016; Montenegro et al., 2017; Varshney et al., 2017; Zhao et al., 2018). Among these structural variations, large insertions/deletions (InDels) are of particular interest since they conduct to the presence or absence of portions of the genome between individuals from the same species, including genes or regulatory regions of genes (Hirsch et al., 2014; Saxena et al., 2014). InDel sequences can be absent at one locus and present elsewhere in the genome or not leading to Presence / Absence Variations (PAVs).

By comparing BAC sequences in maize, Fu and Dooner (2002) and Brunner et al., (2005) first highlighted that gene and TE contents could greatly vary between inbred lines Mo17 and B73. Several studies found multiple copy number variations (CNVs) and PAVs between the reference genome and other maize inbred lines or teosinte by using comparative genomic hybridization technology (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010). More recently, the whole genome assembly of several inbred lines has been carried out, leading to the identification of 2,500 genes that were either present or absent between PH207 and B73 genomes (Hirsch et al., 2016), 10,735 PAVs between F2 and B73 including 417 novel genes in F2 (Darracq et al., 2018) as well as 3,408 insertions and 3,298 deletions regarding the reference genome and K11 and W22 inbred lines (Jiao et al., 2017). By investigating gene annotations, many studies observed within InDels an enrichment in genes involved in stress tolerance, suggesting that InDels might contribute to environmental adaptation of maize (Swanson-Wagner et al., 2010; Hirsch et al., 2016; Jiao et al., 2017; Darracq et al., 2018).

Since InDels could carry genes or host regulatory elements of genes, they can have a strong functional and therefore phenotypic impact (Hirsch et al., 2014; Saxena et al., 2014) but the effect of InDels on phenotypic variations remains poorly known in maize. In maize, Beló et al., (2010) identified in a large CNV region a previously discovered QTL for resistance to sugar-cane mosaic virus as well as a gene (*Rcg1*), conferring resistance to a fungus in a large insertion. Maron et al., (2013) showed that the MATE1 copy number was correlated with aluminium tolerance in maize. In maize, Chia et al., (2012), identified genes in different copy number by sequencing. Genome-wide association study (GWAS) with SNPs on 5 traits (leaf angle, length and width and resistance to southern and northern leaf blight) led to an overrepresentation of associated SNPs that are in high linkage disequilibrium (LD) with genes in different copy number suggesting an effect of these genes on trait variations. Using data from the Hapmap1 and Hapmap2 (Gore et al., 2009; Chia et al., 2012), Wallace et al., (2014) performed GWAS on 28,9 million SNPs and ~800k CNVs on 41 traits relative to plant architecture, development, and disease resistance. They identified 4,484 SNPs and 318 CNVs associated with at least one phenotype and found an enrichment of genic CNVs in GWAS hits in comparison to the whole dataset. More recently, by resequencing 14,129 inbred lines using Genotyping By Sequencing (GBS), Lu et al., (2015) selected 1,1 million mapped tags classified as InDels since they did not align to the reference or did not align within 10Mb of their genetic position. To test the role of these InDels in controlling phenotypic variations, they conducted GWAS on SNPs, in high LD or not with InDels. SNPs associated with InDels exhibited enriched associations with four traits (days to silking, days to anthesis, plant height, and ear height). In addition, other studies also showed that TEs could contribute to phenotypic variations in plant like rice, tobacco or maize by modifying genes or their regulatory regions, notably

in response to biotic and abiotic stress (Zerjal et al., 2012; Grandbastien, 2015; Makarevitch et al., 2015).

Genomic regions involved in adaptation can also be detected by identifying signatures of selection between different genetic groups adapted to contrasted environments (Nielsen, 2005). Many regions targeted by selection were found thanks to SNPs in maize from the 50K Illumina® (van Heerwaarden et al., 2012; Bouchet et al., 2013), 600K Affymetrix® Axiom® genotyping array (Unterseer et al., 2016), genotyping by sequencing (Takuno et al., 2015; Gouesnard et al., 2017) and sequencing (Brandenburg et al., 2017; Romero Navarro et al., 2017). Identification of signatures of selection allowed notably to identify genomic regions targeting genes involved in flowering or stress response (van Heerwaarden et al., 2012; Brandenburg et al., 2017). However, few studies have tested if large InDels were under selection. More CNVs segregated in the wild than cultivated barley accessions suggesting that CNVs may contribute to the molecular mechanisms for adaptation (Muñoz-Amatriaín et al., 2013). The resequencing of 292 pigeonpea accessions allowed to detect 68 CNVs and 1 PAV as putative targets of selection (Varshney et al., 2017). They also found 183 and 29 marker-trait associations mapped on 63 and 19 CNV or PAV across the genome of breeding lines and wild species, respectively, which suggests an effect of SV on trait variation.

The present study was designed to evaluate the interest of integrating PAV in population and quantitative genetic analyses. For this purpose, we genotyped an association panel representing a broad range of diversity in maize (tropical, Corn Belt Dent, and Stiff Stalk, Northern and European Flint inbred lines) with a new InDels Affymetrix® Axiom® high throughput genotyping array (Chapter 2). This panel was also genotyped with three SNPs genotyping technologies: the 50K Illumina® (Ganal et al., 2011), and the 600K Affymetrix® Axiom® (Unterseer et al., 2014) SNPs genotyping arrays as well as genotyping by sequencing (Elshire et al., 2011). We first examined the distribution of presence and absence frequencies in the panel and then compared the estimation of relatedness and structure between inbred lines with InDels and SNPs. In order to determine if putative effects of InDels can be captured by SNPs, we calculated LD between different types of markers and estimated the LD extent according to distance. We then detected signatures of selection by comparing allelic frequencies between flint, dent and tropical maize inbred lines. Finally, we explored the potential of InDels to discover new regions involved in genetic architecture of quantitative trait by performing GWAS on 23 traits relative to phenology, architecture and yield components.

Materials and methods

Plant material

A panel of 364 maize lines representing a broad range of diversity (Camus-Kulandaivelu, 2006; Ducrocq et al., 2008; Bouchet et al., 2013, 2017) was evaluated in 9 environments for 23 traits relative to phenology, architecture and yield component (Supplemental table S3.1). All inbred lines were not measured for all traits in each environment, thus we calculated the adjusted mean for each trait (Bouchet et al., 2013, 2017). We genotyped this panel with four different genotyping technologies (see below). DNA for genotyping was extracted from leaves following a NaBisulfite method modified from Tai and Tanksley (1990) and Dellaporta et al. (1983).

We performed a conformity check to identify putative illegitimate DNAs genotyped with different technologies. We compared relatedness value between inbred lines estimated by these different genotyping technologies. We identified two illegitimate DNA samples that were removed from analysis leading to a final number of 362 inbred lines.

Molecular markers

We genotyped this panel of inbred lines with 49,036, 616,201 and 955,691 markers originated from the 50K SNPs Illumina® genotyping array (Ganal et al., 2011), 600K SNPs Affymetrix® Axiom® genotyping array (Unterseer et al., 2014) and Genotyping By Sequencing (GBS) (Elshire et al., 2011), respectively. Among these markers, we selected 45,821 and 569,622 polymorphic SNPs from 50K array and GBS respectively. We selected only markers from the 600K array classified PolyHighResolution and OffTargetVariants according to AxiomGT1 software leading to 586,071 polymorphic markers. GBS genotyping was already imputed with TASSEL software (Glaubitz et al., 2014). We also genotyped the Mite vgt1 (Salvi et al., 2002, 2007; Ducrocq et al., 2008; Castelletti et al., 2014) and the InDel Dwarf8 (Thornsberry et al., 2001; Camus-Kulandaivelu et al., 2006) using Kaspar technology (KBioscience).

Among 586,071 markers from the 600K Affymetrix® Axiom® array we genotyped 514,375 SNPs, 6,453 small InDels (from 1 to 4bp) called InDel-600K and 65,243 Off Target Variants (OTVs), so-called OTV-600K in analysis. These OTV-600K indicated that probes of these markers did not hybridize with DNA of some individuals. This absence of hybridization could originate either from polymorphisms within probe sequences or its absence. These markers could detect the presence and absence of the underlying sequences carrying probes and could therefore indicate the presence of an InDel (Unterseer et al., 2014). We treated this subset of markers as bi-allelic “presence” and “absence” in following analyses in order to evaluate the complementarity between the 65,243 OTV-600K and the InDels Affymetrix® Axiom® genotyping array. Note that the number of individuals whose the probe sequences were putatively absent was very low for 39,040 OTV-600K markers (<3%).

Based on their physical position on the B73 v2 reference genome, we identified duplicated markers between different genotyping technologies for 71,745 loci. We selected for these loci the marker with the lowest missing data and heterozygosity rates, providing that the similarity between duplicated markers was higher than 95%. When similarity was lower than 95% between markers at the same locus, we removed both markers. We finally removed monomorphic markers leading to a final set of 1,076,794 polymorphic markers: 34,964, 483,657 and 486,475 SNPs from the 50K, 600K arrays

and GBS respectively as well as 65,243 OTV-600K, 6,453 InDel-600K, the markers for MITE vgt1 and Dwarf8 InDels.

We also genotyped this panel with 105,927 InDels using an InDel Affymetrix® Axiom® genotyping array that was developed to genotype specifically presence and absence of InDel sequences (Chapter 2). Briefly, these InDels were discovered by aligning reads from three genomes (F2, PH207 and C103) on the B73 reference genome to find deletions relative to the reference genome (present in B73, absent in at least one other line). We also applied the inverse strategy that consisted of aligning reads of B73 on these three different draft genomes to find insertions regarding the reference genome (absent in B73, present in at least another line). 60% of InDels discovered as insertions regarding the B73 reference genome were not anchored on the reference genome and thus no physical position was available for these InDels (Chapter 2). Physical positions used for the anchored InDels corresponded to the B73 v2 reference.

To build the array, 662,772 probes were designed to genotype 105,927 InDels previously discovered by sequencing approach. Two types of probes were developed: (i) “external” probes which target breakpoints by designing probes on the two flanking and external sequences of a given InDel (BP probes) and (ii) “internal” probes which target the internal InDel sequences by designing probes on polymorphic (OTV probes) or monomorphic sites (MONO probes) (Chapter 2). BP probes revealed presence/absence by analyzing the variation of fluorescence contrast between individuals according to the allele at breakpoint site (Figure S3.1 A) whereas OTV and MONO probes revealed presence/absence by analyzing variation of fluorescence intensity between individuals (Figure S3.1 B and C) (Chapter 2). In addition to presence and absence, OTV probes genotyped the SNP within the sequence, allowing to test the effect of four different genotypes (absence, AA, BB and AB, Figure S3.1 B) in genetic analysis.

The genotyping of these probes was called as “present” (2) vs. “absent” (0), corresponding to the presence or absence of the sequence for the inbred lines genotyped.

479,027 probes (46,263 BPs, 96,867 OTVs, and 335,778 MONOs) passed Affymetrix® quality control and could be used to genotype 89,393 InDels (Chapter 2). Among these probes, we first selected polymorphic probes for presence and absence for which genotyping results were consistent at more than 75% between array genotyping and resequencing for the four inbred lines used to discover InDels (see Chapter 2 for more explanation). Among 46,382 BP probes passing Affymetrix® Quality control, we removed 26,012 BP probes displaying an unexpected OTV cluster due to the absence of probes hybridization. In the end, we kept only 18,190 bi-allelic BP probes which genotyped 15,393 InDels (2,797 InDels were genotyped with both forward and reverse BP probes). We also selected 345,057 OTV and MONO probes which genotyped 60,077 InDels. Note that 50,524 out 60,077 InDels were genotyped with at least two probes.

To combine the genotyping of 50,524 InDels genotyped by several OTV and/or MONO probes, we first removed probes that displayed too divergent genotyping for presence/absence with other probes from same InDel based on their average LD with other probes. We removed probes with an average LD with other probes below 0.1 in a first step and below 0.5 in a second step. InDels genotyped with two divergent probes were removed from the analysis. In this way, we selected 231,077 probes which allowed us to genotype 49,806 InDels. Second, we calculated for each inbred lines the average frequency of presence across the different probes. InDel was genotyped either present or absent or as

missing data if the average frequency of presence over all probes within InDels for one individual is superior, inferior or equal to 0.5, respectively. Finally, we obtained presence/absence genotyping for the 362 inbred lines for 49,731 InDels genotyped with OTV and MONO probes.

In addition, 2,797 out 15,393 InDels were genotyped by two BP probes (a forward and a reverse probes). We observed divergence (dissimilarity >5%) between forward and reverse BP probes genotyping for 807 out 2,797 InDels. It was probably due to complex rearrangements within InDels (Chapter 2). Therefore, we kept the forward and reverse probes for these 2,797 InDels. 3,632 InDels were also genotyped with both BP and internal probes (OTV and/or MONO). For these 3,632 InDels, we kept the genotyping from both (i) BP probes and (ii) OTV and MONO probes because they reveal presence/absence in a different way. Finally, we genotyped presence and absence for 61,492 InDels thanks to 49,731 markers from OTV and MONO probes and 18,190 BP probes.

In the end, the whole panel was genotyped with 1,076,794 markers from three SNP genotyping technologies and with 67,921 InDel markers from the new InDels Affymetrix® Axiom® array. This total set of markers displayed a missing data rate of 2.4%, 2.2%, 24.4% and 1% for 50K, 600K, GBS and InDels array respectively. The heterozygosity rate was 0.45%, 0.24% for 50K, and 600K respectively. Note that for the GBS, heterozygous genotypes were transformed into missing data due to high error rate for heterozygous genotype imputed by TASSEL (Gouesnard et al., 2017; Negro et al., 2018). For the InDels genotyping array, the hemizygous rate was available with BP probes but not with OTV and MONO (Chapter 2). The 18,190 BP probes used displayed 2% of hemizygous genotypes.

Diversity analysis

We calculated the Minor Allelic Frequency (MAF) for the 1,076,794 markers and 67,921 InDel markers as well as the frequency of absence for InDels. For all other analyses, we selected markers based on quality of genotyping: MAF>3%, missing data<20% (<80% for GBS markers) and heterozygosity rate<15%. We selected a total of 873,011 bi-allelic and non-duplicated markers from 50K, 600K and GBS (33,911 SNPs 50K, 472,361 SNPs 600K, 26,196 OTV-600K, 6,325 InDel-600K, and 334,216 SNPs GBS) as well as 63,803 bi-allelic InDel markers targeting 59,927 unique InDels from the InDels Affymetrix® Axiom® array. Since we still had 11.4% of missing genotyping data for SNPs technologies (2.22%, 2.18% and 26.2% for 50K, 600K and GBS respectively), we imputed missing genotyping data with Beagle v3.3.2 (Browning and Browning, 2009, 2007) for the three technologies together. We also imputed missing data for InDels (1%) using Beagle v3.3.2.

We used GenaBEL package (Aulchenko et al., 2007) in R software to calculate identity by state (IBS) and identity by descent (IBD). To estimate population structure, we used ADMIXTURE software (Alexander et al., 2009) for K=3 and K=5. We choose these numbers of genetic groups according to Bouchet et al., (2013) and Camus-Kulandaivelu et al., (2006) which previously studied the genetic structuration of the panel.

Linkage disequilibrium analysis

Linkage disequilibrium (LD) was calculated as the squared correlation between allele doses at two loci for 25,544 InDel markers and 872,324 markers from SNPs genotyping technologies anchored on the reference genome. Note that 687 SNPs from GBS and 38,259 InDel markers were not anchored. Since we used a panel with related inbred lines, we calculated the correlation corrected by relatedness between inbred lines (r2k) (Mangin et al., 2012). To calculate the r2K, we used the identity-by-descent

matrix estimated with 28,530 SNPs panzea. In order to spare computational time, LD was calculated for each chromosome separately using a sliding window of 2 Mbp centered on each marker with imputed genotyping. To avoid a possible bias due to the difference of lines used for SNPs and InDels discovery, we filtered out SNPs which were monomorphic among the four inbred lines used to discover InDels. LD extent was estimated using Hill and Weir model (Hill and Weir, 1988) for both genetic and physical distances. The genetic position of all markers was estimated by projecting the physical position of markers on a consensus genetic map obtained from two connected European NAM obtained by crossing dent and flint lines genotyped with 50K arrays (Giraud et al., 2014). The LD extent for each chromosome was obtained at the abscissa of the intersection between the LD curve and the value of LD = 0.1. We identified clusters of InDels on chromosomes by grouping adjacent InDels with a LD value >0.7.

Mapping InDels with linkage disequilibrium

In order to map 36,568 non-anchored InDel markers, we developed an algorithm based on LD (r2) between 38,259 not anchored InDel markers (candidates) and the anchored markers (reference markers) corresponding to 872,324 markers from SNPs genotyping technologies and 25,544 InDel markers positioned on the B73 genome. We assigned the candidate to the position of the highest correlated reference marker providing that at least one marker has a LD value above 0.5 and that there was a single peak on the genome. We chose this LD threshold after comparing effect of different thresholds from 0 to 0.8 on the number of InDels correctly and erroneously positioned using a first set of 100 anchored InDels that we remapped using LD with SNPs (Table S3.2). We set the LD threshold at 0.5 because it maximized the number of mapped InDels while limiting the number of InDels badly mapped (<4%) (Table S3.2). To remove InDels anchored at several positions, we filtered out non-anchored InDels displaying several peaks with similar LD values (less than 0.1 of difference) on different chromosome. We validated our approach using this LD threshold on 10 other random sets of 100 InDels already positioned and compared observed and real positions. To identify regions with the highest increase in mapped InDels, we calculated the proportion of InDels located on contiguous bins of 20Mbp in each chromosome.

Detection of signatures of selection

Based on Admixture results obtained with 28,530 panzea SNPs, we considered the structure for Q=3 that represent dent, flint and tropical genetic groups to identify the signatures of selection between different groups. We selected 207 inbred lines assigned to a genetic group with a proportion higher than 0.7 (99 dent, 70 flint and 38 tropical inbred lines). We used BayeScan (Foll and Gaggiotti, 2008) to identify loci under selection according to differences in allele frequencies between 3 genetic groups. We selected 63,803 InDel markers and 872,890 markers from SNPs genotyping technologies polymorphic in 207 inbred lines. We considered markers under selection if the logarithm of the posterior odd (PO) was above 1. To analyze allelic variations within groups, we calculated the allelic frequency of each marker putatively under selection within three groups.

The differentiation (Fst) between 3 genetic groups was calculated as the ratio of inter-group diversity over the total genetic diversity according to Nei (1973). The Fst was also estimated with Bayescan (Fstb). To identify genomic regions under selection, we grouped adjacent markers under selection with LD corrected by relatedness (r2K) higher than 0.7. To display Fst variation along chromosomes, we integrated non anchored InDels at the physical position predicted by our linkage disequilibrium approach.

GWAS

We performed GWAS using genotyping data filtered according to MAF (>3%), missing data (<20%, and <80% for GBS markers) and heterozygosity rate (<15%), after imputing missing data with beagle (Browning and Browning, 2009, 2007). We used adjusted means of 23 traits relative to phenology, architecture and yield component (Supplemental table S3.1) that were estimated and analyzed in previous studies (Bouchet et al., 2013, 2017).

We tested the association between 63,803 InDel markers from the Affymetrix® array and 873,011 polymorphisms from the SNP genotyping arrays and GBS (840,490 SNPs, 26,196 OTV-600K, 6,325 InDel-600K) with 23 traits using mixed model described in Yu et al., (2006):

$$Y = \mu + X\beta + Zu + \varepsilon$$

With Y the vector of phenotype, μ the intercept, X the vector of individual marker genotypes, β the marker fixed effect, Z the incidence matrix for lines, *u* the genetic background effect, $u \sim \mathcal{N}(0, K\sigma_g^2)$, where K is a matrix of similarity between lines, and ε the vector of residuals, $\varepsilon \sim \mathcal{N}(0, I\sigma_e^2)$. We used a kinship matrix estimated with 28,530 panzea SNPs belonging to all chromosomes except the chromosome of the marker tested as fixed effect (Rincent et al., 2014). Single marker associations were run with FaST-LMM (Lippert et al., 2011). Associations with false discovery rate (FDR) < 10% (Benjamini and Hochberg, 1995) were considered significant. FDR was calculated with the R function "p.adjust" after filtering out 218,745 duplicated markers in complete LD (r2=1). We grouped significant markers into a same QTL that were adjacent on the physical map and with LD value (r2K) > 0.7. To display Manhattan plots and group markers into QTLs, we used the physical position predicted by our linkage disequilibrium approach for non-anchored InDels. We visualized haplotypes, phenotype variations and assignation of lines to genetic groups, type and origin of markers around association peaks using the heatmap.3 R function.

To evaluate the effect of SNPs within InDel sequences in addition to presence and absence of the sequence, we performed GWAS on 50,605 SNPs within InDels genotyped with the InDels genotyping array (OTV probes). These markers could display four genotypes: AA, AB, BB and absence (Figure S3.1 B). Single marker associations were run with MM4LMM (Laporte et al., 2019). To test only the effect of allele A vs allele B vs absence, we selected 32,737 out 50,605 OTVs displayed only three genotypes: AA, BB and absence. We compared the p-value of the GWAS test obtained for the tri-allelic marker (absence, AA, BB) in comparison to testing effect of presence / absence of sequence for the corresponding InDels.

Results

Genotyping

We genotyped 61,492 InDels from 37bp to 129.7Kbp with 67,921 markers in 362 inbred lines. These InDels were discovered as deleted or inserted regarding the reference genome B73 (insertions and deletions respectively) and three maize inbred lines (F2, C103 and PH207). The majority of InDels had their sequence absent in B73 and present for F2 (Table S3.3). 5,797 InDels had specific sequences that were totally absent from the rest of the genome. In contrary, 9,706 InDels had non-specific sequences, meaning that these sequences were either present or absent at one locus but were present in totality elsewhere in the genome. Other InDels had a variable proportion of specific sequences ranging from 35 to 21,540bp with a median of 197bp.

We also genotyped this panel with 1,076,794 bi-allelic markers that we assembled by combining markers (see methods) from 50K Illumina® (Ganal et al., 2011) and 600K Affymetrix® Axiom® genotyping array (Unterseer et al., 2014) and Genotyping By Sequencing (GBS) (Elshire et al., 2011). Note that 600K allowed to genotype 483,657 SNPs, 65,243 OTVs (OTV-600K) and 6,453 small InDels from 1 to 4bp (InDel-600K).

To evaluate the complementarity of SNP genotyping technologies and the InDels genotyping array, we compared the density of SNPs on the whole genome with the density of SNPs within InDel sequences. Among 14,761 deletions relative to B73 (*e.g* whose sequences were present in B73 and absent from other lines), we observed a lower SNP density within InDel sequences (1 SNP every 3,669bp) than in the whole genome (1 SNP every 1,913 bp). Out of these 14,761 deletions, 3,476 presented SNPs, InDel-600K and/or OTVS-600K (11,933 SNPs, 47 InDel-600K and 537 OTV-600K). Among these markers, 74% were originated from the GBS although they represented only 51% of all markers, indicating that SNP designing arrays probably lead to a higher counter selection of such regions than GBS. Accordingly, we found an enrichment of missing data for SNPs within InDels (28% of missing data) in comparison to others (0.12% of missing data) (pvalue<2.2x10-16).

Comparing InDels and SNPs, for allele frequency spectra and estimations of relatedness and genetic structuration of the panel

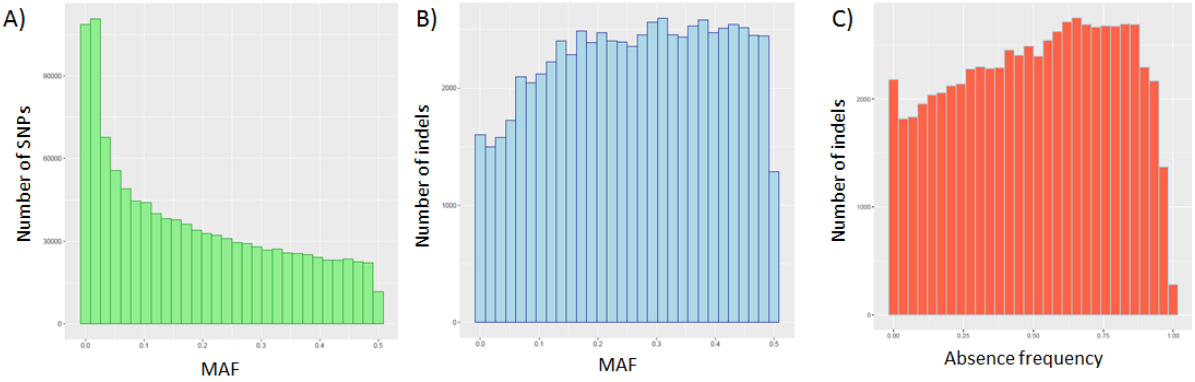


Figure 3.1: Distribution of (A) Minor Allele Frequency of 1,076,794 markers and (B) 67,921 InDel markers and (C) frequency of absence for 67,921 InDel markers in the 362 inbred line panel.

We compared Minor Allelic Frequency (MAF) distribution between 1,076,794 markers from SNPs genotyping technologies and 67,921 InDel markers (Figure 3.1). We found a deficit of rare alleles for InDels in comparison to the other class of frequencies while we found an excess of rare alleles for SNPs (Figure 3.1): 8% of InDels *vs* 25% of SNPs and 2% of InDels *vs* 9% of SNPs displayed a MAF below 0.05 and 0.01, respectively. Note that this difference of MAF distribution between SNPs and InDels came mainly from GBS since 83% of SNPs with MAF below 0.01 were originated from GBS. SNPs from 600K and 50K array displayed a MAF distribution similar to InDels.

We observed a strong asymmetry between absent and present alleles for InDel markers close to the fixation (<0.05 and >0.95). We observed indeed a strong deficit of InDels with an allele absent close to the fixation (allelic frequency of absent > 0.95%). As a consequence, the InDels with "presence" allele quasi-fixed was the most frequent representing 96% of InDels with MAF below 1%.

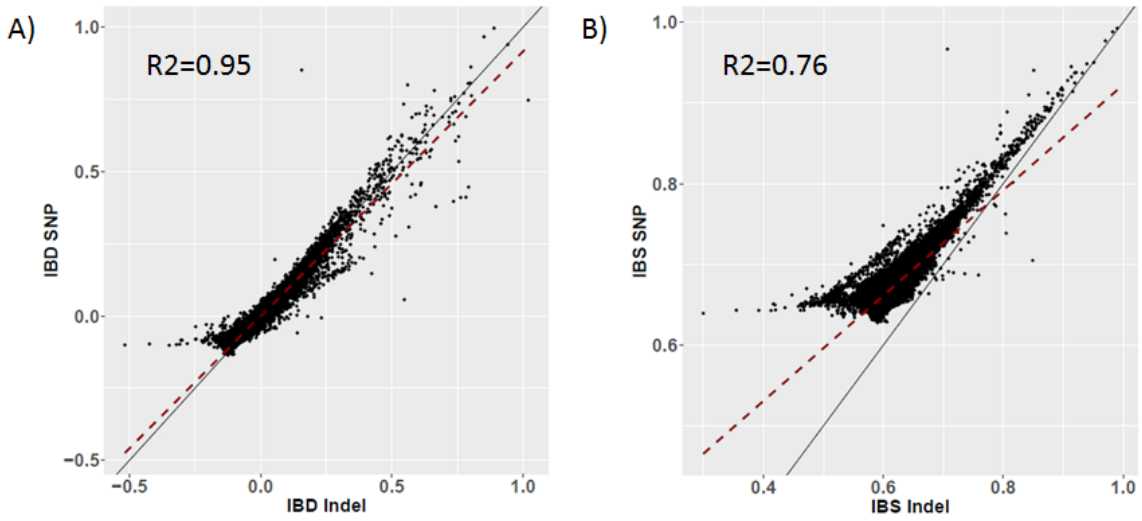


Figure 3.2: Comparison of kinship values for (A) IBD and (B) IBS matrix estimated with 873,011 markers from SNPs genotyping technologies and 63,803 InDel markers. The dotted and solid lines represent the linear regression and the bisector, respectively.

To evaluate the contribution of InDels, we compared the kinship matrices estimated with InDels and SNPs. Kinships were linearly correlated (r2=0.95) for Identity By Descent (IBD) (Figure 3.2 A) but a lower correlation (r2=0.76) was observed for kinship estimated using Identity By State (IBS) (Figure 3.2 B). For both IBD and IBS, we observed a few pairs of lines for which the kinship values estimated with InDels were clearly lower than those estimated with SNPs. These values correspond to pairs of lines including F2, B73 and inbred lines closely related to B73 and F2. It suggested a slight ascertainment bias since the majority of InDels on array were discovered between B73 and F2 inbred lines.

We estimated population structure using Admixture software for K=3 (Dent, Flint and Tropical genetic groups) and K=5 (Stiff Stalk, Corn Belt Dent, European Flint, Tropical and Northern Flint genetic groups) using 63,803 InDel markers and 840,490 SNPs, as well as 26,196 OTV-600K and 6,325 InDel-600K. 97.7% and 95.3% of inbred lines were assigned to the same group with InDels and other markers for K=3 and K=5, respectively. Assignation to genetic groups was strictly identical for inbred lines strongly assigned to a genetic group (>0.8) for K=3 and K=5.

Linkage disequilibrium

Chromosome	InDel-InDel	InDel-SNP	SNP-SNP
1	36,823	39,410	75,998
2	34,526	37,129	81,957
3	45,629	53,718	113,151
4	48,983	78,263	153,665
5	51,609	48,608	71,715
6	61,433	42,868	61,415
7	55,069	50,500	88,718
8	27,365	42,478	117,214
9	39,820	40,862	80,232
10	30,394	40,767	86,188
Whole genome	42,818	47,560	93,024

Table 3.1: Estimation of physical LD extent (bp) for a r2k = 0.1 using Hill and Weir model with r2k measure between InDels (InDel-InDel), InDels and SNPs (InDel-SNP) and between SNPs (SNP-SNP) 1. For the whole genome, we calculated average values weighted by the number of markers on each chromosome.

In order to estimate to what extent the polymorphism at InDels is captured by SNPs, InDel-600K or OTV-600K, we estimated the LD extent for InDels and other markers using Hill and Weir (1988) model according to the physical and genetic distance in a 2 Mbp window. LD extent according to physical and genetic distance was longer between SNPs (SNP-SNP) than between InDels (InDel-InDel) and between InDels and SNPs (InDel-SNP) (table 3.1 and S3.4) except for chromosome 6. For chromosome 6, LD extent was indeed similar between InDels and between SNP for physical distance and higher between InDels than between SNPs for genetic distance. This could be explained by the presence of two clusters of InDels on this chromosome located from 12.8 to 15.6 Mbp (Cluster 1) and from 134.1 to 135.3 Mbp (Cluster 2). These two clusters consisted in (i) 34 deletions from 113 to 11,640 bp totalling 131 kbp including 77.9 kbp of specific sequences (ii) 26 InDels from 121bp to 27.9kbp totaling 87,2kbp including 10,1 kbp of specific sequences.

To estimate the number of InDels whose genetic effect could be putatively captured by SNPs, we calculated LD measures between all anchored SNPs and InDels. Among 21,360 InDels anchored along the maize genome, 10,222 (49%) were in high LD (r2K>0.8) with at least one marker from SNPs genotyping technologies distant from less than 2Mbp. These markers were located at a median distance of 2,411 bp from InDels. We found a slight enrichment of non-specific InDels among those in high LD with at least one marker (39%) in comparison to InDels not in high LD with any marker (30%). For the 11,138 InDels not in high LD (r2k< 0.8), we investigated the presence of SNPs in the vicinity to understand whether these InDels were in regions (i) without genotyped SNPs or (ii) without any SNP in high LD with these InDels. We found a SNP at a higher median distance for InDels not in high LD with any SNP (332bp) in comparison to InDels in high LD with at least one SNP (214bp). We found a significant higher SNP density (p-value>2.2x10-16) in a 2 Mbp window around InDels not in high LD with any SNP (1 SNP/936bp) in comparison to InDels in high LD with at least one SNP (1 SNP/1,110 bp).

InDels mapping using linkage disequilibrium

In order to map not anchored InDels along the maize genome, we developed an algorithm based on LD between not anchored InDels and markers with known position. We calculated the LD (r2) between 897,868 anchored markers (InDels, SNPs, InDel-600K and OTV-600K) and 36,568 not anchored InDels (see Figure S3.2 for an example). After testing different LD thresholds on a set of 100 InDels with known position, we chose a threshold of LD value equal to 0.5 which maximized the number of mapped InDels and minimized the number of false positives (Table S3.2). We then further tested our approach with the chosen threshold by sampling 10 times 100 anchored InDels as candidates. We mapped 74% of InDels at a median difference of 2,985 bp between their predicted and true positions. Using this approach, we mapped 26,738 non anchored InDels (72% of them). Among these InDels, 13% were mapped thanks to anchored InDels, which represent only 3% of the whole set of markers anchored used (p-value<2.2x10-6).

We investigated the density of InDels along the genome before and after adding InDels mapped based on LD through bins of 20Mb along each chromosome. We found a significant enrichment of InDels mapped in peri-centromeric regions for all chromosomes except chromosome 10 (Figure S3.3, Table S3.5).

Identification of loci under selection

Using 207 inbred lines assigned to a genetic group with a membership > 0.7 for K=3 (99 dent, 70 Flint and 38 tropical), we identified 127 InDels under selection among which 30 were under decisive selection (log10(PO) > 2) (Table S3.6). All alpha values were positive, indicating a diversifying selection. Fst (Nei) and Fstb (Bayescan) for these 127 InDels ranged from 0.31 to 0.82 and from 0.33 to 0.42, respectively. Figure S3.4 illustrated the distribution of Fst (Nei) along the ten chromosomes as well as the distribution of InDels under selection. We calculated allelic frequencies of presence for these 127 putative InDels under selection within three genetic groups (Dent, Flint, and Tropical). Among them, 62 InDels were monomorphic in one group (38, 23 and 1 InDels were monomorphic in tropical, flint and dent groups respectively) whereas they segregated in other groups. Note that the MAF distributions for InDels were different between three genetic groups with a deficit of rare alleles for dent group whereas they were quite similar for SNPs (Figure S3.5). We found a slight enrichment of specific InDels (PAVs) among those under selection relative to their proportion in the entire InDel dataset (12% vs 10%).

We identified also 1,292 SNPs, 11 InDel-600K and 1 OTV-600K under selection among which 96 were under decisive selection (log10(PO) > 2) including the Mite Vgt1 (Table S3.7). Fst (Nei) calculated for these 1,305 markers were ranged between 0.31 and 0.9 and from 0.33 to 0.51 when estimated with BayeScan (Fstb). Among these markers, 483 were monomorphic in one or two groups (308, 170 and 23 monomorphic markers within the tropical, flint and dent genetic groups respectively) whereas they segregated in at least one other group. 17 SNPs were fixed in both tropical and flint groups and 1 SNP was fixed in both tropical and dent groups. We found a significant enrichment of InDels under selection in comparison to SNPs (0.22% vs 0.15%, p-value = 0.00241).

To investigate the interest of using InDels in addition to SNPs to identify signatures of selection, we grouped InDels and SNPs under selection according to the LD between adjacent markers. We identified 188 genomic regions under selection. We found 28, 132 and 28 regions under selection thanks to InDels, SNPs and both InDels and SNPs respectively. Among 28 regions identified with InDels

and SNPs, the maximum of log10(PO) was found for 13 and 15 of them with a SNP and an InDel respectively. InDels detected 30% of genomic regions under selection although they represented only 7% of markers. The enrichment of genomic regions under selection detected by InDels as compared to SNPs was strongly amplified when considering the ratio of the number of regions detected per number of markers (0.00085 vs 0.00019, pvalue<2.2*10-16).

We identified notably a region at 13Mbp in the chromosome 2 composed of 10 SNPs and one InDel close (7kbp) to a gene producing a protein (ZmAsr2) previously identified in a region associated with tolerance to drought (Virlouvet et al, 2011). The sequence of this InDel is totally absent of tropical inbred lines analyzed and present in 10%, and 65% of flint and dent inbred lines respectively.

GWAS

Table 3.2: All significant associations with InDels. Chr Chromosome, Position of the InDels, Log – log10(p-value), SV_Type InDel type regarding the reference genome, LENGTH the InDel size, FRAC_SPE proportion of InDel sequence present elsewhere in the genome, FGS_DIST distance to the closest gene.

Trait	Chr.	Position	Log	MAF	SV_TYPE	LENGTH	FRAC_SPE	FGS_dist	Closest gene
KW	1	254,558,578	5,66	0,34	DEL	321	NA	1,425	GRMZM2G045678
FFLW8	3	158,903,190	5,71	0,34	INS	400	0	821	GRMZM2G171600
FFLW8	3*	158,903,190*	5,59	0,34	INS	127	1	NA	
MFLW8	3	159,556,505	6,78	0,31	INS	191	0	6,349	AC188753.3_FG005
FFLW8	3	159,556,505	6,21	0,31	INS	191	0	6,349	AC188753.3_FG005
LFNB	3	159,556,505	5,66	0,31	INS	191	0	6,349	AC188753.3_FG005
LFNBb	3	159,556,505	5,66	0,31	INS	191	0	6,349	AC188753.3_FG005
SL	4*	26,490,017*	6,00	0,41	INS	609	0.27	NA	
SL	4*	26,490,081*	5,80	0,36	INS	309	0.16	NA	
BRNB	4	36,042,865	5,93	0,41	DEL	831	0	534	GRMZM2G131378
BRNB	4	36,045,322	6,12	0,49	INS	142	0	0	GRMZM2G131329
HL	4*	170,796,605*	7,00	0,03	INS	391	1	NA	
SL	7	159,137,426	6,78	0,36	DEL	274	0.24	144	AC197013.3_FG008
LFNB	8	123,512,295	5,89	0,24	INS	118	0	0	GRMZM2G479987
LFNB	8	123,512,412	6,08	0,24	INS	118	0	0	GRMZM2G479987
LFNBb	8	123,512,412	5,62	0,24	INS	118	0	0	GRMZM2G479987
MFLW8	8	131,213,501	6,13	0,24	DEL	4464	0.02	0	GRMZM2G064758
FFLW8	8	131,213,501	6,20	0,24	DEL	4464	0.02	0	GRMZM2G064758
HL	NA	NA	5,84	0,15	INS	1479	0.14	NA	
HL	NA	NA	5,90	0,31	INS	607	0.13	NA	
TKW	NA	NA	5,69	0,06	INS	2173	0.20	NA	
TL	NA	NA	6,22	0,35	INS	478	0.88	NA	
TaAk	NA	NA	5,67	0,49	INS	6010	0.83	NA	

In order to analyze the effect of InDels on phenotypic variations, we performed GWAS for 63,803 InDel markers and 840,490 SNPs, as well as 26,196 OTV-600K and 6,325 InDel-600K, with MAF>3% and 23 agronomic traits relative to phenology, yield component and architecture. We found 23 significant associations (Table 3.2) between phenotypic variations and InDels (0.03% of InDel markers) with a FDR 10%. In comparison, we performed also GWAS on 873,011 markers from SNPs genotyping technologies and found 546 significant associations with trait variation, 515 with SNPs, 28 with OTV-600K, 2 with InDel-600K and the Mite Vgt1 (0.05% of the 873,011 markers) (Table S3.8).

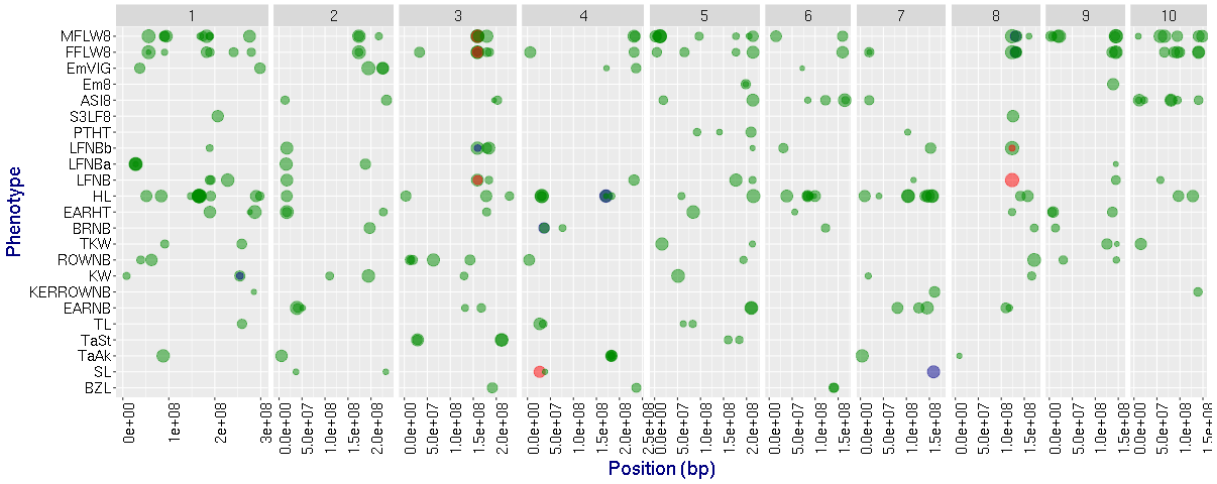


Figure 3.3: Projection of QTLs for 23 trait analyzed in GWAS (see table S1 for trait description). QTLs were obtained by grouping contiguous associated markers r2k (linkage disequilibrium) superior to 0.7 Green, blue and red circles represent QTLs composed uniquely of SNPs, InDels, and both SNPs and InDels, respectively. Size of circles correspond to strength of association.

Grouping adjacent associated markers from a same trait with LD (r2k) upper to 0.7 led to 294 Quantitative Trait Loci (QTLs) (Figure 3.3). Among them, 6 QTLs were identified by both SNPs and InDels whereas 255 QTLs and 8 QTLs were composed uniquely of SNPs and InDels respectively. In addition, 22 and 2 other QTLs were identified thanks to OTV-600K and InDel-600K respectively as well as one QTL thanks to the Mite Vgt1 (-log(p-value)=6.36). We found a slight enrichment of OTV-600K in comparison to all significant SNPs (p-value =0.0288). Note that 5 InDels associated were not taken into account in QTLs grouping since we did not map them and they were not in LD (r2k>0.7) with any associated and anchored markers. Thus these 5 InDels probably identified 5 other QTLs. For the 6 QTLs identified by both SNP and InDels, InDel showed stronger associations for 4 of them. We investigated the presence of SNPs in the vicinity of the 8 QTLs identified only with InDels and found a SNP at less than 1Kbp from all these QTLs.

We found no significant difference for QTLs detected by InDel markers and by SNPs relative to the respective number of markers (ratios of 0.000298 vs 0.000311, p-value=0.9525).

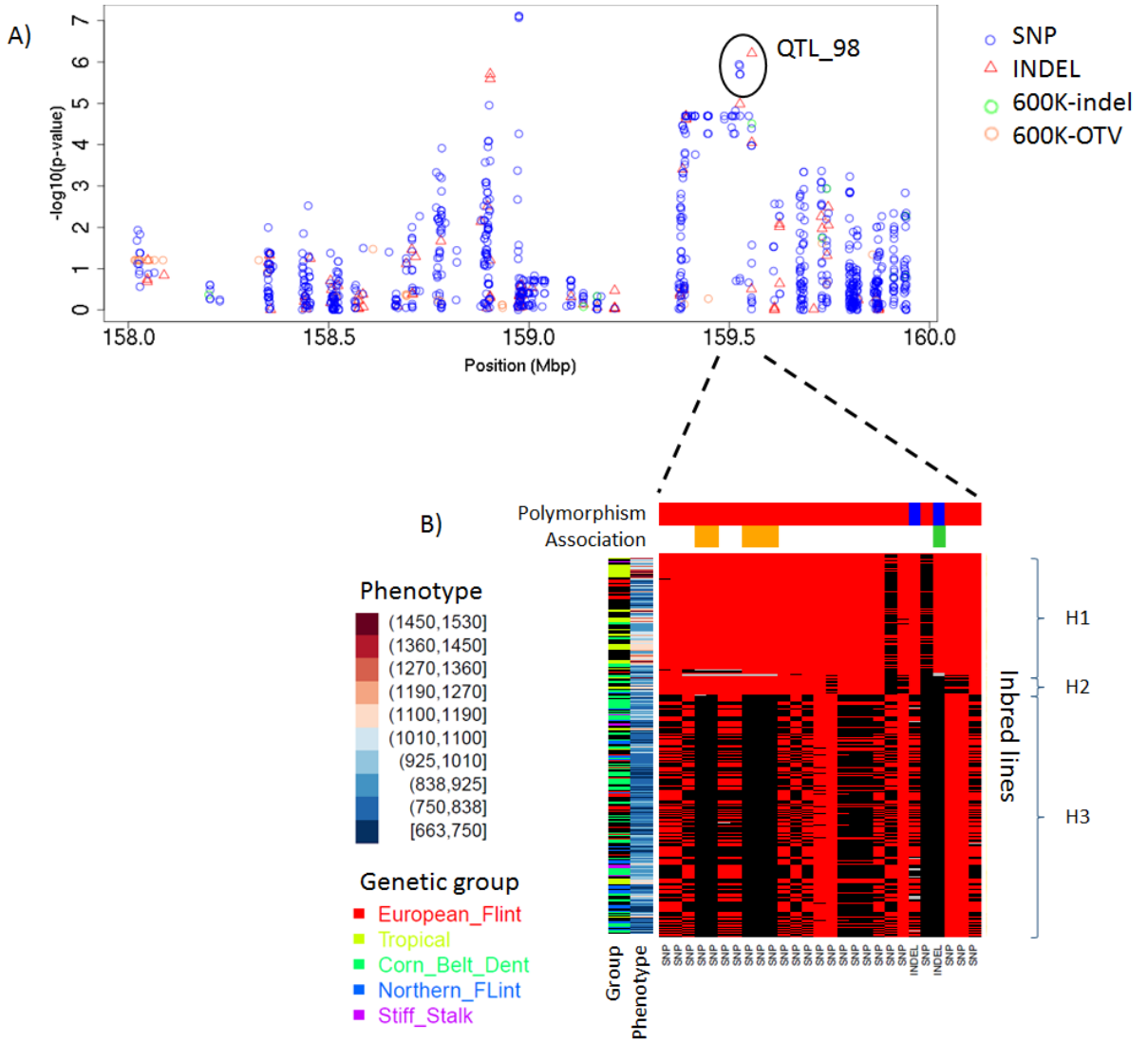


Figure 3.4: Complementarity of InDels and SNPs for identifying 4 main QTLs involved in variation female flowering time in a 2 Mbp region of chromosome 3. (A) Manhattan plot of the –log10(p-value) according to physical position Markers were colored according to their types :SNPs in blue, InDel-600K in green, OTV-600K in orange and InDels in red (B) Local haplotype of markers in a genomic region of 80kbp centered on QTL_98 associated with FFLW8. Lines (in row) were ordered according to haplotypes given by 6 associated markers (1 InDel and 5 SNPs). Above the haplotypes were displayed the marker types (Polymorphism) with SNPs in red and InDels in blue and the markers associated (Association) with SNPs associated in orange, InDels associated in green. Right to the haplotypes were displayed the genetic group of inbred lines (Group) and the flowering time variation in growing degree day for the inbred lines (Phenotype). For phenotype, inbred lines were colored using a colored gradient according to their earliness with early lines colored in blue to late lines colored in brown. For genetic group, inbred lines were colored according to their group assignation with black representing the admixed inbred lines (assignation probability < 0.7). Genotypes were colored according to B73 genotypes for SNPs (black = similar to B73 genotype) and allelic state for InDels (absent in black and present in red)

To illustrate the interest of using InDels genotyping in GWAS analysis, we focused on one QTL in a genomic region in the vicinity of *vgt3* (Salvi et al., 2007) (1Mbp around the highly associated SNP) (Figure 3.4 A). This QTL is composed of 5 SNPs and 1 InDel which displayed the strongest association (Figure 3.4). We identified 3 major haplotypes in this QTL based on genotyping of associated SNPs and InDel (H1, H2 and H3 in Figure 3.4 B). SNPs differentiated the H1 and H2 haplotypes from H3 whereas the InDel differentiated the H1 from the H2 and H3 haplotypes. Since 18 lines with H2 haplotype were earlier than lines with H1 haplotype, it conducted to a stronger association with flowering time for the InDel than for the SNPs.

Our array provided also information of polymorphism for SNPs within InDels. In order to understand the effect of SNP within InDel on trait variation, we tested the effect of three alleles: the absence of the sequence and two alleles discriminated by the SNP in case of presence. By taking into account SNP within InDels in GWAS, we found 39 loci associated with phenotypic variations (–log10(p-value)>5). These loci were located within 11 InDels which were not found by testing the effect of presence and absence in GWAS analysis. An interesting example was an association between a SNP within an InDel on chromosome 10 at 80,6Mbp and anthesis-to-silking date intervals (ASI8). Testing three genotypes, the –log10(p-value) of the SNP was 8.30 while testing only presence and absence, the –log10(pvalue) was 2.32. This stronger association was due to a difference for ASI8 between (i) lines homozygous for presence allele A and (ii) lines homozygous either for absence or presence allele B (Figure S3.6).

Among 18 InDels and 405 SNPs associated with phenotypic variations, 4 InDels (22%) and 27 SNPs (6%) were under selection. Among these four InDels, two were associated with flowering time and were located close to a region (Vgt3) well known on chromosome 3 (Salvi et al., 2007) and two other were associated with leaf number and were located on chromosome 8 in the vicinity of Vgt2 (Bouchet et al., 2013). The 25 SNPs under selection were associated with flowering time and leaf number and were located on the chromosome 8, close to the gene Zcn8. Two other SNPs under selection were associated with kernel weight and were located on chromosome 2 and 5.

Discussion

Do the InDels bring new information as compared to SNPs?

We found little difference between genetic structuration and relatedness obtained with InDels with those obtained with SNPs, except for IBS relatedness estimator which revealed slightly different pictures. This last estimator is more sensitive to recent mutations than IBD estimator which incorporates weighting by allele frequency (Astle and Balding, 2009). Our results suggested that these polymorphisms followed globally a similar evolutionary trajectory and that evolutionary forces act globally similarly on them. This result was relatively unexpected because these two types of polymorphism originated indeed from different mutation mechanisms and therefore occurred at different rates (Stankiewicz and Lupski, 2010). We could also expect that selection acts differently on InDels and SNPs since InDels could lead to a change in gene dosage or a deletion of genes whereas SNPs have a lower probability to produce drastic effects, by causing for instance stop codons.

Different reasons could explain the small difference between InDels and SNPs for genetic structuration and relatedness. First the linkage disequilibrium between InDels and their nearby SNPs certainly blurred the detection of evolutionary trajectory differences (Conrad et al., 2010). This hypothesis was well supported by the observation that 49% of anchored InDels were in high LD with at least one nearby SNP in a 2Mbp window (r2k>0.8) and that 72% of non-anchored InDels were in intermediate LD (r2>0.5) with at least one SNP along the genome. Second, InDels from our array were discovered on four temperate inbred lines, conducting to some ascertainment bias (Clark, 2005; Ganal et al., 2011). As a consequence, only a small subset of InDels segregating in our diversity panel were genotyped. We can expect that these cryptic InDels were partially captured by LD with some SNPs blurring the comparison between InDels and SNPs. Finally, Swanson-Wagner et al., (2010) observed that a majority of InDels were common to maize and teosinte, suggesting that most of these variants predate domestication and are probably neutral. As a consequence, the evolution of most InDels is probably driven only by genetic drift.

Although little difference between InDels and SNPs was observed for kinship and structure, 13 out 280 QTLs as well as 26 out 190 regions under selection were exclusively identified by InDels, although SNPs were found in the vicinity of all these regions. Also, 51% of 21,360 anchored InDels were not in high LD (r2K>0.8) in a 2Mbp window with any SNP and 28% of non-anchored InDels could not be mapped as their LD with anchored marker was too low (r2<0.5). Altogether, these results suggest that InDels probably targeted new genomic regions not already captured by SNPs and are complementary to SNP genotyping. Different hypotheses could explain this result. Firstly, the density of SNPs used may not be sufficient to well capture all InDel effects by LD. This could be reinforced by the higher density of InDels in telomeric than in peri-centromeric regions. Since the telomeric regions displayed shorter LD than peri-centromeric regions due to higher recombination rate, SNP density could be not sufficient in telomeric regions to capture the effect of these InDels by LD. Accordingly, Negro et al., (2018) showed that the SNP density obtained by combining SNPs from GBS, 600K and 50K arrays is not sufficient to capture all QTLs by GWAS in telomeric regions considering the LD in a diversity panel of dent inbred lines. In agreement with this hypothesis, we also observed that non anchored InDels mapped preferentially in peri-centromeric regions rather than telomeric regions using LD mapping. This pattern could also be explained by the inability to anchor InDels at the discovery step

due to high repeat contents in peri-centromeric regions (Darracq et al., 2018; Chapter 2). Secondly, InDels and SNPs could tag differently haplotypes involved in adaptation or trait variations as well exemplified by InDels involved in flowering time variation in Vgt3 (Salvi et al., 2007) region (Figure 3.4). An association with flowering time was also found for an InDel in the vicinity of Vgt1 (Salvi et al., 2002, 2007; Ducrocq et al., 2008), slightly lower than Vgt1 itself. This result suggests that InDels and SNPs could target different causal polymorphisms or haplotypes. Accordingly, we observed also that anchored InDels not in high LD (r2k>0.8) with any SNP have a SNP in their vicinity (median of 332bp). In the same way, Negro et al., (2018) showed that GBS and 600K arrays identified complementary genomic regions involved in trait variation by tagging different haplotypes rather than different regions.

Distribution of InDels suggested recurrent rearrangements

We observed a lower InDel density in peri-centromeric regions than telomeric regions and the distribution of InDels along the chromosome followed globally the SNPs distribution. The same pattern was observed in wheat with insertions and deletions detected in high recombination regions of the genome (Akhunov et al., 2003). In contrary, Lu et al., (2015) found InDels more often present in peri-centromeric regions and with a density positively correlated with repeat density. InDels from our array were discovered preferentially in low-copy regions which could explain this different repartition (Schnable et al., 2009).

We found a significant enrichment of not anchored InDels mapped thanks to anchored InDels rather than SNPs. It suggested the presence of many clusters of InDels in high LD. Accordingly, we identified two very large clusters of InDels on chromosome 6. One of these clusters contained 34 InDels in the short arm of the chromosome 6. All these InDels were discovered as deleted regarding the reference genome B73. Using CGH approach, these regions were previously identified as present in B73 and absent in Mo17 by Springer et al. (2009), then absent in 14 out 25 domesticated lines and in 11 out 14 teosinte lines by Swanson-Wagner et al., (2010). This region carries 25 genes, some with a strong variation in expression across lines (Hirsch et al., 2014). Additionally, this region displayed very large haplotypes (result not shown) suggesting either strong selection or a reduction in recombination. The presence of two large clusters of InDels on Chromosome 6 suggested that some regions were probably more susceptible to recurrent chromosomal rearrangements (Cooper et al., 2007; Stankiewicz and Lupski, 2010).

One reference genome is not sufficient

Among the 61,492 InDels genotyped by our array, 38,839 have their sequence absent from B73 reference genome (Table S3). As a consequence, these sequences have not been considered to design the 600K and 50K arrays (Ganal et al., 2011; Unterseer et al., 2014), for which SNPs were discovered by aligning reads from sequenced inbred lines on B73 reference genome. Similarly, SNPs from GBS were called using AllZeaGBSv2.7 database that was obtained by aligning reads from 17,280 lines on the B73 reference genome (Glaubitz et al., 2014). Although GBS can be used to discover and call SNPs in sequence that are not present in B73 reference genome (Lu et al., 2015), it was difficult to localize

precisely these variants on B73 reference genome and therefore impute them correctly limiting their use considering the high level of missing data for GBS. Classical genotyping tools could not target genomic regions not in the B73 reference genome and therefore conducted to an ascertainment bias (Hurgobin and Edwards, 2017). In maize, it could be very important considering the large amount of the sequence that are not present in B73 genome (1/3 of genome in Lu et al., 2015). In our study, 11,127 SNPs were genotyped within 91.4 Mbp sequence not present at all in B73 reference genomes. This allowed to identify new regions of interest (see below).

Additionally, density of SNPs genotyped by arrays and GBS was significantly lower in the sequences of 14,761 deletions (sequences present in B73 reference genome but absent from PH207, C103 or F2) than in the whole genome. Among these deletions, 3,476 were targeted by 11,933 SNPs from 50K, 600K and GBS whereas 8,057 InDels were targeted by 36,964 SNPs from our array. Within these InDel sequences, we also noted a strong enrichment of SNPs originated from GBS compared to 50K and 600K. Since no selection was applied for GBS, this suggests that some SNPs located in these InDels were filtered out during the selection of probes for designing arrays. These SNPs probably displayed a low call rate due to absence of hybridization of probe sequences for inbred lines for which the sequence was absent, leading to elimination of these probes during the design (Unterseer et al., 2014). Some SNPs were also probably not discovered or eliminated due to their bad quality at sequencing step because reads from some inbred lines sequenced did not or badly align to the reference genome due to presence of InDels.

InDels are therefore very important to take into account in genetic analysis in addition to SNPs from GBS and arrays. First, absence of InDels information increases genotyping errors because it conducts to impute a SNP genotype instead of absence of sequence for a missing genotype. Second, testing jointly the effect of SNP polymorphism and the absence of sequence on phenotypic variations could reveal new QTLs as well exemplified by a SNP within InDels associated with ASI8 on chromosome 10. All these results reinforce the idea that one reference genome is not sufficient to discover all SNPs within one species and that InDels bring new interesting information for genetic analysis.

Are InDels involved in adaptation to contrasted environment?

We detected a significant higher proportion of InDels under selection (0.22%) in comparison to SNPs (0.15%). This enrichment was reinforced when markers under selection were grouped according to LD. InDels were indeed involved in 29.8% of genomic under selection although they represented only 7.3% of markers. Interestingly, the highest number of associations for InDels under selection were found for flowering time (4 associations), that plays a key role in adaptation of maize to higher latitude (Tenaillon and Charcosset, 2011; Bouchet et al., 2013). Among these InDels, two were located on chromosome 3 close to a QTL for flowering time previously found (Salvi et al., 2011). We also identified an InDel under selection on the chromosome 2 at 7kbp of a gene implicated in drought tolerance (Virvoulet et al., 2012). The higher frequency of the presence of this InDel sequence in dent inbred lines suggests a specific selection for this group. Genes affected by InDels were reported to be often associated to biotic and abiotic stress in many species (Beló et al., 2010; Chia et al., 2012; Muñoz-Amatriaín et al., 2013; Saxena et al., 2014; Golicz et al., 2016; Montenegro et al., 2017; Gabur et al., 2018; Hübner et al., 2019). Altogether, these results suggested that InDels probably affected dispensable genes implicated in adaptation to contrasted environments.

Among InDels under selection, we found 38, 23 and 1 InDels as well as 308, 170 and 23 SNPs fixed in tropical, flint or dent inbred lines but segregating in other groups, respectively. Gouesnard et al., (2017) found also a majority of SNPs that were fixed in tropical inbred lines and segregated among flint and dent inbred lines. This result was expected since agro-climatic conditions of tropical inbred lines are strongly different from those for flint (notably for Northern Flint) and in a lesser extent for dent, which are both more adapted to temperate climates. Surprisingly, we found only one InDel that was fixed in dent lines and segregated among tropical and flint inbred lines. It may be explained by the fact that the dent group was originated from a hybridization between tropical and flint groups (Tenaillon and Charcosset, 2011, Brandenburg et al., 2017). As a consequence, we expected that dent lines would display intermediate allelic frequencies between tropical and flint lines.

Why were the InDels not more strongly associated than the SNP?

InDels were slightly less frequently associated with trait variation than SNPs (0.03% against 0.05%) but this difference was not significant. This slight difference was attenuated when considering the QTL number. This result was surprising since we could expect a stronger effect of InDels on phenotypic variations since they could affect gene content and therefore affect trait variation more than SNP.

Different hypotheses could explain this result. First, only 8% of InDels were present/absent at one locus and not present elsewhere in the genome (PAVs). Conversely, 92% of InDels did not carry a specific sequence. Therefore, effect of gene loss could be partially compensated by the presence of another copy elsewhere in the genome. Swanson-Wagner et al., (2010) found that genes affected by SVs are often part of genes families. Therefore, the loss of one single gene copy could be tolerated thanks to the compensation by genes from the same family. Moreover, maize is an ancient polyploid and many genes remain duplicated, facilitating this functional compensation (Schnable et al., 2011). This compensation phenomenon was well known in polyploid context since homoelogous chromosomes could compensate partially or totally chromosome loss (Hurgobin et al., 2018).

Secondly, we certainly reduced the genotyping complexity of large InDels. In our study, we first removed internal probes displaying too divergent genotyping with other within InDel. We then averaged allelic frequencies of probes across InDel, and discretizing this allelic frequencies as present/absent. Probe genotyping divergences within large InDels could indicate that several recurrent and different rearrangements occurred in the genomic region leading to deletion/insertion of different part of the genomic regions (Chapter 2). Consequently, taking into account this genotyping complexity by analyzing the effect of different haplotypes on trait variations could help to capture different events with different phenotypic impact. Note that divergences between internal probes (OTV and Mono) could also indicate the presence of residual heterozygosity for presence/absence (hemizygous) since Affymetrix® algorithms do not currently call hemizygous individuals for these probes, whereas they do for BP probes (Chapter 2).

Thirdly, InDels with a strong deleterious effect on phenotype have probably been partially purged from our population. Indeed, inbred lines were derived by selfing from highly heterozygous landraces, which probably purged highly deleterious alleles (Duvick, 2005; Charlesworth and Willis, 2009). The majority of our InDels (97%) may not have a strong deleterious effect since there sequences are present in at least 18 inbred lines (5%) in the panel. On the other hand, we observed a strong deficit of loci with a quasi-fixation of the absent allele, suggesting that absence of sequence was partially

deleterious and counterselected. Considering their putative deleterious effect due to gene loss, we also expected that residual heterozygosity was more maintained for InDels than SNPs during selfing, conducting to stretches of heterozygosity. Accordingly, we observed a higher heterozygosity rate for presence/absence than SNPs. Also, we found a higher proportion of InDels under selection and thus more differentiated than SNP between genetic groups. Highl differentiation is known to reduce power of GWAS (Rincent et al., 2014) so that some InDels having passed a strong differential selection pressure across environments may not be detectable in association studies.

Fourthly, as suggested above, InDels seem to be more involved in adaptive traits. Except for flowering time, which showed a high number of associations with InDels, traits measured and analyzed in GWAS in our study were not strongly related to stress tolerance. Therefore, we could find more associations by testing the effect of InDels on traits relative to abiotic and biotic stress. We could consider in further experiments appropriate panels measured on different environments with a wide range of climatic conditions. By this way, we will probably identify InDels implicated in adaptation to specific environments. We identified one cluster of 34 InDels in the short arm of the chromosome 6. This region was not detected under selection or associated with any trait variation. However, Millet et al., (2016) performed GWAS on 29 field experiments in contrasting conditions across Europe with SNPs. They found a QTL close to this cluster with a large effect on yield and grain number especially under hot climatic conditions and carrying a gene involved in drought tolerance (ZmASR8). This region was also identified as involved in yield variation with pleiotropic effect on plant and ear height in a multiparental population of recombinant lines derived from eight founders by Dell'Acqua et al. (2015).

Finally, previous studies reported that InDels complementation in hybrids could lead to a substantial hybrid vigor according to dominance hypothesis (Fu and Dooner, 2002; Beló et al., 2010; Sun et al., 2018). Regarding the high number of InDels in maize, it would be very difficult to create inbred lines containing all genes (Swanson-Wagner et al., 2010). Moreover, Yang et al., (2017) showed an important role of incomplete dominance of deleterious alleles to explain heterosis. If we assume that the loss of sequence could be deleterious, we should find more InDels associated with trait variations in GWAS by testing the dominance effect instead of testing the additive effect of InDels on hybrid phenotypic values. Incorporating dominance effect estimated with InDels on genomic prediction models could also improve the phenotypic predictions in plant breeding scheme.

Acknowledgments

We thank Pierre Dubreuil for his contribution to the mapping of InDels and for helpful discussion. We thank Cyril Bauland for the field experiment. This work was supported by the project CNV-MAIZE (ANR-10-GENM-003) and the project Investment for the future AMAIZING ANR-10-BTBR-01 (ANR-PIA AMAIZING) and France Agrimer. C. Mabire PhD student is jointly funded by the program CNV4sel of the INRA metaprogram SelGen and by the Plant Biology and Breeding department of the French National Institute for Agricultural Research (INRA).

Supplementary figures

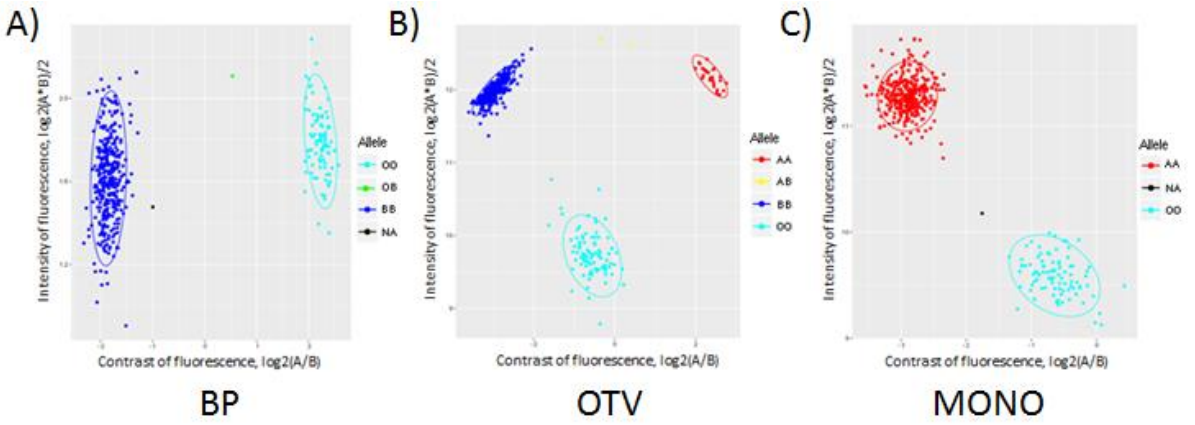


Figure S3.1: Clustering produced by Affymetrix® algorithm for BP, OTV and MONO probes from InDel based on both fluorescence contrast (X axis) and intensity (Y axis) of the 362 inbred lines. Red, blue and yellow dots indicated the presence of the sequence (genotype "present") either homozygous for allele A (AA), or allele B (BB) or heterozygous (AB), respectively. Cyan and green indicated that the sequence was absent in the individual (OO), or only in one copy of the sequence, e.g hemizygous for presence/absence (OB or OA). Black dots indicated individuals for which no genotype could be assigned (Missing data).

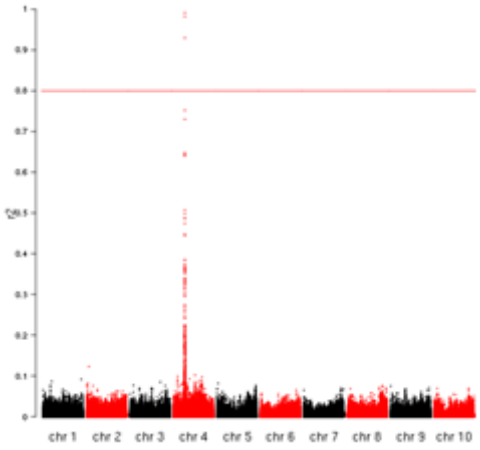


Figure S3.2: Manhattan plot of LD measures (r2) between a candidate marker and 897,868 anchored markers. The estimated position of the candidate is the position of the marker(s) with the highest LD measure(s).

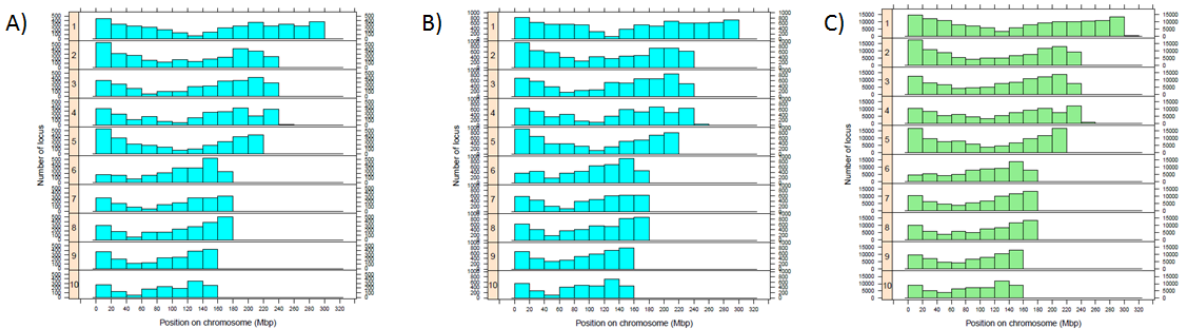


Figure S3.3: Distribution along 10 chromosomes of 25,544 anchored InDel markers (A) before and (B) after mapping 27,479 non anchored InDels and (C) 872,324 anchored SNPs

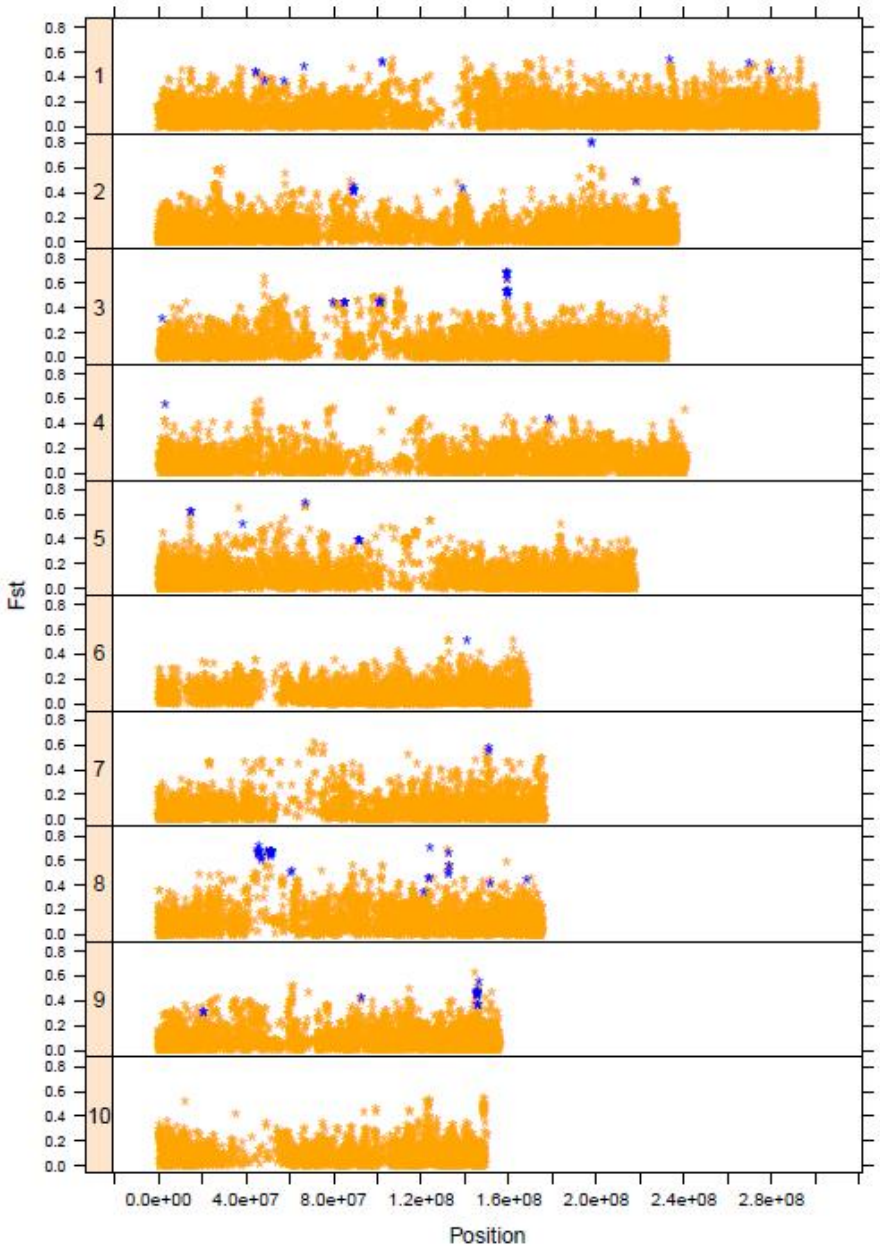
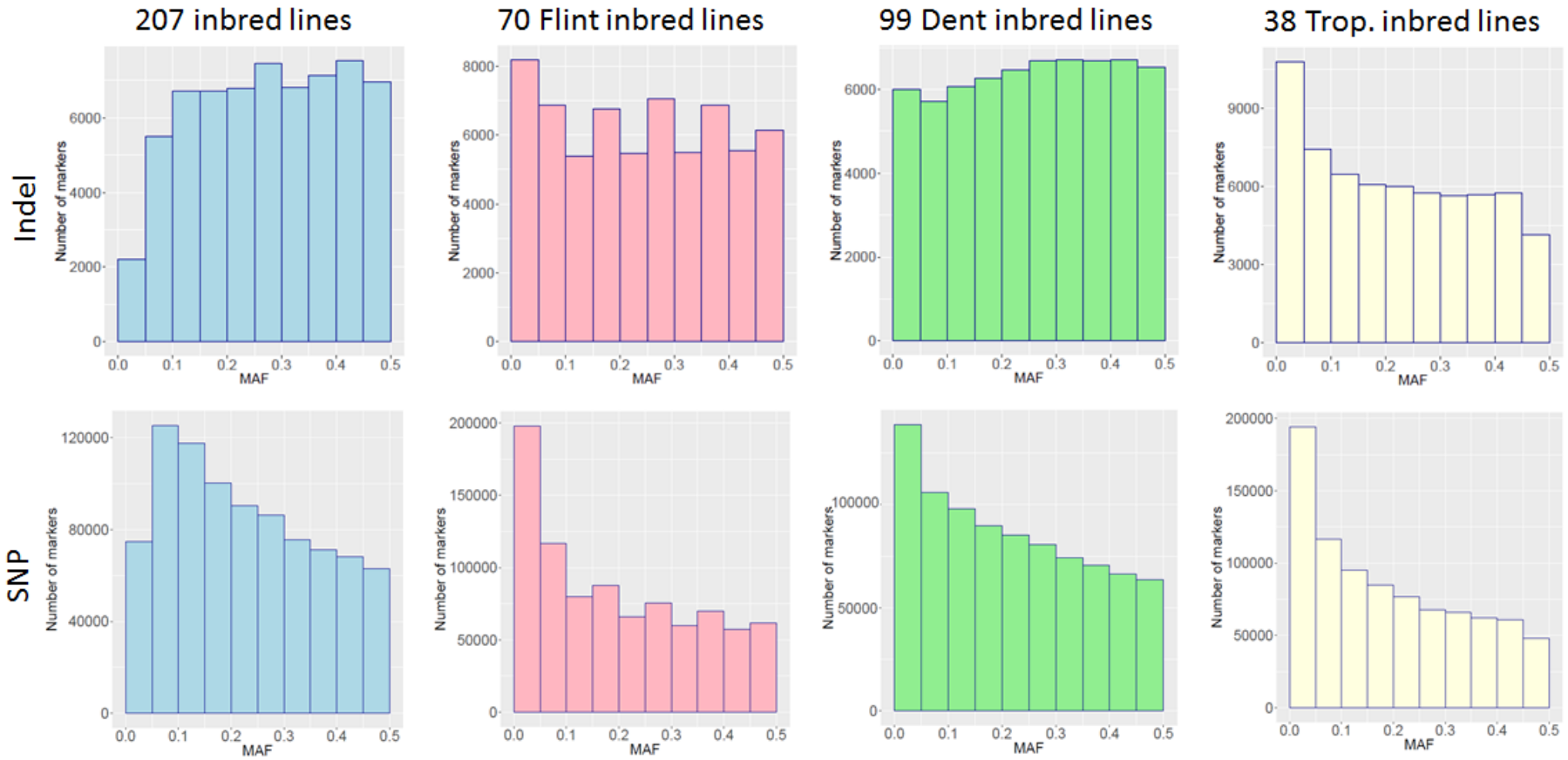


Figure S3.4: Variation of Fst (Nei) for 53,023 anchored InDels along the 10 chromosomes. Blue dots represent InDels under selection according to Bayescan analysis (log10(PO) superior to 1).

Figure S3.5: Distribution of minor allele frequency (A) for 63,803 InDel markers and (B) 873,011 markers from 50K, 600K and GBS for 207 inbred lines, 70 flint, 99 dent and 38 tropical inbred lines assigned to three genetic groups (Flint, Dent and Tropical) with an assignment probability superior to 0.7.



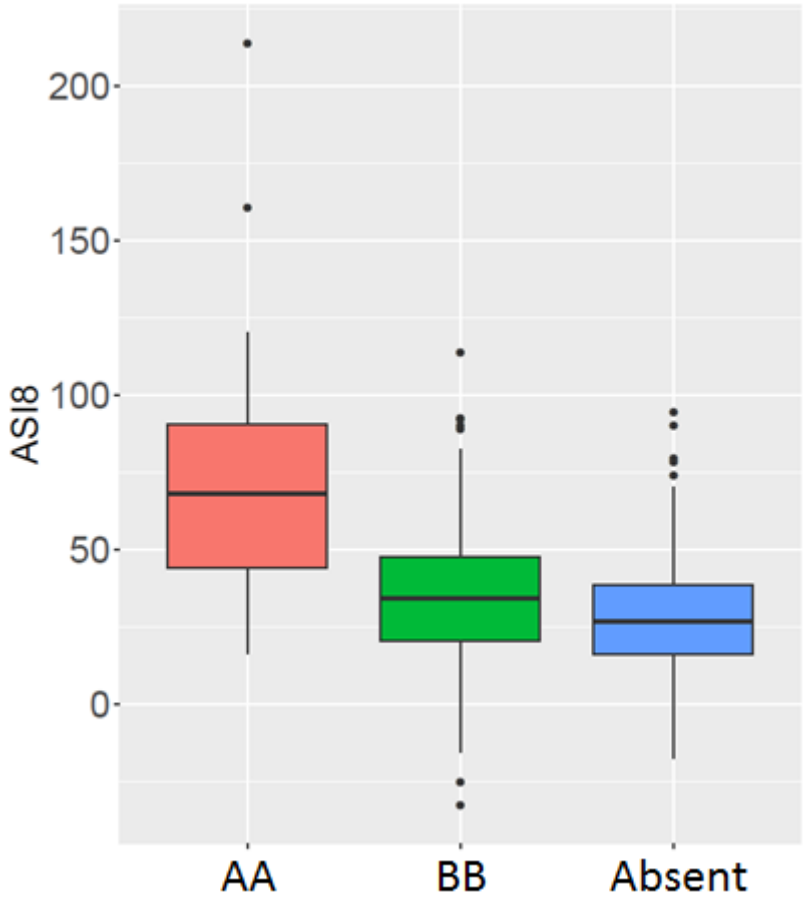


Figure S3.6: Boxplot of ASI8 according to three genotypes at SNP AX-114385991 within InDel CNVMAIZE_DEL_34922 located on chromosome 10 at 80,6 Mbp. 26, 203 and 133 lines were homozygous for absence (blue), for presence with allele A (red) or allele B (green) at SNP respectively. Black horizontal line represented median.

Supplementary tables

Table S3.1: Description of 23 traits analyzed in GWAS. Growth Degree Days (GDD) correspond to trait calculated according to thermal time.

Abbreviation	Name
Yield	
EARNB	Ears per plant
KERROWNB	Kernels per row
KW	Kernel weight per plant (g)
ROWNB	Kernel row number
TKW	Thousand kernel weight (g)
Phenology	
ASI8	Anthesis silking interval (GDD)
Em8	Emergence (GDD)
EmVIG	Emergence vigor (qualitative)
FFLW8	Female flowering (GDD)
MFLW8	Male flowering (GDD)
S3LF8	3rd leaf stage (GDD)
Architecture	
BRNB	Tassel branch number (qualitative)
EARHT	Ear insertion height (cm)
HL	Ear husk leaf length (cm)
LFNB	Leaf number
LFNBa	Leaf number above top ear
LFNBb	Leaf number below top ear
TaAk	Tassel secondary branches angle (qualitative)
TL	Tassel length (cm)
SL	Tassel length above top secondary branches (cm)
BZL	Length from the first branch to the top branch of the tassel (cm)
TaSt	Tassel floppiness (qualitative)
PTHT	Plant height (cm)

Table S3.3: Number of sequences present or absent for 61,492 InDels for four inbred lines used to discover InDels.

Inbred lines	Present		Absent		Hemizygous		Missing	
B73	20,755	(33.8%)	38,839	(63.2%)	228	(0.4%)	1,670	(2.7%)
F2	42,971	(69.9%)	17,420	(28.3%)	206	(0.3%)	895	(1.5%)
C103	31,739	(51.6%)	28,398	(46.2%)	317	(0.5%)	1,038	(1.7%)
PH207	31,374	(51.0%)	28,910	(47.0%)	347	(0.6%)	861	(1.4%)

Table S3.2: Testing of InDels mapping approach with 9 different linkage disequilibrium thresholds by sampling randomly 100 InDels candidates with known position. For each threshold, we accounted for number of InDels mapped at a wrong position, correctly mapped and not mapped. For InDels correctly mapped n, we calculated the average genetic and physical distance between the predicted and real position

Threshold	Nbr of InDels correctly mapped	Nbr of InDels mapped at a wrong position	Nbr of not mapped InDel	Genetic distance (cM) between real and predicted position (mean)	Physical distance (bp) between real and predicted. position (mean)
0.8	52	1	47	0.0006	1,905
0.7	59	2	39	0.0011	2,059
0.6	67	2	31	0.0011	2,359
0.5	74	4	22	0.0013	2,436
0.4	79	7	14	0.0017	2,688
0.3	79	7	14	0.0017	2,688
0.2	79	7	14	0.0017	2,688
0.1	79	7	14	0.0017	2,688
0	79	7	14	0.0017	2,688

Table S3.4: Estimation of genetic LD extent (cM) for a r2k = 0.1 using Hill and Weir model with r2k measure between InDels (InDel-InDel), InDels and SNPs (InDel-SNP) and between SNPs (SNP-SNP) 1. For the whole genome, we calculated average values weighted by the number of markers on each chromosome.

Chromosome	InDel-InDel	InDel-SNP	SNP-SNP
1	0.0205	0.0230	0.0443
2	0.0202	0.0242	0.0479
3	0.0307	0.0333	0.0703
4	0.0220	0.0320	0.0659
5	0.0272	0.0246	0.0359
6	0.0656	0.0390	0.0511
7	0.0325	0.0302	0.0572
8	0.0253	0.0400	0.0903
9	0.0310	0.0354	0.0709
10	0.0285	0.0367	0.0573
Whole genome	0.0290	0.0306	0.0573

Table S3.5: Variation of InDels number along the 10 chromosomes before and after mapping using linkage disequilibrium. . For each chromosome and 20Mbp bins, we displayed the number of anchored InDels before mapping, after mapping and the total number of InDel. We tested the enrichment of InDel in each bin after mapping (p-value). Gray bins indicated position of centromere.

Chr.		0-20	20-40	40-60	60-80	80-100	100-120	120-140	140-160	160-180	180-200	200-220	220-240	240-260	260-280	280-300	300-320	Total
1	Before	438	326	276	253	199	137	63	148	247	279	364	279	316	282	373	5	3,985
1	Mapped	370	317	297	307	335	151	28	233	298	222	312	327	298	349	353	8	4,205
1	Total	808	643	573	560	534	288	91	381	545	501	676	606	614	631	726	13	8,190
1	p-value	0.0891	0.5848	0.9239	0.3626	0.002	0.8717	0.0393	0.0253	0.3904	0.087	0.1442	0.4831	0.4565	0.2671	0.4331	0.8445	
2	Before	547	307	262	160	124	174	129	182	232	419	356	244	-	-	-	-	3,136
2	Mapped	410	347	321	243	145	245	211	266	235	317	394	377	-	-	-	-	3,511
2	Total	957	654	583	403	269	419	340	448	467	736	750	621	-	-	-	-	6,647
2	p-value	0.001	0.9734	0.5557	0.0794	0.8771	0.1832	0.0439	0.1075	0.5628	0.0042	0.9558	0.0219	-	-	-	-	
3	Before	346	262	171	54	103	111	208	225	333	349	414	292	-	-	-	-	2,868
3	Mapped	342	321	199	115	125	149	336	286	337	320	449	199	-	-	-	-	3,178
3	Total	688	583	370	169	228	260	544	511	670	669	863	491	-	-	-	-	6,046
3	p-value	0.4218	0.5113	0.821	0.0145	0.7307	0.3926	0.012	0.3927	0.533	0.1846	0.8858	0.0039	-	-	-	-	
4	Before	353	236	120	182	81	52	154	280	291	368	205	351	16	-	-	-	2,689
4	Mapped	291	304	210	227	74	55	197	321	219	334	286	301	21	-	-	-	2,840
4	Total	644	540	330	409	155	107	351	601	510	702	491	652	37	-	-	-	5,529
4	p-value	0.0955	0.2033	0.0088	0.358	0.6553	1	0.3237	0.5982	0.042	0.2891	0.0845	0.1583	0.8066	-	-	-	
5	Before	539	344	208	189	143	78	99	189	268	368	416	-	-	-	-	-	2,841
5	Mapped	406	320	176	194	162	56	174	297	275	349	400	-	-	-	-	-	2,809
5	Total	945	664	384	383	305	134	273	486	543	717	816	-	-	-	-	-	5,650
5	p-value	0.0341	0.6969	0.4295	0.8738	0.534	0.3463	0.0072	0.0042	0.8419	0.7896	0.8562	-	-	-	-	-	
6	Before	174	158	78	155	208	314	319	527	241	-	-	-	-	-	-	-	2,174
6	Mapped	183	273	94	218	229	323	365	404	209	-	-	-	-	-	-	-	2,298
6	Total	357	431	172	373	437	637	684	931	450	-	-	-	-	-	-	-	4,472
6	p-value	1	0.0039	0.6661	0.1244	0.8449	0.8796	0.5927	0.0128	0.2668	-	-	-	-	-	-	-	
7	Before	289	173	92	51	138	186	293	284	323	-	-	-	-	-	-	-	1,829
7	Mapped	279	274	116	51	243	266	313	339	291	-	-	-	-	-	-	-	2,172
7	Total	568	447	208	102	381	452	606	623	614	-	-	-	-	-	-	-	4,001
7	p-value	0.1797	0.0837	0.8469	0.6689	0.0272	0.2697	0.4877	1	0.0639	-	-	-	-	-	-	-	

Chr.		0-20	20-40	40-60	60-80	80-100	100-120	120-140	140-160	160-180	180-200	200-220	220-240	240-260	260-280	280-300	300-320	Total
8	Before	310	179	59	164	167	239	284	382	495	-	-	-	-	-	-	-	2,279
8	Mapped	301	226	115	196	239	312	242	436	389	-	-	-	-	-	-	-	2,456
8	Total	611	405	174	360	406	551	526	818	884	-	-	-	-	-	-	-	4,735
8	p-value	0.5011	0.3854	0.0256	0.6065	0.1092	0.2164	0.1511	0.6765	0.0153	-	-	-	-	-	-	-	
9	Before	368	201	113	131	234	250	374	423	-	-	-	-	-	-	-	-	2,094
9	Mapped	303	214	176	217	252	328	342	387	-	-	-	-	-	-	-	-	2,219
9	Total	671	415	289	348	486	578	716	810	-	-	-	-	-	-	-	-	4,313
9	p-value	0.0885	1	0.065	0.0184	0.9543	0.1612	0.3063	0.2812	-	-	-	-	-	-	-	-	
10	Before	269	121	44	177	227	189	359	263	-	-	-	-	-	-	-	-	1,649
10	Mapped	283	144	72	230	255	262	350	195	-	-	-	-	-	-	-	-	1,791
10	Total	552	265	116	407	482	451	709	458	-	-	-	-	-	-	-	-	3,440
10	p-value	0.8688	0.7105	0.2223	0.3273	0.8704	0.155	0.4644	0.0326	-	-	-	-	-	-	-	-	

Table S3.6: List of 30 InDels under decisive selection (log10(PO) > 2) found by BayeScan analysis between dent, flint, and tropical inbred lines. For each InDel, we calculated two different Fst measures and the frequency of the presence for the three genetic groups (Flint, Dent and Tropical). * indicate predicted position using LD approach for non-anchored InDels

Chr	Position	log10(PO)	Fstb	Fst (Nei)	Flint	Dent	Tropical	SV_TYPE	SV_LENGTH	FRAC_SPE	Dist. to gene	Closest_gene
1*	102,240,644*	3	0.42	0.52	0.97	0.37	0	INS	427	0.54	NA	NA
1*	102,240,644*	2.92	0.42	0.53	0.97	0.35	0	INS	641	0.89	NA	NA
2	89,239,710	2.15	0.41	0.45	0.1	0.27	1	DEL	4,449	0.09	0	AC177933.2_FG003
2*	197,680,674*	2.55	0.41	0.8	0.89	0.02	0	INS	740	0.07	NA	NA
2*	197,690,647*	3.1	0.41	0.82	0.06	0.98	0.92	INS	1,254	0.05	NA	NA
6	140,944,729	2.07	0.38	0.51	0.01	0.66	0.95	DEL	3,346	0.12	0	GRMZM2G001572
8	45,767,369	2.27	0.39	0.72	0.04	0.95	0.79	DEL	621	0.5	44309	GRMZM2G077259
8*	51,330,228*	2.07	0.37	0.67	0.96	0.07	0.24	INS	857	0.31	NA	NA
8*	51,330,228*	2.01	0.37	0.67	0.96	0.07	0.24	INS	3,606	0.2	NA	NA
8*	51,330,228*	2.01	0.37	0.67	0.96	0.07	0.24	INS	289	0.25	NA	NA
8*	51,330,228*	2.14	0.37	0.67	0.96	0.07	0.24	INS	4,143	0.32	NA	NA
8*	51,330,228*	2	0.37	0.67	0.96	0.07	0.24	INS	795	0.31	NA	NA
8*	51,330,228*	2	0.37	0.67	0.96	0.07	0.24	INS	1,077	0.13	NA	NA
8*	51,330,228*	2	0.37	0.67	0.96	0.07	0.24	INS	281	0.51	NA	NA
8*	51,330,228*	2.02	0.37	0.67	0.96	0.07	0.24	INS	359	0.19	NA	NA
8	61,052,290	2.16	0.39	0.51	1	0.21	0.68	INS	264	0	19085	GRMZM2G582918
8*	132,674,942*	2.47	0.39	0.56	0.99	0.27	0.05	INS	497	0.66	NA	NA
9*	145,520,432*	2.27	0.4	0.46	0	0.58	0.89	INS	628	0.15	NA	NA
9*	145,520,432*	2.44	0.41	0.47	0	0.61	0.89	INS	694	0.71	NA	NA
9*	145,520,432*	2.35	0.4	0.46	0	0.59	0.89	INS	904	0.86	NA	NA
9*	145,520,432*	2.37	0.4	0.46	0	0.59	0.89	INS	993	0.79	NA	NA
9*	145,520,432*	2.44	0.4	0.47	0	0.6	0.89	INS	362	0.29	NA	NA
9*	145,520,432*	2.42	0.41	0.46	0	0.59	0.89	INS	773	1	NA	NA
9*	145,520,432*	2.3	0.4	0.47	0	0.6	0.89	INS	316	1	NA	NA
9*	145,520,432*	2.23	0.4	0.46	0	0.58	0.89	INS	310	0.6	NA	NA
9*	145,520,432*	2.7	0.41	0.48	0	0.59	0.92	INS	510	1	NA	NA
9*	145,520,432*	2.34	0.4	0.47	0	0.6	0.89	INS	310	1	NA	NA
9*	145,520,432*	2.3	0.4	0.46	0	0.58	0.91	INS	205	0	NA	NA
9	145,594,253*	2.22	0.4	0.45	1	0.43	0.11	DEL	588	0.17	0	GRMZM2G356579
NA	NA	2.85	0.41	0.52	0.47	0.05	0.97	INS	950	0.17	NA	NA

Table S3.7: List of 96 SNPs and the Mite Vgt1, putatively under decisive selection (log10(PO) > 2) found by BayeScan analysis between dent, flint, and tropical inbred lines. For each region, we calculated two different Fst measures (Fstb and Fst(Nei)), the frequency of the presence for the three genetic groups (Flint, Dent and Tropical).

Chr	Position	log10(PO)	Fstb	Fst (Nei)	Flint	Dent	Tropical	Dist. to gene	Closest_gene
1	47,990,736	2.07	0.39	0.5	0.63	0.09	0.97	0	GRMZM2G017536
1	47,991,072	2.13	0.39	0.49	0.61	0.09	0.97	0	GRMZM2G017536
1	47,992,619	2.03	0.39	0.49	0.61	0.09	0.97	0	GRMZM2G017536
1	48,093,341	3.4	0.43	0.56	0.53	0.03	0.97	0	GRMZM2G057027
1	48,094,360	3.4	0.43	0.55	0.44	0.03	0.97	0	GRMZM2G057027
1	48,095,528	3.7	0.43	0.55	0.44	0.03	0.97	0	GRMZM2G057027
1	48,530,468	3.7	0.44	0.67	0	0.81	0.03	3,227	GRMZM2G317487
1	102,240,644	2.8	0.42	0.51	0.96	0.35	0	7,601	GRMZM2G068917
1	102,245,022	2.4	0.41	0.5	0.96	0.36	0	0	GRMZM2G068917
1	287,201,827	2.03	0.41	0.49	1	0.35	1	1,049	GRMZM2G031062
1	287,201,853	2.04	0.42	0.49	1	0.35	1	1,023	GRMZM2G031062
2	20,556,142	3.4	0.41	0.73	0.91	0.03	0.82	8,579	GRMZM2G475678
2	89,273,822	2.4	0.42	0.46	0.91	0.73	0	6,251	GRMZM2G173738
2	89,321,974	2.3	0.42	0.46	0.91	0.73	0	54,403	GRMZM2G173738
2	89,417,861	2.49	0.42	0.47	0.91	0.74	0	47,324	GRMZM2G492768
2	194,506,047	2.27	0.41	0.56	0.94	0.23	1	3,794	GRMZM2G123973
2	197,680,674	2.34	0.42	0.8	0.89	0.02	0	0	GRMZM2G315902
2	197,683,873	1000	0.45	0.9	0.97	0.02	0.03	0	AC185415.3_FG005
2	197,684,098	3.7	0.46	0.9	0.97	0.02	0.03	0	AC185415.3_FG005
2	197,684,153	2.18	0.42	0.8	0.89	0.02	0	0	AC185415.3_FG005
2	197,685,091	3.1	0.42	0.84	0.96	0.02	0.08	0	AC185415.3_FG005
2	197,685,162	3.1	0.42	0.84	0.96	0.02	0.08	0	AC185415.3_FG005
2	197,685,427	3.4	0.43	0.84	0.96	0.02	0.08	0	AC185415.3_FG005
2	197,685,944	2.85	0.43	0.84	0.96	0.02	0.08	0	AC185415.3_FG005
2	197,686,210	3.22	0.43	0.84	0.96	0.02	0.08	0	AC185415.3_FG005
2	197,686,340	3	0.42	0.84	0.96	0.02	0.08	0	AC185415.3_FG005
2	197,686,627	3.4	0.42	0.84	0.96	0.02	0.08	0	AC185415.3_FG005
2	197,690,647	2.92	0.43	0.84	0.96	0.02	0.08	1,141	GRMZM5G851990
2	197,731,486	2.47	0.41	0.82	0.94	0.02	0.08	5,238	AC233910.1_FG010
2	197,733,696	3.1	0.43	0.84	0.96	0.02	0.08	3,028	AC233910.1_FG010
2	197,734,054	3.4	0.44	0.86	0.97	0.02	0.08	0	AC233910.1_FG010
2	199,714,822	2.08	0.42	0.42	0.19	0.21	1	12,907	GRMZM2G173162
3	82,620,376	2.8	0.45	0.65	0.14	0	0.89	448,722	GRMZM2G081380
5	51,214,667	2.04	0.4	0.56	0.97	0.26	1	0	GRMZM2G134156
5	135,622,653	2.05	0.4	0.54	0.99	0.29	1	20,022	GRMZM2G069596
6	97,413,823	2.04	0.4	0.45	0	0.57	0.89	529	GRMZM2G512400
7	150,514,000	2.58	0.41	0.58	1	0.26	0.97	0	GRMZM2G437675
7	150,791,000	2.62	0.42	0.59	1	0.25	0.97	0	GRMZM2G175799
7	150,978,352	2.03	0.4	0.55	1	0.29	0.97	6,178	GRMZM2G161809
8	45,880,166	2.16	0.39	0.72	0.96	0.05	0.21	0	GRMZM2G326945
8	61,520,087	2.23	0.41	0.55	1	0.18	0.68	0	GRMZM2G164419
8	123,505,878	1000	0.51	0.53	1	0.68	0	3,026	GRMZM2G179274
8	123,505,879	1000	0.51	0.53	1	0.68	0	3,025	GRMZM2G179274
8	123,505,881	1000	0.51	0.53	1	0.68	0	3,023	GRMZM2G179274
8	123,505,897	1000	0.51	0.53	1	0.68	0	3,007	GRMZM2G179274
8	123,506,087	1000	0.51	0.53	1	0.67	0	2,817	GRMZM2G179274

Chr	Position	log10(PO)	Fstb	Fst (Nei)	Flint	Dent	Tropical	Dist. to gene	Closest_gene
8	123,506,089	1000	0.51	0.53	1	0.67	0	2,815	GRMZM2G179274
8	123,506,141	1000	0.51	0.53	1	0.67	0	2,763	GRMZM2G179274
8	123,507,284	1000	0.51	0.53	1	0.68	0	0	GRMZM2G179274
8	123,509,445	1000	0.46	0.51	0.99	0.67	0	0	GRMZM2G479987
8	123,509,765	2.16	0.41	0.46	1	0.68	0.08	0	GRMZM2G479987
8	123,510,186	1000	0.51	0.54	1	0.69	0	0	GRMZM2G479987
8	123,510,268	2.16	0.4	0.46	1	0.69	0.08	0	GRMZM2G479987
8	123,510,344	1000	0.51	0.53	1	0.68	0	0	GRMZM2G479987
8	123,510,397	2.15	0.41	0.46	1	0.68	0.08	0	GRMZM2G479987
8	123,510,555	2.18	0.41	0.46	1	0.68	0.08	0	GRMZM2G479987
8	123,510,742	1000	0.51	0.54	1	0.69	0	0	GRMZM2G479987
8	123,510,776	2.7	0.43	0.49	1	0.68	0.05	0	GRMZM2G479987
8	123,510,820	2.66	0.43	0.49	1	0.68	0.05	0	GRMZM2G479987
8	123,511,186	2.66	0.43	0.49	1	0.68	0.05	0	GRMZM2G479987
8	123,532,955	2.13	0.41	0.66	0.93	0.14	0	23,948	GRMZM2G479987
8	123,786,068	2.42	0.4	0.74	0.96	0.1	0.03	25,398	GRMZM2G180668
8	123,907,065	2.74	0.4	0.72	0.97	0.11	0.05	0	GRMZM2G136369
8	123,908,182	2.7	0.4	0.74	0.97	0.1	0.05	0	GRMZM2G136369
8	131,985,083	3.1	0.44	0.47	1	0.47	0.05	0	Mite Vgt1
8	132,067,949	2.34	0.41	0.51	0.96	0.34	0	23,746	GRMZM2G393039
8	132,092,798	2.58	0.42	0.51	0.96	0.34	0	3,955	GRMZM2G393039
8	132,202,886	2.01	0.38	0.51	0.97	0.34	0.03	0	GRMZM2G173763
8	132,203,990	2.8	0.46	0.84	0.89	0	0	0	GRMZM2G173763
9	145,520,365	2.15	0.4	0.45	0	0.56	0.89	0	GRMZM2G093254
9	145,520,432	2.12	0.4	0.46	0	0.58	0.89	0	GRMZM2G093254
9	145,529,036	2.16	0.41	0.46	0	0.58	0.89	0	GRMZM2G528283
9	145,529,583	2.14	0.4	0.46	0	0.58	0.89	1,561	GRMZM2G528283
9	145,681,844	2.18	0.41	0.46	0	0.58	0.89	63,978	GRMZM5G884709
9	145,683,950	2.35	0.41	0.46	0	0.58	0.89	66,084	GRMZM5G884709
9	145,731,569	2.08	0.41	0.46	0	0.58	0.89	36,302	GRMZM2G328268
9	145,765,011	2.01	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,765,201	2.18	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,765,347	2.13	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,765,520	2.09	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,765,889	2.25	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,766,205	2.14	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,766,303	2.32	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,766,428	2.19	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,766,484	2.08	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,766,510	2.22	0.4	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,766,617	2.07	0.4	0.44	0	0.58	0.87	0	GRMZM2G328268
9	145,766,793	2.27	0.41	0.45	0	0.57	0.89	0	GRMZM2G328268
9	145,766,810	2.09	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,766,939	2.28	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,766,999	2.14	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,767,029	2.06	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,767,102	2.12	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,767,138	2.15	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,767,240	2.25	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,767,296	2.27	0.4	0.46	0	0.58	0.89	0	GRMZM2G328268
9	145,767,339	2.25	0.41	0.46	0	0.58	0.89	0	GRMZM2G328268

Table S3.8: List of most significant associated SNPs (-log10(pvalue) >7). For each association, we display the -log10(pvalue) ("Log"), the minor allelic frequency ("MAF"), the distance to the closest gene and the name of the closest gene ("FGS_DIST" and "Closest gene" respectively) and the origin of the SNP ("Origin").

Trait	Chr.	Position	-Log(pvalue)	MAF	FGS_DIST	Closest gene	Origin
LFNBa	1	27,901,951	9.03	0.07	0	GRMZM2G040561	600K
HL	1	165,252,698	7.51	0.08	0	GRMZM2G335438	GBS
HL	1	165,412,333	10.26	0.06	2,216	GRMZM5G875002	600K
HL	1	165,413,285	7.56	0.1	3,168	GRMZM5G875002	600K
HL	1	165,414,523	10.26	0.06	4,406	GRMZM5G875002	600K
HL	1	165,414,651	7.56	0.1	4,534	GRMZM5G875002	600K
HL	1	167,595,857	9.82	0.08	0	GRMZM5G831313	GBS
FFLW8	1	56,077,700	8.02	0.11	4,557	GRMZM2G119850	GBS
MFLW8	1	56,077,943	7.98	0.11	4,800	GRMZM2G119850	GBS
MFLW8	1	182,089,229	7.28	0.05	0	GRMZM2G023906	GBS
LFNB	1	228,051,746	7.22	0.16	0	GRMZM5G870629	600K
EARHT	1	287,027,200	7.61	0.04	28,150	GRMZM2G017299	GBS
KW	2	194,506,047	7.01	0.31	3,794	GRMZM2G123973	600K
EmVIG	2	194,485,292	8.58	0.06	0	GRMZM2G124011	GBS
FFLW8	2	173,832,780	8.01	0.14	0	GRMZM2G134998	GBS
MFLW8	2	173,832,780	7.24	0.14	0	GRMZM2G134998	GBS
EARHT	2	16,940,666	7.67	0.21	0	GRMZM2G039455	600K
EARNB	2	38,988,722	8.94	0.03	0	GRMZM2G151418	600K
FFLW8	3	158,974,594	7.06	0.48	0	GRMZM2G171622	600K
FFLW8	3	158,974,756	7.11	0.36	0	GRMZM2G171622	600K
MFLW8	3	158,974,594	7.85	0.48	0	GRMZM2G171622	600K
MFLW8	3	158,974,756	7.38	0.36	0	GRMZM2G171622	600K
TaSt	3	211,604,808	7.59	0.09	7,641	GRMZM2G121753	600K
HL	4	30,473,933	7.99	0.2	0	GRMZM2G126253	GBS
HL	4	31,522,383	8.27	0.28	0	GRMZM2G052670	GBS
KW	5	51,214,667	7.6	0.3	0	GRMZM2G134156	600K
HL	5	214,938,599	8.05	0.07	1,554	GRMZM2G512422	GBS
MFLW8	5	11,799,889	7.47	0.03	17,913	GRMZM2G024973	GBS
MFLW8	5	11,799,892	8.05	0.02	17,916	GRMZM2G024973	GBS
MFLW8	5	11,799,924	7.47	0.03	17,948	GRMZM2G024973	GBS
EARHT	5	84,088,080	7.29	0.06	1,523	GRMZM2G017802	GBS
EARNB	5	210,476,290	7.52	0.07	0	GRMZM2G154332	GBS
ASI8	6	164,373,431	7.41	0.03	0	GRMZM2G038284	GBS
HL	7	103,544,585	7.95	0.08	3,184	GRMZM5G838060	GBS
HL	7	156,012,926	7.42	0.1	1,378	GRMZM2G019916	GBS
ROWNB	8	171,542,858	7.52	0.48	0	GRMZM2G591605	GBS
FFLW8	8	123,510,268	10.64	0.3	0	GRMZM2G479987	600K
MFLW8	8	123,509,765	10.99	0.3	0	GRMZM2G479987	600K
LFNB	8	123,510,268	12.5	0.3	0	GRMZM2G479987	600K
LFNBb	8	123,506,087	12.5	0.32	2,817	GRMZM2G179274	GBS
FFLW8	9	144,880,649	7.51	0.03	0	GRMZM2G021573	GBS
MFLW8	9	21,813,228	7.1	0.04	2,457	GRMZM2G144421	GBS
MFLW8	9	144,880,649	9.31	0.03	0	GRMZM2G021573	GBS
MFLW8	9	144,881,290	8.25	0.03	0	GRMZM2G021573	GBS
MFLW8	10	57,554,765	7.51	0.02	12,435	GRMZM2G009681	GBS

Chapter 4

Contribution of large InDels in addition to SNPs to identify new QTLs with additive and dominant effects and predict hybrid values in maize

Clément Mabire, Romane Guilbaud, Alain Charcosset, Laurence Moreau, Cyril Bauland, Delphine Madur, Valérie Combes, Stéphane Nicolas, Julie B. Fiévet

Key words: hybrid value, InDels, SNP, diallel, maize, Present/Absent Variants

Introduction

Since more than a century, hybrid vigor is largely exploited by breeders to produce high-performing varieties in maize (Zirkle, 1952). Shull (1908) and East (1908) studied inbreeding depression and hybrid vigor in maize and first proposed the concept of hybrid varieties. Shull (1914) defined the superiority of the hybrids in comparison to their parents as heterosis. There are three non-exclusive genetic causes for hybrid vigor: *complementary dominance*, which results in masking of the effects of deleterious alleles at several loci (Davenport, 1908; Bruce, 1910; Jones, 1917); *overdominance*, which postulates an advantage *per se* of heterozygosity at one locus over both homozygotes (Hull, 1946) and *epistasis*, where interactions between alleles at different loci in the hybrid have favorable effects (Bateson, 1908; Dooner et al., 1991).

QTL detection has been an important step to understand the genetic basis of quantitative traits and hybrid value. Previous studies on bi-parental linkage mapping populations concluded that apparent overdominance, *i.e.* QTLs closely linked in repulsion phase, was the major cause for heterosis in maize (Stuber et al., 1992; Lu et al., 2003; Frascaroli et al., 2007, 2009; Schön et al., 2010) and that most of these QTLs were located in the centromeric region where recombination is limited. The question of heterosis and hybrid performance was also investigated in populations with a larger genetic diversity. In Larièpe et al. (2012), the analysis of an extended NCIII design concluded that QTLs detected for grain yield displayed apparent overdominance and that QTLs for plant height and silking date displayed additive and dominance effects. They also showed a high correlation between the heterozygosity of the hybrids and heterosis, confirming the role of dominance in the occurrence of heterosis.

For many years, the challenge for breeders has been to predict the phenotypic values of untested hybrids. Predictions of hybrid performances were initially attempted based on molecular marker heterozygosity. Charcosset et al. (1991) and Bernardo (1992) showed that many conditions were necessary for effective prediction of hybrid performances with this approach: (i) a large number of loci displaying a dominance effect on hybrid performances, (ii) the existence of linkage disequilibrium between these QTLs and markers used to estimate heterozygosity and (iii) a high heritability of the trait under study. Condition (ii) was further refined by Charcosset and Essioux (1994) who investigated the effect of organization of diversity into heterotic groups and showed that prediction of between group hybrids by this approach was very unlikely to be efficient. The importance of such hybrids prompted the search for alternative approaches. Bernardo (1994, 1996) first proposed to model specific and general combining abilities by calculating best linear unbiased predictions (BLUP) including the information on molecular markers or maize pedigree. Several methods were then developed to predict hybrid performances based on the genetic covariance between individuals estimated with molecular markers data (VanRaden, 2008; Maenhout et al., 2010). Using a single genomic relationship matrix (VanRaden, 2008), the dominance genetic variance was partially integrated (Varona et al., 2018) but it was not possible to distinguish between the additive and dominance genetic effects. Su et al. (2012) proposed a method to construct a dominance relationship matrix calculated with SNPs markers and used it in models integrating simultaneously additive and dominance genomic relationship matrices. This approach was then used by integrating a dominance deviation in the model in addition to breeding value (Vitezica et al., 2013) or by decomposing the genetic variation of a trait into additive and dominant genotypic effects for traits in pigs (Vitezica et al., 2016). In maize, the inclusion of dominance genetic effect in addition to additive genetic effect calculated with SNPs significantly improved predictive abilities (Technow et al., 2012; dos Santos et al., 2016).

In addition to SNPs polymorphism, other molecular variations between two genomes can be involved in heterosis (Springer and Stupar, 2007). In addition to SNPs, maize displays a high amount of Structural Variations (SVs) between individuals, some of them corresponding to insertions or deletions (InDels) of genomic sequences. InDels correspond to the presence or absence of genomic sequences from one to several thousands of base pairs and thus could change gene dosage between individuals. These InDel sequences could be present or totally absent from the genome leading to presence-absence variations (PAVs). Fu and Dooner (2002) first suggested that complementation of genes present in one parent and absent in the other, could explain hybrid vigor. This hypothesis has been put forward by several studies (Beló et al., 2010; Lai et al., 2010; Swanson-Wagner et al., 2010; Chia et al., 2012; Sun et al., 2018).

In their pioneer study, Fu and Dooner (2002) compared the sequence of two maize inbred lines for homologous regions around the gene *bz1*. They showed that these two lines did not share the same gene content. Brunner et al., (2005) first studied the extent of indels in the whole genome and identified that 30% of the genome (corresponding to a total of 2.8Mbp) was not shared between the two maize lines B73 and Mo17. Springer et al., (2009) found 1,783 segments present in B73 but absent in Mo17. More recently, thousands of InDels were identified in maize using comparative genomic hybridization (CGH) array (Beló et al., 2010; Lai et al., 2010; Swanson-Wagner et al., 2010; Liu et al., 2015) or by sequencing inbred lines (Lai et al., 2010; Hirsch et al., 2016; Jiao et al., 2017; Darracq et al., 2018). However, high

throughput genotyping of InDels is more complicated than SNP genotyping, thus limiting the genome wide studies of InDels involvement in adaptation, trait variations and hybrid performance. To address this issue, Chia et al. (2012) resequenced 103 maize inbred lines. They discovered 55 million of SNPs as well as genomic regions in different copy number by identifying difference in reads counts. Performing genome-wide association study (GWAS) with SNPs on five traits (leaf angle, leaf length, leaf width and resistance to southern and northern leaf blight), they found an overrepresentation of associated SNPs in high LD with reads in different copy number, suggesting an effect of the number of gene copies on the variation of the traits of interest. In the same way, Lu et al. (2015) sequenced 14,129 inbred lines using GBS and identified 1.1 million mapped tags classified as InDels. To assess the role of these InDels in controlling phenotypic variations, they conducted GWAS on 2.661 inbred lines and four traits (days to silking, days to anthesis, plant height, and ear height) with SNPs. They found an enrichment of SNPs in high LD with InDels among associated SNPs for these four traits. Lyra et al., (2018) first proposed to include InDels in GS prediction models but only addressed their additive contribution. They did not observe a significant gain for predictive ability except for plant height. To our knowledge, no study tested the effects of both InDels and SNPs on trait variations by GWAS in a hybrid experimental design in maize with the decomposition their contribution into additive and dominance effects.

In this study, we used a hybrid maize panel to first, evaluate the fraction of additive and dominance genetic effects in the genetic variation of traits. Second, we tested the contribution of InDels in addition to SNPs to hybrid traits variation using GWAS models addressing additive and dominance effects for putative QTLs and the polygenic genetic background effect. Third, we compared the predictive ability of the additive and dominance GBLUP models with the genomic additive and dominance relationship matrix calculated with SNPs or InDels genotyping. The hybrid panel we developed consisted in 287 hybrids obtained from an incomplete diallel between 210 lines from two complementary heterotic groups (Dent and Flint) often used for hybrid breeding in northern Europe. Lines were genotyped with 61,492 large InDels from a new InDels genotyping array and with two SNPs genotyping arrays (Ganal et al., 2011; Unterseer et al., 2014) and by GBS (Elshire et al., 2011).

Materials and methods

Plant material

We selected 210 inbred lines adapted to grain production in French climatic conditions from a collection of 375 inbred lines previously used in association mapping experiments (Camus-Kulandaivelu, 2006; Ducrocq et al., 2008; Manicacci et al., 2009; Bouchet et al., 2013). These inbred lines were common to those used by Larièpe et al. (2017). Among them, 101 and 109 were from the dent and flint groups, respectively. These maize inbred lines were crossed following an incomplete diallel design to produce a total of 284 hybrids among which 146 were intra-heterotic group (75 dent x dent hybrids and 71 flint x flint hybrids) and 141 were inter-heterotic group hybrids (dent x flint).

Genotyping

The 210 parental inbred lines were genotyped with 49,036 markers from the 50K Illumina® (Ganal et al., 2011) and 616,201 markers from the 600K Affymetrix® Axiom® (Unterseer et al., 2014) SNPs genotyping arrays and also with 955,691 markers from genotyping by sequencing technology (GBS) (Elshire et al., 2011). Genotyping from GBS was imputed with TASSEL (Glaubitz et al., 2014). We assembled all genotyping matrices and selected one marker by locus based on genotyping quality as described in chapter 3 leading at the end of this process to a matrix of 1,076,794 markers. This matrix was imputed with Beagle v3.3.2 (Browning and Browning, 2007, 2009).

The parental inbred lines were also genotyped with 662,772 probes from the InDels Affymetrix® Axiom® genotyping array (Chapter 2) allowing to target 105,927 InDels from 35 to 129.7 Kbp. The probes were designed at the breakpoints or within InDel sequences previously discovered by the comparison of four inbred lines. We selected 231,077 probes allowing the genotyping of 61,492 InDels polymorphic and biallelic for presence and absence (0=absence of the sequence, 2=presence of the sequence, 1=hemizygous) (Chapter 3).

All the markers of the study were mapped on the B73 reference genome v2 (Schnable et al., 2009). Since 60% of the InDels were not anchored on the physical map, we mapped 72% of these InDels thanks to LD with markers which positions were known on the genome (Chapter 3). In order to include them in GWAS, the 28% remaining InDels that were not anchored on the genome were assigned with a virtual position to the chromosome 11 that has no biological meaning.

The genotypes of the hybrids were reconstructed according to the genotypes of the parental inbred lines. The heterozygous loci in the parental lines were treated as missing values. We filtered out markers with minor allele frequency (MAF) < 5% and markers with less than five hybrids in at least one of the three genotypic classes. In the SNP set of markers, those that were in complete LD with another SNP were also discarded. We tested the hypothesis of panmixis on the hybrid panel for each marker. Note that only biallelic markers were selected and Off Target Variants (OTV) from the 600K genotyping array were

filtered out due to their third allele (absent). A marker-by-marker Chi-square test showed a deviation from panmixis (pvalue < 0.05) only for a very low number of markers (4 InDels and 474 SNPs). In the end, two sets of markers were defined: the 469,267 SNPs set and the 51,844 InDels set.

The 51,844 InDels dataset allowed to genotype 47,332 unique InDels because 2,474 InDels were genotyped with breakpoint (BP) probes and internal probes and 2,038 InDels were genotyped with both forward and reverse BP probes (see chapter 3). Among these 47,332 InDels, 7,157 (15%) were totally specific, meaning that these sequences are either present or totally absent from the genome.

The inbreeding of each hybrid was estimated as $h = \frac{number\ of\ homozygous\ markers}{total\ number\ of\ markers}$ (Xiang *and al*. 2016, Silió *and al*. 2013). Note that this value is equivalent to the Identity By State (IBS) proposed by Astle & Balding (2009). We calculated the *pvalues* for the Pearson correlation coefficient between inbreeding and traits with the R function *cor.test*.

Phenotypic evaluation

The hybrids were evaluated in five French locations in 2009 and 2010 jointly with the hybrid design analyzed in Larièpe et al. (2017). Due to the large range of expected flowering time of our hybrids (based on the parental *per se* values), the panel was divided into two sub-panels. The early-flowering sub-panel (EP) was evaluated in 2 northern France locations (Mons in 2009 & 2010 and Gif-sur-Yvette in 2009) whereas the late-flowering sub-panel (LP) was evaluated in 2 southern French locations (Lusignan in 2009 & 2010 and Satolas in 2010). Both sub-panels were also evaluated simultaneously for two years (2009 & 2010) in the INRA station of Saint Martin de Hinx that allows the evaluation of panels with a large range of flowering time. Each hybrid was evaluated on average in 6.4 different year x location combinations.

Each trial was made of 224 elementary two-row plots of 9.28 m². Planting density was settled according to the usual practice of each location. To avoid competition between hybrids showing different level of inbreeding, each trial was divided into 14 blocks. To assign the hybrids to the different blocks, we calculated the Roger's genetic distance between the parental inbred lines of each hybrid with 55 SSRs as described in Camus-Kulandaivelu, et al. (2006). Each block was made of 16 elementary two-row plots: 13 for hybrids with close Roger's genetic distances, 2 for hybrids repeated in another block and 1 for a hybrid belonging to the alternative sub-panel (late-flowering hybrids for the EP trial and, early-flowering hybrids for the LP trial, respectively). The positions of the plots within the blocks and of the blocks within the trials were randomized. Each genotype was repeated on average 1.2 times per trial.

In each trial, we measured plant height (PH in cm) as the average of five plants, the female flowering date (FFD in days after the 1st of January) as the day at which 50% of the plants exhibited silks, and the grain yield at 15% moisture (GY in qx/ha) calculated from the fresh weight (FW in kg) with the following formula: $GY = \frac{FW}{9.28} \times 100 \times \frac{(100-GM)}{(100-15)}$. The total kernel weight was measured in all environments except in Mons 2009.

Field data analysis

Three hybrids were eliminated due to a too low number of plants at harvest in all trials (below the median of the number of plants at harvest – 15 plants). Aberrant data were also identified by visual inspection of trait distribution and they were manually eliminated as they corresponded to typing errors (1.3% of the data).

In each trial, we analyzed field heterogeneity using the mixed model:

$$P_{ixyr} = \mu + \mathrm{G}_i + \mathrm{R}_y + C_x + \varepsilon_{ixyr}$$

where P_{ixyr} is the phenotype of the repetition *r* of the hybrid *i* evaluated in the row y and the column x of the trial. The random terms are assumed to be independant and distributed as: $\mathrm{G} \sim \mathcal{N}\left(0, I\sigma^2_{Z_i}\right)$, $\mathrm{R} \sim \mathcal{N}\left(0, I\sigma^2_{Z_y}\right)$, $\mathrm{C} \sim \mathcal{N}\left(0, I\sigma^2_{Z_x}\right)$, $\varepsilon \sim \mathcal{N}(0, I\sigma^2_{\varepsilon})$.

We used the BLUP of row and column effects to correct data for field heterogeneity as:

$$Y_{ir} = P_{ixyr} - Z_y\hat{\mathrm{R}} - Z_x\hat{\mathrm{C}}$$

Based on corrected data, we estimated the relative contribution of genotypic and environmental variances to the phenotypic variation. This analysis was carried out on the data from each trait in each of the 10 trials and on the data of all trials with Asreml-R using the following mixed- models:

$$Y_{ir} = \mu + \mathrm{G}_i + \varepsilon_{ir} \qquad \text{(one-trial)}$$

$$Y_{irl} = \mu + \tau_l + \mathrm{G}_i + \varepsilon_{irl} \qquad \text{(multi-trial)}$$

where Y_{irl} (or Y_{ir}) is the vector of corrected data of the repetition *r* of the genotype *i* evaluated in the trial l, μ is the intercept, *G* is the random effect of genotype, τ the fixed effect of the trial (only for the multi-trial analysis) and ε the error. The random terms are assumed to be independant and distributed as: $G \sim \mathcal{N}(0, I\sigma^2_G)$, $\varepsilon \sim \mathcal{N}(0, I\sigma^2_{\varepsilon})$.

The broad sense heritability of each trait in each trial was calculated with the following formula: $h^2 = \frac{\sigma^2_G}{\sigma^2_G + \frac{\sigma^2_{\varepsilon}}{N}}$ with *N* the average number of replicates of the hybrid genotypes in the trial.

Using corrected data, we calculated adjusted means for each trait with the R function *lsmeans* (from the "lsmeans" R package) with a trial effect as fixed effect (multi environments model) on (1) the data of the two trials of SMH 2009 (further referred to as environment SMH09), (2) the data of the two trials of SMH 2010 (environment SMH10), (3) the data of the 4 trials of SMH (environment SMH0910) and (4) the data from all the trial (environment GLOBAL).

Decomposition of the genetic variance

The genetic value of the hybrids was decomposed into additive effects (A_i) and dominance effects (D_i) according to the model proposed by Vitezica *et al.*, 2016:

Model 1: $Y_{gilr} = \mu + \tau_l + \gamma_g + \theta \times h_{i(g)} + A_{i(g)} + D_{i(g)} + (\tau A)_{li(g)} + (\tau D)_{li(g)} + \varepsilon_{gilr}$

with the fixed and random terms symbolized respectively with lowercase Greek letters and capital Latin letters; Y_{gilr} corresponds to the performance of the repetition *r* of the hybrid *i* of the genetic group *g* evaluated in the environment *l* ; μ is the intercept, τ_l is the fixed environment effect, γ_g is the fixed effect of the genetic group of the hybrid (DD, DF or FF), h is the inbreeding fixed effect calculated on the SNPs dataset (Xiang *and al*. 2016, Silió *and al*. 2013), θ the regression coefficient of h, $A_{i(g)}$ and $D_{i(g)}$ are the random additive and dominant effects, $(\tau A)_{li(g)}$ and $(\tau D)_{li(g)}$ are the random interaction effects between additive or dominance and the environment τ for hybrid *i* and the environment *l*, and ε_{gilr} is the environment specific error. The random terms are assumed to be independant and distributed as:

$A \sim \mathcal{N}(0, K_a\sigma_a^2)$, $D \sim \mathcal{N}(0, K_d\sigma_d^2)$, $\tau A \sim \mathcal{N}(0, I_l \otimes K_a\sigma_{\tau a}^2)$, $\tau D \sim \mathcal{N}(0, I_l \otimes K_d\sigma_{\tau d}^2)$ and $\varepsilon \sim \mathcal{N}(0, I\sigma_\varepsilon^2)$ with I_l the identity matrix of size *l*. According to Vitezica et al., 2016, the covariance matrix for additive effects K_a was constructed as $K_a = \frac{ZZ'}{\{tr[ZZ']\}/n}$; were *Z* contains the genotypes of the hybrids for each marker coded as {-1 ;0 ;1}. For SNP, -1 correspond to hybrids homozygotes for the B73 allele, 1 for hybrids that were homozygotes for the alternative allele, and 0 for the heterozygotes. For the InDels genotypes, -1 corresponds to the absence of the allele and 1 to the presence of the allele in two copies and 0 to the heterozygotes (one copy of the allele). The covariance matrix for dominance effects K_d was constructed as $K_d = \frac{WW'}{\{tr[WW']\}/n}$ where *W* is a matrix with 0 for homozygous hybrids and 1 for heterozygote.

We compared the effects of the inbreeding, the group of the hybrid and the interaction between genetic effects and the environment by comparing the complete model 1 with models 2 (without inbreeding effect), 3 (without group effect), 4 (without group and inbreeding effect) and 5 (without interaction with environment):

Model 2: $Y_{gilr} = \mu + \tau_l + \theta h_{gi} + A_{i(g)} + D_{i(g)} + (\tau A)_{li(g)} + (\tau D)_{li(g)} + \varepsilon_{gilr}$

Model 3: $Y_{gilr} = \mu + \tau_l + \gamma_g + A_{i(g)} + D_{i(g)} + (\tau A)_{li(g)} + (\tau D)_{li(g)} + \varepsilon_{gilr}$

Model 4: $Y_{gilr} = \mu + \tau_l + A_{i(g)} + D_{i(g)} + (\tau A)_{li(g)} + (\tau D)_{li(g)} + \varepsilon_{gilr}$

Model 5: $Y_{gilr} = \mu + \tau_l + \gamma_g + \theta h_{gi} + A_{i(g)} + D_{i(g)} + \varepsilon_{gilr}$

For both SNPs and InDels genotyping datasets, the genotypic variance was decomposed into the additive and the dominance components. Note that model 5 is similar to model 1 but without interaction effects with the environment.

Genome Wide Association Study

The analysis was made on the *LS-means* data (SMH09, SMH10, SMH0910 and GLOBAL) with the following model for each marker (m):

$$Y = 1\mu + Z_m\alpha_m + W_m\beta_m + \theta \times h + A + D + \varepsilon$$

where A and D are the vectors of random additive and dominant effects distributed as $A \sim \mathcal{N}(0, K_a\sigma_a^2)$, $D \sim \mathcal{N}(0, K_d\sigma_d^2)$; *Y* the *ls-means* vector, α is the additive fixed marker effect and β is the dominance fixed effect at the marker m, Z_m and W_m are the vectors of the hybrid genotypes (the column m of the Z and W matrices defined above) of the hybrids encoded as {-1 ; 0 ; 1} and {0 ; 1 ; 0} and h is the inbreeding fixed effect calculated on the SNPs dataset (Xiang *and al*. 2016, Silió *and al*. 2013), θ the regression coefficient of h, and ε the residual with $\varepsilon \sim \mathcal{N}(0, I\sigma_\varepsilon^2)$. According to Rincent et al. (2013) we excluded the chromosome of the tested marker to estimate the covariance matrices.

We performed GWAS on both InDels and SNPs genotyping data. In each case we calculated the covariance matrices for the Additive and Dominance effects, K_a and K_d, using the corresponding marker data.

QQplots and Manhattan plots were made for additivity and dominance effects with the R package "qqman". The –log(*pvalue*) threshold for declaring an association as significant was fixed to 5.

Associated markers were grouped into QTLs according to their linkage disequilibrium. We choose the LD-corrected measure proposed by Mangin et al. (2012) to tackle the problem of the relatedness between genotypes. We grouped into a same QTL significant adjacent markers with LD-corrected value, $r_K^2 > 0.7$. The kinship matrix used for the correction were an Identity By Descent matrix (Astle and Balding, 2009) calculated with both SNPs and InDels dataset.

Genomic prediction models

We compared InDels and SNPs for the prediction of the hybrid performance. We used a GBLUP model as a base model. Three models were compared: *GP1*, the additive standard GBLUP model; *GP2*, the additive-dominance GBLUP model (based on Vitezica et al., 2016); and *GP3*, an additive-dominance model integrating the vector of inbreeding as fixed effect. For each model, the kinship between individuals was calculated on InDels, SNPs or by considering jointly both types of genotypes (InDels+SNPs).

$$y = 1\mu + A + \varepsilon \qquad (GP1)$$

$$y = 1\mu + A + D + \varepsilon \qquad (GP2)$$

$$y = 1\mu + A + D + \theta \times h + \varepsilon \qquad (GP3)$$

with μ is the intercept, *A* and *D* the additivity and dominance random effects with $A \sim \mathcal{N}(0, K_a\sigma_a^2)$, $D \sim \mathcal{N}(0, K_d\sigma_d^2)$, h is the inbreeding fixed effect calculated on the SNPs dataset (Xiang *and al*. 2016, Silió *and al*. 2013), θ the regression coefficient of h, ε the vector of errors, with $\varepsilon \sim \mathcal{N}(0, I\sigma_e^2)$. The covariance matrices for the Additive and Dominance effects, K_a and K_d, were calculated as described above.

Prediction with the different models was evaluated with 100 cross-validations by randomly sampling 4/5 of the hybrids population as training set to predict 1/5 of population as the validation set. BLUPs were calculated according to parameters estimated with the training set and predictive abilities were calculated as the correlation between observed and estimated phenotypic values of the validation set. We compared accuracies of prediction according to the genotyping dataset used to estimate the covariance matrices K_a and K_d and the model of prediction.

Results

Statistical analysis of the phenotypic data

We estimated the relative contribution of genotypic and environmental variances to the phenotypic variation and calculated the broad sense heritability for each trait. The heritabilities varied across all trials (Table S4.1). Except for some particular trait – trial combinations with very low heritabilities (0.52 for grain yield (GY) in LS10 and 0.45 for plant height (PH) in SMHL09), the heritabilities were high for each trait (between 0.69 and 0.95 for female flowering date (FFD), 0.6 and 0.92 for PH, and between 0.67 and 0.91 for GY). On the whole data set (GLOBAL), the heritabilities were higher, ranging from 0.88 for GY to 0.95 for FFD.

For further analyses, we focused on 4 year x trial combinations that will be further referred to as the four environments: SMH09 (early and late flowering panels), SMH10 (early and late flowering panels), SMH0910 (the 4 SMH trials) and GLOBAL (all of the 10 trials). The relative contribution of genotypic and environmental variances to the phenotypic variation and the broad sense heritability were calculated for each trait (Table 4.1).

Table 4.1: Variance decomposition and heritability. The heritability (h^2) of each trait was calculated from the genetic variance component (σ^2G) and the residual variance ($\sigma^2\varepsilon$).

		FFD	PH	GY
GLOBAL	**σ^2G**	8.77	247.54	134.86
	$\sigma^2\varepsilon$	2.56	143.27	115.88
	h^2	0.95	0.91	0.87
SMH09	**σ^2G**	5.06	207.45	284.38
	$\sigma^2\varepsilon$	1.47	133.64	49.01
	h^2	0.82	0.67	0.88
SMH10	**σ^2G**	3.82	297.56	177.31
	$\sigma^2\varepsilon$	1.78	97.68	64.77
	h^2	0.74	0.80	0.78
SMH0910	**σ^2G**	6.08	243.14	201.36
	$\sigma^2\varepsilon$	1.09	126.94	88.60
	h^2	0.93	0.83	0.85

Covariance matrices of the genetic effects and inbreeding

To estimate the correlation between the two-genotyping datasets (SNPs and InDels), we calculated the covariance matrices for additive (K_a) and dominance effects (K_d) and the inbreeding of the hybrids.

K_a and K_d matrices showed a high correlation between SNP and InDels (r^2 =0.95). Even if the results showed that both types of genotypic information were very close, these little differences could point out

some particular features of each genotyping datasets in further statistical analyses (variance decomposition and GWAS models).

The inbreeding of each hybrid, h, calculated with the SNPs ranged from 0.54 to 0.82 (mean = 0.59) while the inbreeding calculated with the InDels ranged from 0.42 to 0.81 (mean = 0.57). Thus, the inbreeding of each hybrid calculated with the SNPs tended to be slightly higher than the inbreeding calculated with the InDels. In addition, the inbreeding based on SNPs and InDels were correlated (r^2=0.91) (Table S4.2).

We studied the effect of inbreeding on each trait by calculating the correlation between the phenotype and the inbreeding vectors calculated with InDels only or with both genotyping datasets (SNPs + InDels) (Table S4.3). We did not analyze the correlation between the inbreeding vectors calculated on the SNPs dataset since the correlation with the SNPs+InDels genotyping datasets was very high (0.999, Table S4.2). Whether the rate of inbreeding was calculated with the SNPs genotyping set or the InDels genotyping set, the correlations between inbreeding and the trait were not significant for FFD but significant and negatives for PH and GY (Table S4.3). Whatever the trait or the environment under study, the correlations between inbreeding estimated with the SNPs + InDels genotyping set and phenotypic values were higher than those with the Indels genotyping set.

Regarding the level of inbreeding within each group of hybrids, whether the inbreeding was calculated on the SNPs or the InDels genotyping matrix, the DF inter-group pairs of lines had a lower average inbreeding value than the intra-groups pairs DD and FF (Figure S4.1). While DD and FF had a similar inbreeding rate, the DD hybrids displayed a higher GY compared to FF (Figure S4.1).

Variance decomposition

We decomposed the genotypic variance of phenotype to estimate the effect of different parameters. In the following, the variance decomposition was performed for the GLOBAL environment.

Table 4.2: Significance level of Wald tests (*p-values*) on the SNP dataset for the GLOBAL phenotyping dataset. The fixed effect tested were the environment (τ), the group (ϒ) of the hybrid and the inbreeding (θ)

		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
		$\tau + \gamma + \theta$	$\tau + \theta$	$\tau + \gamma$	τ	$\tau + \gamma + \theta$ *without interactions*
FFD	*pvalue* τ	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.20×10^{-16}	2.00×10^{-16}
	pvalue ϒ	0.77		0.76		0.72
	pvalue ϑ	0.88	0.69			0.95
PH	*pvalue* τ	2.20×10^{-16}	2.20×10^{-16}	2.20×10^{-16}	2.20×10^{-16}	2.20×10^{-16}
	pvalue ϒ	4.00×10^{-3}		0.01		0.01
	pvalue ϑ	0.01	2.96×10^{-5}			3.00×10^{-3}
GY	*pvalue* τ	0.05	0.05	0.05	0.05	2.20×10^{-16}
	pvalue ϒ	0.1		0.1		2.10×10^{-2}
	pvalue ϑ	1.67×10^{-7}	3.30×10^{-8}			1.12×10^{-8}

The *p-values* of the statistical tests for the fixed effects were very close for SNPs and InDels (Table 4.2 and Table S4.4). The environment effect, τ, was significant for each model for Female Flowering Date (FFD) and Plant Height (PH) whereas it was only significant for model 5 for Grain Yield (GY). The effect of the group of the hybrid, γ, was significant only for PH and GY only with the model 5. The differences for GY observed between the groups (a higher GY for DD than FF hybrids, Figure S4.1) were not confirmed by the statistical modeling as far as a GxE effect was declared in the model (models 1 and 3).The inbreeding effect was significant and negative for PH and GY but not significant for FFD.

The variances of the random effects were very close between SNPs and InDels (Table 4.3 and Table S4.5). The variance of additivity was higher than the variance of dominance for all the traits and all the models (from 3 times to 6 times more in model 1). For all traits, the interaction effect additivity x environment ($a \times \tau$) was significant and its variance was of the same order of magnitude as the variance of dominance: 1.36 ($\sigma^2_{a\times\tau}$) vs 2.14 (σ^2_d) for FFD, 52.07 vs 52.04 for PH and 27.99 vs 39.83 for GY with the model 1. The dominance x environment ($d \times \tau$) effect was higher than dominance effect only for GY and its variance was of the same order of magnitude as the variance of additive effect: 98.18 (σ^2_a) vs 73.37 ($\sigma^2_{d\times\tau}$). For the model without the interaction effects (model 5), the variance of additivity was slightly higher whereas the variance of dominance was slightly lower than in the complete model (model 1). For GY, the introduction of inbreeding decreased the part of variance for dominance (model 1 vs models 3 and 4) whereas the variance of additivity slightly increased.

Table 4.3 – Genotypic Variance decomposition based on the SNPs genotyping dataset. Variances (and their standard error) were indicated for each model.

		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
		$\tau + \gamma + \theta$	$\tau + \theta$	$\tau + \gamma$	τ	$\tau + \gamma + \theta$ *without interactions*
FFD	σ^2_a	10.06 (1.43)	9.96 (1.41)	10.09 (1.43)	10.01 (1.41)	10.56 (1.45)
	σ^2_d	2.14 (0.46)	2.13 (0.45)	2.11 (0.45)	2.1 (0.45)	1.78 (0.45)
	$\sigma^2_{a\times E}$	1.36 (0.24)	1.36 (0.24)	1.36 (0.24)	1.36 (0.24)	
	$\sigma^2_{d\times E}$	0.30 (0.36)	0.30 (0.36)	0.3 (0.36)	0.30 (0.36)	
	σ^2_E	1.44 (0.13)	1.44 (0.13)	1.44 (0.13)	1.44 (0.13)	2.54 (0.1)
	Likelihood	-1913.49	-1912.6	-1915.18	-1914.2	-2008.74
PH	σ^2_a	309.64 (45.5)	304.54 (45.13)	318.08 (47.24)	301.62 (46.58)	316.93 (46.46)
	σ^2_d	52.04 (16.14)	53.54 (16.18)	55.11 (16.57)	69.85 (17.96)	50.04 (16.87)
	$\sigma^2_{a\times E}$	52.07 (14.07)	52.17 (14.06)	51.47 (14.05)	51.45 (14.00)	
	$\sigma^2_{d\times E}$	3.16 (24.09)	2.29 (24.12)	4.94 (24.04)	3.77 (24.09)	
	σ^2_E	109.81 (9.42)	110.21 (9.46)	109.13 (9.37)	109.47 (9.41)	142.42 (5.32)
	Likelihood	-5360.01	-5363.19	-5367.39	-5374.56	-5388.44
GY	σ^2_a	98.18 (19.37)	100.53 (19.33)	93.68 (20.76)	94.22 (20.58)	102.97 (19.25)
	σ^2_d	39.83 (11.70)	38.88 (11.52)	56.90 (13.59)	59.31 (13.53)	35.74 (11.75)
	$\sigma^2_{a\times E}$	27.99 (10.80)	27.91 (10.8)	27.55 (10.76)	27.50 510.74)	
	$\sigma^2_{d\times E}$	73.37 (16.63)	73.65 (16.62)	74.52 (16.52)	74.61 (16.48)	
	σ^2_E	55.83 (5.24)	55.76 (5.23)	55.16 (5.16)	54.98 (5.14)	115.68 (4.33)
	Likelihood	-5057.355	-5060.487	-5073.095	-5077.043	-5128.776

For all these models, we compared the dominance/additivity ratio of variances. This ratio was higher from GY (0.35 to 0.63 depending on the model) than for FFD (0.17 to 0.21) or for PH (0.16 to 0.23). For GY this $\frac{\sigma_d^2}{\sigma_a^2}$ ratio increased from 0.4 to 0.6 when the inbreeding effect was removed from the model. The group effect removal did not lead to difference in the $\frac{\sigma_d^2}{\sigma_a^2}$ ratio. Estimation of variances with the InDels genotyping dataset led to similar results (Table S4.5).

Genome Wide Association Studies

Performing GWAS, we found 23 InDels and 362 SNPs markers associated with phenotypic variations: 0 InDels and 69 SNPs markers for FFD, 9 InDels and 131 SNPs markers for PH and 14 InDels and 162 SNPs markers for GY.

We detected a total of 38 unique QTLs for FFD with SNPs. No InDel marker was found significantly associated with the Female Flowering Date. Among these 38 QTLs, 22 displayed a significant additive effect and 16 with a dominance effect (Table 4.4). Among the QTLs with a significant additive effect, 8 out of 22 (36%) have a negative effect. Among the QTLs with a significant dominance effect, 5 out of 16 (31%) have a negative effect. This indicated that the heterozygotes for these QTLs loci have an earlier flowering date than the average of homozygotes. No QTLs was found with both an additive and a dominance effect. Among the 38 QTLs very few were detected in common in several environments (4, 5 and 1 were detected respectively in two, three and four environment datasets, Table 4.4) and they were mainly QTLs with an additive effect suggesting that QTLs with dominance effect were more susceptible to GxE effects.

We detected a total of 78 unique QTLs for Plant Height: 70 QTLs with the SNPs dataset, 7 QTLs with the InDels genotyping dataset (Table S4.6) and 1 with both SNPs and InDels (Table 4.4). Among these 79 QTLs, 36 displayed an additive effect and 43 displayed a significant dominance effect. We did not detect any QTL with both additive and dominance effects. This equilibrium between the numbers of QTLs detected with an additive or with a dominance effect was found for both SNPs and InDels. Half of QTL with additive and dominant effect for PH displayed a positive effect. Most of the QTLs were detected in only one environment and 2 QTLs were common to two environments (against 9, 2 and 1 in respectively 2, 3 and 4 environments for QTLs with an additive effect).

A total of 94 unique QTLs were detected for Grain Yield: 81 QTLs with the SNPs dataset, 10 QTLs with the InDels genotyping dataset (Table S4.6) and 3 with both SNPs and InDels (Table 4.4). Among these 94 QTLs, 20 displayed an additive effect and 74 displayed a significant dominance effect. These dominant QTLs had mostly positive effects (92%) while the additive QTLs had mostly negative effects (85%). Only 14 QTLs out of the 94 were detected in more than 1 environment. With InDels, we found only two QTLs for GY that were detected in two environments: one QTL on chromosome 6 at 14Mbp with an additive effect (Figure S4.2) and one QTL on chromosome 4 at 37Mbp with a dominant effect.

The lists of the 5 highly associated SNPs markers with each trait were detailed in tables S4.7.

Table 4.4: Summary of the QTLs found for Female Flowering Date (FFD), Plant Height (PH) and Grain Yield (GY) in the four studied environments SMH09, SMH10 SMH0910 and Global. The QTLs can be detected with only SNPs, only InDels or with both (SNP and InDels). The last five columns indicated if the QTLs have been detected in 1, 2, 3 or 4 environments.

Trait	Markers	Type of effect	SMH09	SMH10	SMH0910	Global	specific to 1 environment	common to 2 environments	common to 3 environments	common to 4 environments	Total number of unique QTLs
FFD	**SNPs**	**A**	7	10	9	8	15	2	5	0	22
		D	1	8	5	7	13	2	0	1	16
	InDels	**A**	0	0	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0	0	0
	SNP and InDels	**A**	0	0	0	0	0	0	0	0	0
		D	0	0	0	0	0	0	0	0	0
	Total	**A**	7	10	9	8	15	2	5	0	22
		D	1	8	5	7	13	2	0	1	16
PH	**SNPs**	**A**	12	9	13	13	21	8	2	1	32
		D	9	16	3	20	34	3	1	0	38
	InDels	**A**	2	0	1	1	2	1	0	0	3
		D	0	2	0	2	4	0	0	0	4
	SNP and InDels	**A**	0	0	0	0	0	0	0	0	0
		D	0	1	0	0	1	0	0	0	1
	Total	**A**	14	9	14	14	23	9	2	1	35
		D	9	19	3	22	39	3	1	0	42
GY	**SNPs**	**A**	9	3	2	6	12	4	0	0	16
		D	21	9	9	33	57	6	2	0	65
	InDels	**A**	2	0	1	0	2	1	0	0	3
		D	2	0	0	5	7	0	0	0	7
	SNP and InDels	**A**	1	0	0	0	1	0	0	0	1
		D	0	1	1	1	1	1	0	0	2
	Total	**A**	12	3	3	6	15	5	0	0	20
		D	23	10	10	39	65	7	2	0	74

Predicting hybrids phenotypic values using InDels or SNPs genotyping set

We estimated the additive (*Ka*) and dominance (*Kd*) relationship matrices with SNPs, InDels or both of them to evaluate if taking into account InDels genotyping, in addition to SNPs, could improve the predictive abilities (PA). We then compared the PA according to the relationship matrices used. We also evaluated the added value of integrating dominance in the prediction models in addition to the additive effect.

According to the model used and whatever the genotyping used, we found PA between 0.714 and 0.822 for FFD (Figure PGFF), between 0.583 and 0.768 for PH (Figure PGPH) and between 0.588 and 0.768 for GY (Figure 4.1). The PA estimated with InDels, SNPs or both genotyping datasets were very close for GY and FFD whatever the model used (Figure 4.1 and S4.3). For PH, we observed a slight decrease of PA estimated with InDels in comparison to SNPs (Figure S4.4).

Integrating dominance into the model increased PA from 0.06% to 1.6% and from 1.5% to 5.6% for PH and GY respectively, according to the environment or the marker type (comparison of models GP2 and GP1 in Figure S4.4 and 4.1). For FFD, PA were similar whatever the models and the genotyping datasets, with a slight advantage to the additive model GP1 (Figure S4.3). Integrating inbreeding slightly improved the PA for GY, especially when inbreeding was estimated with InDels (model GP3 vs model GP2 in Figure 4.1). For FFD and PH, we found little difference by adding inbreeding in the model except for SMH10 and GLOB for which we observed a slight increase for PH (~0.4%) (model GP2 vs model GP3 in Figure PGFF and PGPH).

We observed an increase of PA for the GLOBAL environment in comparison to the three other environments for all traits whatever the model used. The highest increase was observed for FFD (Figure S4.3). The PA were higher for SMH0910 than for SMH09 and SMH10 for FFD and PH. For GY the higher PA was for the SMH09 environment (in comparison to SMH10 and SMH0910).

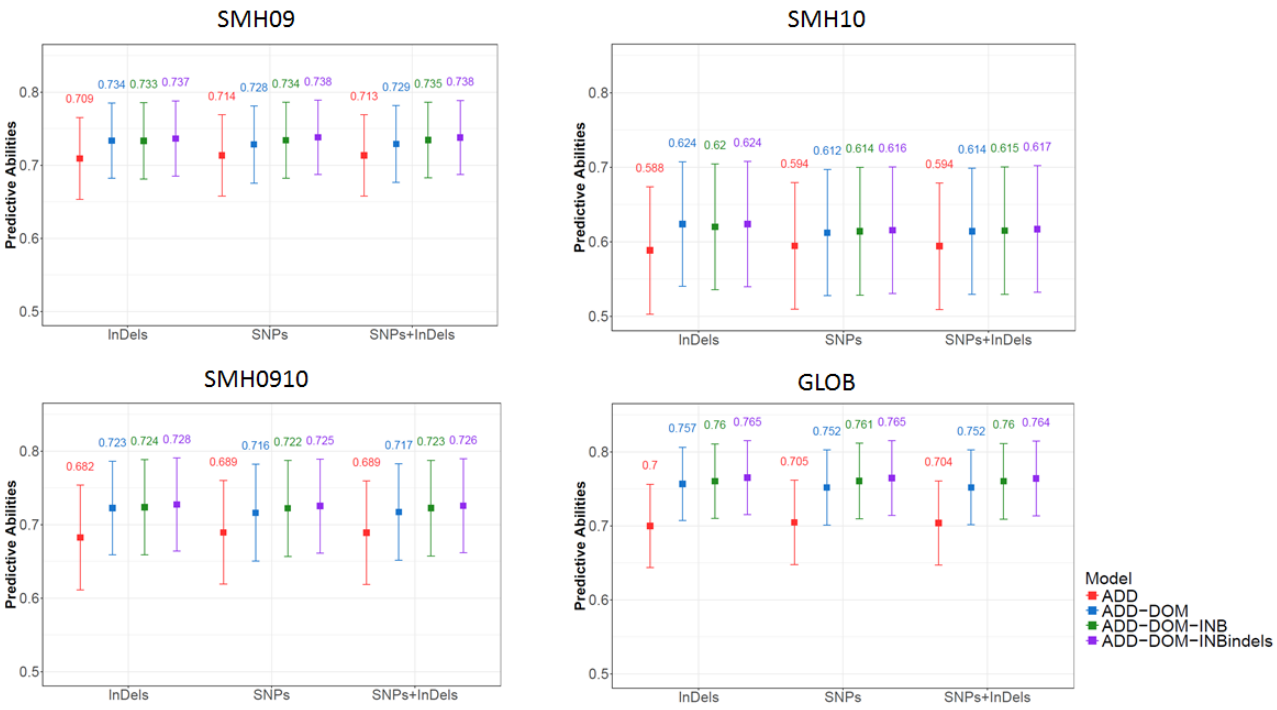


Figure 4.1: Accuracy of prediction for GY. We used four different phenotypes calculated by adjusted means from different environments. Three different models were compared additivity (model 1-red), additivity and dominance (model 2-blue), additivity, dominance and inbreeding (model 3 – green) with inbreeding estimated with SNPs and InDels (green) or only with InDels (purple). We tested each model only with InDels dataset, only with SNPs dataset and with both.

Discussion

Effects of hybrid group and inbreeding in variance decomposition and hybrid value predictions

We compared five different models to evaluate the relative part of additivity and dominance and their interactions with the environment in the genetic variance. The main differences between these models stand in the fixed effect tested: the inbreeding, the genetic group of the hybrids and the environment. Whatever the marker types used (SNPs or InDels dataset), we found very similar results for variance decomposition.

The effect of the hybrid group was significant for PH but did not influence the FFD of hybrids and GY. This result is surprising since we used hybrids from dent and flint inbred lines and these two genetic groups are complementary: flint group contributes to early vigor and cold tolerance and dent group contributes to high productivity of the hybrids (Unterseer et al., 2016). Larièpe et al. (2017) showed that the population structure has a significant effect on specific and general combining abilities (SCA and GCA respectively) for PH but also for GY. In our study, as the Figure S4.1 showed that the hybrid group affected the GY, mostly for DD and FF hybrids, we suggest that the effect of the hybrid group was included in the additivity and dominance genetic effect via the kinship matrices.

Inclusion of inbreeding in the models led to a decrease of the dominance variance and a slight increase of the additive variance, especially for GY. As expected, this inbreeding effect was significantly negative for GY and PH but it had no effect on FFD (Table S4.3). The effect of inbreeding is well known in maize (Barrett and Charlesworth, 1991; Charlesworth and Willis, 2009) but has not yet been considered in genomic prediction to our knowledge. In our study, integrating inbreeding conferred a slight increase in PA for GY, especially when inbreeding was estimated with InDels. Accordingly, Larièpe et al. (2017) found a decrease in SCA and an increase of GCA when inbreeding was included in the model of variance decomposition. In dairy cattle, the dominance deviation was slightly negatively correlated with traits suggesting that individuals with larger positive dominance deviation were less inbred (Aliloo et al., 2016). In pig, the effect of inbreeding is well known to have a strong effect on the phenotype, therefore, adding inbreeding in GBLUP model increases the PA (Xiang et al., 2016).

Dominance vs additivity

Integrating dominance in prediction models greatly improved predicting accuracy for PH and GY but not for FFD, whatever the marker type used. These results suggest significant dominance genetic effects for PH and GY and a preponderance of additive genetic effects for FFD. Accordingly, we found a higher part of variance explained by additive genetic effects with the model 1 for FFD (66%) than PH (59%) and GY (33%). The GWAS results gave the same results: the majority of QTLs for GY and PH were QTLs with a dominance effect whereas QTLs for FFD were mainly additive. In maize, increases of PA by integrating dominance in addition to additive genetic effect in prediction model were previously found

for GY (Technow et al., 2012), for PH, ear architecture, kernel weight (dos Santos *et al*., 2016) and for GY and PH (Lyra *et al*., 2018). According to the adjusted mean tested, we found from 1.5% to 5.6% of increase of predictive abilities for GY whereas Lyra et al. (2018) found an increase from 3% to 21%. We found a slight increase of PA for PH including dominance from 0.06% to 1.6%, whereas they found an increase from 1% to 4%. However, their PA was generally lower than ours for GY and PH.

Indels vs SNPs

To our knowledge, testing indels effect on hybrid performances by GWAS approach, decomposing additive and dominant genetic effect, was never performed before. The QTLs detected with InDels genotyping represented 6.9% of the total number of QTLs detected (whereas they represented 10% of total number of markers) indicating a slight depletion in QTLs detected with InDels compared SNPs. This result may be surprising regarding the complementation hypothesis (Fu and Dooner, 2002). However, this hypothesis is more relevant when the sequence of the InDel is present or totally absent from the genome (presence absence variations, PAVs) which represents 15% of InDels. 75% of InDels were partially specific and among these InDels, 9% displayed at least 1Kbp of specific sequence and could therefore carry genes which could be present or absent in parental inbred lines. As reported by Swanson-Wagner *et al* (2010), genes involved in CNVs and PAVs are often part of large gene families. Therefore, the loss of one copy of the gene probably only has a slight effect, due to the compensation of other genes of the family. Consequently, heterosis could be the result of restoring the full functionality of gene families within hybrids. We therefore hypothesize that InDels information should be considered at the level of gene families to improve the genomic predictions of the hybrid value. Lyra *et al*., (2018) showed that the number of CNVs in hybrids was correlated with plant height, suggesting a quantitative effect of number of sequence copies on phenotype, which supports this hypothesis.

We found little differences in PA when considering InDels, SNPs or both. This result can be related to the fact that the additive and dominance relationship matrices estimated with SNPs or InDels were highly correlated. Thus, both types of markers gave access to the same polygenic effect, and probably have similar evolutionary trajectories, as reported in chapter 3. Lyra *et al*., (2018) previously combined the additive relationship matrix estimated with hundreds of copy number variations (CNVs) with additivity and dominance relationship matrices estimated with SNPs and showed that CNVs did not increase the PA except for PH in one environment. These results could be due to the lower density of InDels compared to SNPs (50K vs 450K), which impact negatively the PA as previously reported in maize (Technow et al., 2012; Crossa et al., 2014; Zhang et al., 2015).

Specific QTLs, GxE interactions

Although our hybrids were evaluated in different environments, our field experiment was decomposed in earliest and latest hybrids evaluated in northern or southern locations respectively leading to a limited number of hybrids in each environment, except for one location (SMH). We performed GWAS analysis by calculating adjusted mean for SMH09 and SMH10 because all hybrids were evaluated at the same time only for this location, and also for SMH09 and SMH10 in a joint analysis (2 location x year combinations) and finally for the global design (10 environment x year combinations). The QTLs detected in several environments were in most cases, detected in the global design. Only 10, 15 and 14 QTLs were detected in at least two different environments (SMH09, SMH10, SMH0910 and GLOBAL) for FFD, PH and GY, respectively, indicating that most of QTLs detected were environment specific. These results suggest a significant genotype by environment (GxE) effect supported by the part of additive and dominance by environment interaction (AxE and DxE) found in variance decomposition for FFD (10% of the variance), PH (10% of the variance) and especially GY (35% of the variance). Moreover, we found a significant effect of the environment for all traits and all models. For GY, the results also pointed out a DxE variance twice higher than the dominance variance suggesting that QTLs for GY are more dependent on the environment than QTLs for PH or FFD.

We found a very similar part of AxE and DxE in variance decomposition for InDels and SNPs. This result was surprising since many studies reported an effect of InDels on stress tolerance (Chia *et al*., 2012; Saxena *et al*., 2014) and an enrichment in genes involved in stress tolerance within InDels (Swanson-Wagner et al., 2010; Hirsch et al., 2016; Jiao et al., 2017; Darracq et al., 2018). These studies suggested an important effect of the environment on InDels effect. Therefore, testing the effect of InDels in environments subject to stress conditions could increase the number of QTLs detected with InDels compared to SNPs.

Features of the QTLs detected

We analyzed the effects of QTLs identified by SNPs considering the sign of the effects. We first noticed that additive effects at QTls involved in flowering time and yield mostly displayed negative effects. This is consistent with (i) SNP alleles coding considering the B73 allele as the reference (0) and (ii) B73 being late flowering inbred line displaying high yield. In that context, hybrids carrying the B73 allele are expected to display late flowering, to be taller and with a higher GY than hybrids that carried the alternative allele. For QTLs with a significant dominant effect detected with SNPs, the heterozygosity mostly was generally associated to taller height and higher GY than homozygous genotypes. This confirms a general trend towards an advantage of heterozygosity for fitness traits.

For the QTLs for GY detected with InDels that showed a significant additive effect, hybrids that carried the present allele had a higher GY than those carrying the absent allele, suggesting that the presence of the sequence was advantageous for hybrids. Concerning the seven QTLs for GY detected with a significant dominant effect, hybrids that carried the hemizygous genotype had a higher GY than those carried the homozygous genotypes (absent or present) suggesting a positive dominant effect of InDels for these QTLs, consistent with what was observed for SNPs.

The association study made with the SNP genotyping revealed several QTLs found in other studies. Among the 38 QTLs detected for FFD with SNPs, 1 and 5 QTL were close to Vgt3 and Vgt1 respectively, two genes well known for their implication in flowering (Salvi et al., 2002, 2007; Castelletti et al., 2014). Two other flowering QTLs were previously identified by Laporte (2018). We detected 70 QTLs for PH and 81 for GY. Frascaroli *et al*. also identified 8 QTLs in the same region for PH and 8 others for GY. Two other QTLs were detected by Schön *et al*. (2010) for PH and 4 for GY.

Despite a lower proportion of QTL detected with InDels than SNPs, 7 and 10 QTLs for PH and GY respectively, were detected only with InDels genotyping dataset. Among these, two QTLs identified for PH only with InDels on chromosome 4 were previously identified by Frascaroli *et al*., (2009) as well as one QTL for the GY on the chromosome 6 previously identified with a panel of inbred lines by Millet *et al*., (2016) in hot conditions. A gene was identified by Gustafson *et al*. in (2018) inside an InDel at the same position than this QTL. This gene, Scmv1 is a resistance gene for Sugarcane mosaic virus (SCMV) and Maize dwarf mosaic virus (MDMV) that can cause great yield loss in maize. They showed that the absence of the sequence was linked with susceptibility to the virus. Our results are in accordance with this result as we found a positive effect of our QTL, meaning that the grain yield was higher when the sequence was present.

As recommended by Rio *et al*, (2019) on maize inbred lines, estimating the effects of the markers within genetic groups could improve the predictive ability for maize inbred lines. In a hybrid context, we could expect that a dominance effect depending on the group of the hybrid could also improve PA and also be helpful to find QTLs specifics to the hybrid's groups.

Supplementary figures

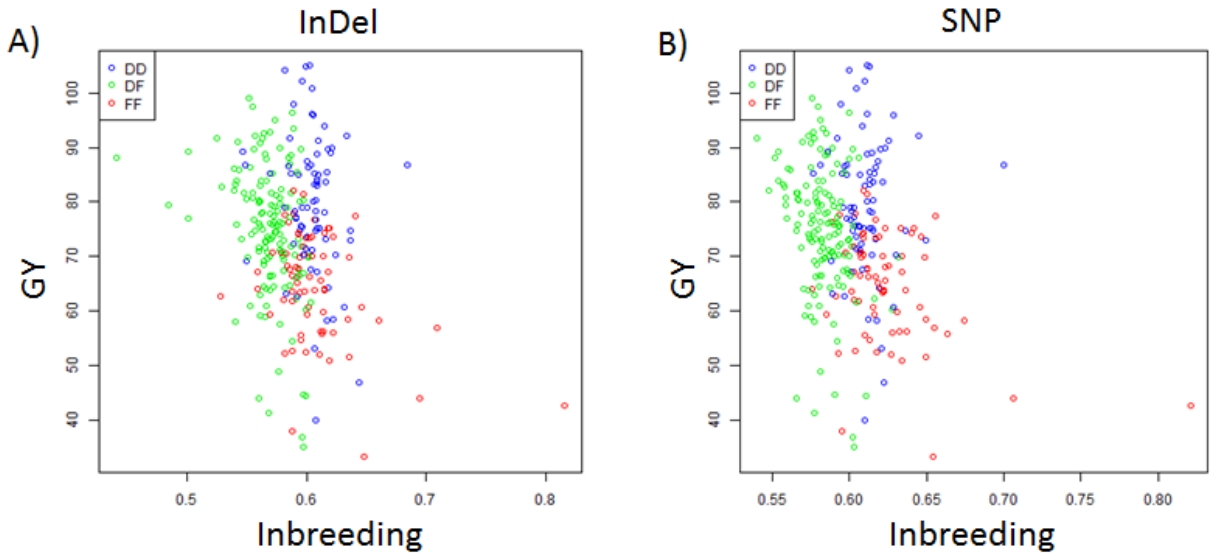


Figure S4.1: Correlation between the Inbreeding and GY (adjusted mean calculated for all environments GLOBAL) for each group calculated with InDels (A) or SNP and InDels (B). The blue empty dots correspond to intra-group hybrids DD, the red ones to intra-group FF hybrids and the green ones to inter-group DF hybrids.

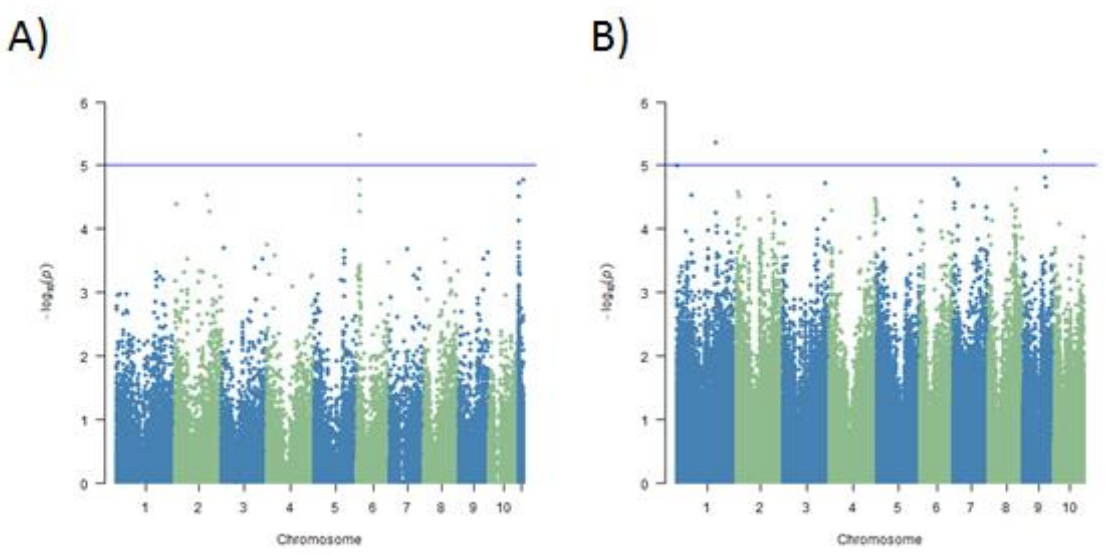


Figure S4.2: Manhattan plots of for GY for SMH0910 with A) the InDels dataset and B) the SNPs dataset. The blue line indicating the -log10(pvalue) threshold of 5.

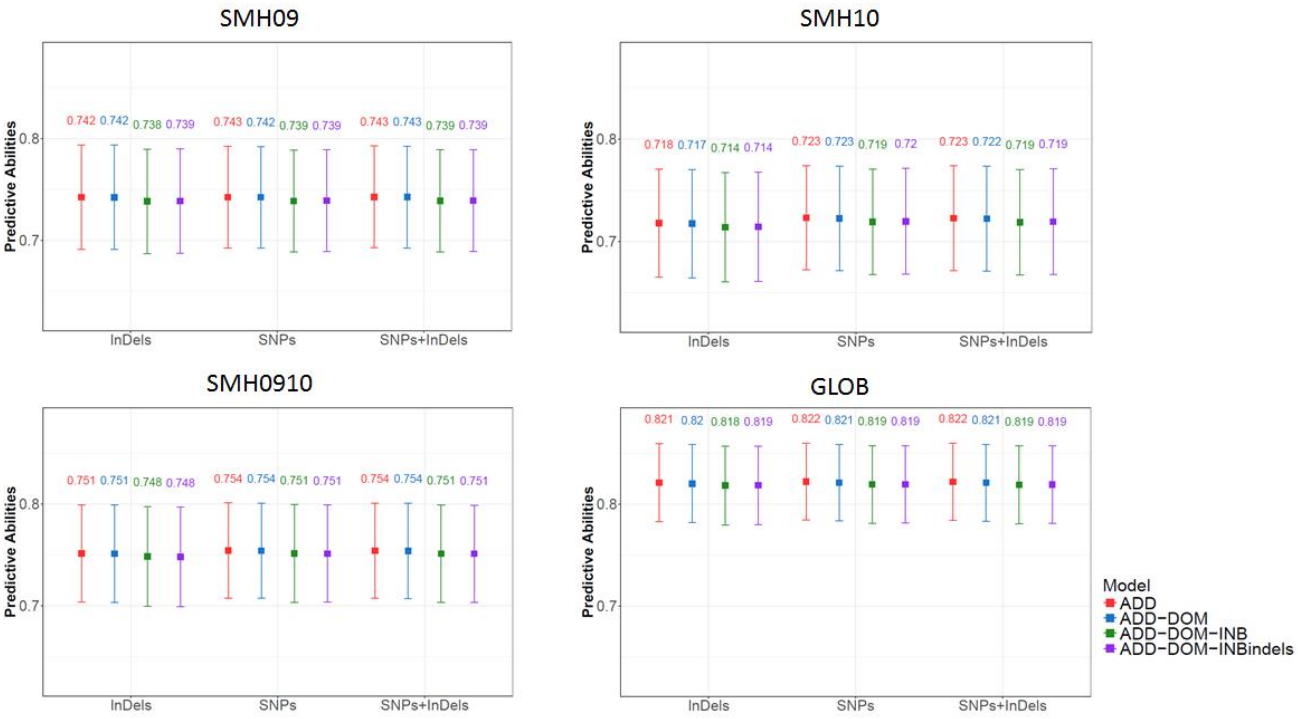


Figure S4.3: Accuracy of prediction for female flowering date. We used four different phenotypes calculated by adjusted means from different environments. Three different models were compared additivity (model 1-red), additivity and dominance (model 2-blue), additivity, dominance and inbreeding (model 3 – green) with inbreeding estimated with SNPs and InDels (green) or only with InDels (purple). We tested each model only with InDels dataset, only with SNPs dataset and with both.

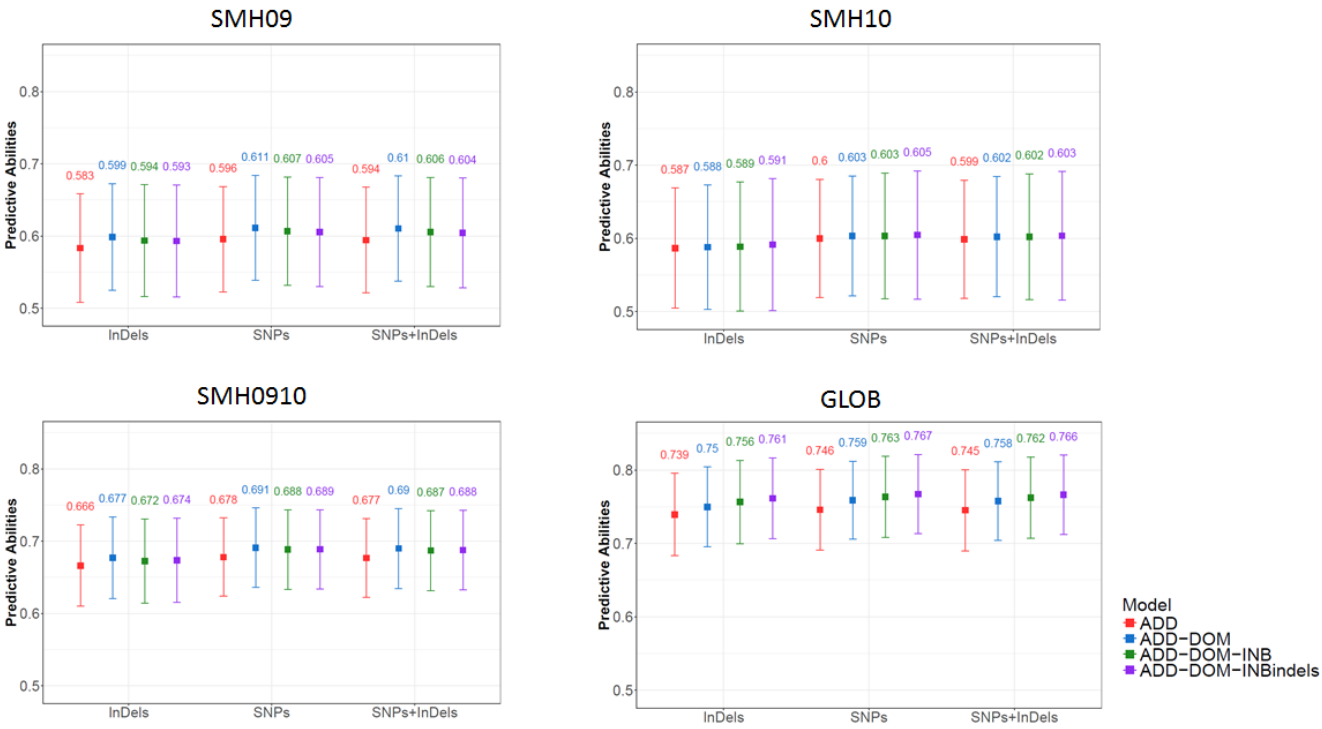


Figure S4.4: Accuracy of prediction for plant height. We used four different phenotypes calculated by adjusted means from different environments. Three different models were compared additivity (model 1-red), additivity and dominance (model 2-blue), additivity, dominance and inbreeding (model 3 – green) with inbreeding estimated with SNPs and InDels (green) or only with InDels (purple). We tested each model only with InDels dataset, only with SNPs dataset and with both.

Supplementary tables

Table S4.1: Variance decomposition and heritability. The heritability (h^2) of each trait was calculated with the formula: h^2=(σ_G^2)/(σ_G^2+(σ_ε^2)/N). This was estimated for all of the environments (2 years: 2009 and 2010; 5 Locations: LS: Lusignan, ML: Moulon, MONS: Mons, SAT: Satolas, SMHE: Saint Martin de Hinx Early panel, SMHL: Saint Martin de Hinx Late panel)

		FFD	PH	GY
Global	σ^2G	8.77	247.54	134.86
	$\sigma^2\varepsilon$	2.56	143.27	115.88
	h^2	0.95	0.91	0.88
LS09	σ^2G	8.21	195.38	129.83
	$\sigma^2\varepsilon$	4.17	145.72	51.68
	h^2	0.69	0.6	0.74
LS10	σ^2G	3.62	196.86	70.1
	$\sigma^2\varepsilon$	1.66	132.71	76.25
	h^2	0.72	0.63	0.52
ML09	σ^2G	11.71	306.49	232.35
	$\sigma^2\varepsilon$	0.86	170.04	96.32
	h^2	0.94	0.67	0.73
MONS09	σ^2G	11.19	221.83	154.53
	$\sigma^2\varepsilon$	3.06	85.3	60.65
	h^2	0.8	0.74	0.74
MONS10	σ^2G	6.47	495.35	168.33
	$\sigma^2\varepsilon$	1.46	48.38	85.32
	h^2	0.83	0.92	0.69
SAT09	σ^2G	7.58	286.64	115.42
	$\sigma^2\varepsilon$	0.8	34.93	46.13
	h^2	0.91	0.9	0.74
SMHE09	σ^2G	4.97	334.4	298.89
	$\sigma^2\varepsilon$	0.64	114.76	31.17
	h^2	0.9	0.76	0.91
SMHE10	σ^2G	5.52	356.57	235.02
	$\sigma^2\varepsilon$	0.6	57.68	29.52
	h^2	0.92	0.88	0.9
SMHL09	σ^2G	5.55	102.6	274.97
	$\sigma^2\varepsilon$	0.45	137.93	30.9
	h^2	0.93	0.45	0.91
SMHL10	σ^2G	3.28	245.13	134.08
	$\sigma^2\varepsilon$	1.19	111.63	74.89
	h^2	0.76	0.72	0.68

Table S4.2: Correlations between inbreeding vectors estimated with InDels, SNPs or both InDels and SNPs (All) genotyping.

	Inbreeding All	**Inbreeding InDels**	**Inbreeding SNPs**
Inbreeding All	1.000	0.908	0.999
Inbreeding indels	-	1.000	0.885
Inbreeding SNPs	-	-	1.000

Table S4.3: Correlations between inbreeding vectors estimated with InDels only or with both InDels and SNPs (All) for three traits, FFD, PH and GY in 4 different environments. The significance of the correlations was calculated with a p-values for the Pearson correlation coefficient between inbreeding and traits with: ns not significant, * $P \leq 0.01$, ** $P \leq 0.001$ and *** $P \leq 0.0001$.

		Environments			
		SMH09	**SMH10**	**SMH0910**	**GLOB**
FFD	Inbreeding InDels	-0.0335 ns	-0.0427 ns	-0.0422 ns	-0.0305 ns
	Inbreeding All	-0.0631 ns	-0.0564 ns	-0.0633 ns	-0.0755 ns
PH	Inbreeding InDels	-0.1793*	-0.1708*	-0.1853*	-0.2352***
	Inbreeding All	-0.2207**	-0.2038**	-0.2220**	-0.2486**
GY	Inbreeding InDels	-0.2051**	-0.2371***	-0.2420***	-0.2793***
	Inbreeding All	-0.2239**	-0.2787***	-0.2766***	-0.2961***

Table S4.4: Significance level of Wald tests (pvalues) on the InDels dataset for the GLOBAL phenotyping dataset. The fixed effect tested were the environment (τ), the group (ϒ) of the hybrid and the inbreeding (θ)

		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
		$\tau + \gamma + \theta$	$\tau + \theta$	$\tau + \gamma$	τ	$\tau + \gamma + \theta$ *without interactions*
FFD	**pvalue τ**	2.00×10^{-16}	2.00×10^{-16}	2.00×10^{-16}	2.20×10^{-16}	2.00×10^{-16}
	pvalue ϒ	0.91		0.91		0.86
	pvalue θ	0.95	0.80			0.99
PH	**pvalue τ**	2.20×10^{-16}	2.20×10^{-16}	2.20×10^{-16}	2.20×10^{-16}	2.20×10^{-16}
	pvalue ϒ	0.01		0.01		0.01
	pvalue θ	6.00×10^{-3}	4.53×10^{-5}			0.01
GY	**pvalue τ**	0.05	0.05	0.05	0.05	2.20×10^{-16}
	pvalue ϒ	0.08		0.13		3.20×10^{-2}
	pvalue θ	8.64×10^{-7}	1.93×10^{-7}			6.34×10^{-8}

Table S4.5: Genotypic Variance decomposition based on the InDels genotyping dataset. Variances (and their standard error) were indicated for each model.

		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
		$\tau + \gamma + \theta$	$\tau + \theta$	$\tau + \gamma$	τ	$\tau + \gamma + \theta$ *without interactions*
FFD	σ_a^2	9.51 (1.34)	9.42 (1.32)	9.53 (1.33)	9.44 (1.32)	9.97 (1.36)
	σ_d^2	2.00 (0.46)	1.99 (0.46)	1.97 (0.46)	1.96 (0.45)	1.63 (0.45)
	$\sigma_{a\times E}^2$	1.29 (0.22)	1.29 (0.22)	1.29 (0.22)	1.29 (0.22)	
	$\sigma_{d\times E}^2$	0.30 (0.35)	0.30 (0.35)	0.30 (0.35)	0.30 (0.35)	
	σ_E^2	1.430 (0.13)	1.43 (0.13)	1.43 (0.13)	1.43 (0.13)	2.54 (0.10)
	Likelihood	-1918.86	-1917.84	-1920.55	-1919.4	-2009
PH	σ_a^2	306.21 (44.56)	301.80 (43.93)	310.89 (45.73)	294.70 (45.02)	312.31 (45.10)
	σ_d^2	47.57 (16.57)	48.92 (16.59)	52.04 (17.22)	67.24 (18.60)	45.82 (17.37)
	$\sigma_{a\times E}^2$	49.06 (12.51)	49.04 (12.49)	48.53 (12.49)	48.28 (12.44)	
	$\sigma_{d\times E}^2$	3.73 (23.42)	3.01 (23.43)	5.32 (23.40)	4.48 (23.42)	
	σ_E^2	109.53 (9.32)	109.89 (9.35)	108.90 (9.27)	109.16 (9.30)	142.55 (5.32)
	Likelihood	-5371.46	-5374.64	-5378.61	-5385.51	-5392.85
GY	σ_a^2	94.94 (18.56)	97.20 (18.50)	89.76 (19.60)	90.25 (19.36)	98.48 (18.35)
	σ_d^2	40.34 (12.11)	39.20 (11.89)	57.22 (13.93)	59.30 (13.79)	36.81 (12.17)
	$\sigma_{a\times E}^2$	29.24 (9.78)	29.18 (9.78)	28.83 (9.74)	28.84 (9.73)	
	$\sigma_{d\times E}^2$	70.59 (16.67)	70.86 (16.67)	71.80 (16.57)	71.84 (16.53)	
	σ_E^2	56.69 (5.31)	56.61 (5.30)	55.98 (5.23)	55.81 (5.21)	115.63 (4.33)
	Likelihood	-5067.94	-5070.907	-5082.217	-5085.93	-5131.71

Table S4.6: Significant QTLs identified with InDels. For each QTL we displayed the Chr Chromosome, Position of the highly associated InDels, Log –log10(pvalue), the effect of the QTL, the Minor Allele Frequency, the environment where the QTL was significant, the genetic effect (A=additive, D=Dominant), the indels type (DEL=Deletion, INS=Insertion), the length of the InDels, FRAC_SPE proportion of indel sequence present elsewhere in the genome, the distance to the closest gene and the closest gene.

Chr.	Pos.	Trait	Log (p)	Effect	MAF	Env.	Genetic effect	InDel type	InDel length	Specific frac.	Dist. to closest gene	Closest gene
1	45121579	PH	5.88	-18.2	0.16	SMH10	D	DEL	4527	0.15	3287	GRMZM2G053669
1	59462760	PH	5.83	-12.7	0.14	SMH09; SMH0910	A	INS	869	0.06	NA	
1	224236439	GY	5.32	3.96	0.35	GLOBAL	D	DEL	318	0.16	0	GRMZM2G031572
1	287112264	PH	5.03	-15.3	0.19	SMH10	D	DEL	172	0.23	372	GRMZM2G006047
2	163729138	PH	5.03	7.69	0.35	SMH09	A	DEL	569	0.87	0	GRMZM2G087921
2	163729138	GY	5.49	7.35	0.35	SMH09	A	DEL	569	0.87	0	GRMZM2G087921
4	37022456	GY	7.19	11.04	0.21	SMH10 ; SMH0910	D	INS	118	1	NA	
4	156932880	PH	5.02	4.6	0.28	GLOBAL	D	INS	303	0.31	NA	
4	157508576	PH	5.02	4.44	0.4	GLOBAL	D	INS	251	1	NA	
5	27953268	GY	5.82	5.21	0.27	GLOBAL	D	INS	1594	0	740	AC184866.3_FG005
6	14199941	GY	5.6	10.13	0.24	SMH09 ; SMH0910	A	DEL	1778	0.87	0	GRMZM2G014055
6	101232357	GY	5.01	-S4.65	0.46	GLOBAL	D	INS	1838	0.56	NA	
7	45118094	GY	5.82	-5.82	0.29	GLOBAL	D	DEL	1585	0.38	79	GRMZM5G857404
7	153843728	GY	5.73	-7.91	0.34	SMH09	A	INS	280	0.13	NA	
8	152588666	GY	5.01	5.45	0.34	SMH09	D	INS	110	1	205	GRMZM2G360219
9	151982938	GY	5.47	4.91	0.43	SMH09	D	INS	394	0.28	NA	
10	622963	GY	5.13	5.21	0.2	GLOBAL	D	INS	135	1	NA	
NA	NA	PH	5.04	-12.5	0.21	SMH10	D	INS	1021	0.29	NA	
NA	NA	GY	5.14	3.68	0.5	GLOBAL	D	INS	520	0.47	NA	
NA	NA	PH	6.08	8.28	0.46	GLOBAL	A	INS	356	0	NA	

Table S4.7: Major significant QTLs identified with SNP. We chose the 5 most highly associated SNPs markers in each GWAS (trait x Environment). For each QTL we displayed the Chr Chromosome, Position of the highly associated InDels, Log –log10(pvalue), the effect of the QTL, the Minor Allele Frequency, the environment where the QTL was detected, the genetic effect (A=additive, D=Dominant), the distance to the closest gene and the closest gene.

Chr.	Pos .	Trait	Log (p)	Effect	MAF	Env.	Genetic effect	Dist. to closest gene	Closest gene
1	8,625,682	GY	5.79	8.35	0.29	SMH09; SMH0910	D	0	GRMZM2G171324
1	14,509,246	FFD	5.02	1.34	0.38	SMH0910	A	1929	GRMZM2G091916
1	16,045,372	FFD	5.25	-1.96	0.14	SMH0910	A	2768	GRMZM2G178852
1	16,120,644	FFD	5.03	-1.36	0.33	SMH09; SMH10; SMH0910	A	0	GRMZM2G399072
1	24,600,929	GY	5.45	7.65	0.14	GLOBAL; SMH09; SMH0910	D	0	GRMZM2G119150
1	40,419,605	PH	5.2	-8.78	0.33	SMH0910	A	0	GRMZM2G087600
1	52,462,456	GY	5.34	8	0.28	SMH10	D	0	GRMZM2G009253
1	59,467,431	PH	6.36	-11.3	0.2	SMH09; SMH0910	A	0	AC205502.4_FG004
1	63,540,741	GY	5.95	9.23	0.19	SMH09	D	0	GRMZM2G703598
1	72,819,282	FFD	5.05	-1.04	0.33	SMH10	D	19329	AC177908.3_FG002
1	86,359,841	GY	5.14	4.44	0.33	GLOBAL	D	0	GRMZM2G113761
1	155,993,763	GY	5.26	10.81	0.07	GLOBAL	D	0	GRMZM2G399325
1	156,070,750	FFD	5.07	1.41	0.4	SMH09	A	86165	AC210816.3_FG006
1	159,758,006	FFD	5.06	-1.28	0.49	SMH10	A	25964	GRMZM2G366843
1	182,101,938	PH	5.59	-11.7	0.24	SMH10	D	0	GRMZM2G023591
1	196,884,558	GY	5.94	6.63	0.25	GLOBAL; SMH0910	A	78359	GRMZM2G472671
1	282,899,897	FFD	5.66	1.01	0.43	GLOBAL	D	0	GRMZM2G170727
1	287,112,172	PH	5.02	-17.7	0.15	SMH10	D	1276	GRMZM2G303972
1	293,133,416	PH	5.21	15.86	0.13	SMH09	D	5345	GRMZM2G061900
1	20107707-20109105	GY	5.29	6.66	0.22	GLOBAL	D	0	GRMZM2G028136
2	1,816,959	PH	5.31	-7.68	0.45	GLOBAL; SMH09	A	0	GRMZM2G084928
2	3,583,157	PH	5.01	-11.9	0.15	SMH09	A	0	GRMZM2G062354
2	6,052,505	GY	5.54	11.83	0.15	SMH10	D	0	GRMZM2G021635
2	9,002,083	PH	5.01	8.02	0.18	GLOBAL	A	1577	GRMZM2G300497
2	9,943,419	GY	5.28	8.17	0.22	SMH0910	D	0	GRMZM2G337532
2	12,683,354	GY	5.14	8.47	0.24	SMH09	D	3266	GRMZM5G862947
2	13,879,665	FFD	5.95	-0.91	0.46	GLOBAL; SMH09; SMH10; SMH0910	D	0	GRMZM2G032554
2	15,483,131	PH	5.29	4.2	0.46	GLOBAL	D	1329	GRMZM2G318408
2	32,492,387	PH	5.01	-7.89	0.46	SMH09	A	0	GRMZM2G154397

Chr.	Pos .	Trait	Log (p)	Effect	MAF	Env.	Genetic effect	Dist. to closest gene	Closest gene
2	39,311,569	FFD	5.16	-1.04	0.3	SMH0910	D	4186	AC185655.3_FG005
2	52,326,247	PH	5.14	5.64	0.24	GLOBAL	D	10827	GRMZM2G107588
2	54,340,358	PH	5.25	-11.1	0.21	SMH10; SMH0910	A	20689	GRMZM2G134432
2	55,680,198	PH	5.17	-6.66	0.36	GLOBAL; SMH10; SMH0910	A	0	GRMZM2G136306
2	65,683,170	PH	5.5	8.91	0.11	GLOBAL	D	146423	GRMZM2G067833
2	143,272,601	PH	5.04	-10	0.18	SMH09	A	8656	GRMZM2G028183
2	144,833,059	FFD	5.09	-1.29	0.3	SMH10	A	32568	GRMZM2G092581
2	163,794,398	GY	5.06	-7.24	0.32	SMH09	A	3128	AC183661.3_FG004
2	202,403,573	PH	6.13	-14.3	0.22	SMH10	D	16071	GRMZM2G064655
2	204,173,138	FFD	5.18	-0.93	0.38	SMH0910	D	0	GRMZM5G803981
2	210,357,046	PH	5.49	-9.86	0.21	SMH0910	A	1644	GRMZM2G427807
2	211,727,109	GY	5.74	10.84	0.19	SMH10	D	0	GRMZM2G056564
2	215,041,990	PH	5.01	-6.18	0.4	GLOBAL	A	886	GRMZM2G350773
2	215,946,319	FFD	5.04	-1.34	0.28	SMH10	A	1098	GRMZM2G458437
2	217,228,648	PH	5.42	-9.44	0.29	SMH09	D	0	GRMZM2G151983
2	219,980,765	FFD	5.96	1.06	0.38	SMH10	D	24065	GRMZM2G137546
2	221,082,873	FFD	5.34	-1.87	0.18	SMH10	D	0	GRMZM2G112839
2	228,083,027	PH	6.38	4.28	0.34	GLOBAL	D	0	GRMZM2G094959
2	232,465,618	PH	5.67	5.12	0.37	GLOBAL	D	0	GRMZM2G144890
2	233,734,033	GY	6.4	5.81	0.43	SMH09; SMH0910	D	0	GRMZM2G037923
2	234,718,641	FFD	5.12	-0.85	0.43	GLOBAL; SMH10	D	0	GRMZM2G324507
2	14893290-14893339	GY	5.46	-7.11	0.32	SMH09	A	0	GRMZM2G129399
2	17220746-17221442	GY	5.89	9.78	0.22	SMH10	D	12150	GRMZM2G585837
2	45716861-45766180	PH	5.33	10.12	0.32	SMH09	D	1429	GRMZM2G021459
2	55557627-55558392	PH	5.04	-8.63	0.2	SMH10; SMH0910	A	0	GRMZM2G128322
2	7621298-7623176	GY	5.1	-5.65	0.31	GLOBAL; SMH09	A	4926	GRMZM2G411046
3	1,406,710	PH	5.02	-6.83	0.41	GLOBAL; SMH10	A	3607	GRMZM5G835677
3	2,948,238	GY	5.1	-7.34	0.29	SMH10	A	0	GRMZM2G061005
3	15,440,932	PH	5.23	4.62	0.34	GLOBAL; SMH10; SMH0910	D	1830	GRMZM2G401581
3	29,095,681	FFD	6.04	1.55	0.27	SMH10	A	0	GRMZM2G166524
3	58,715,786	GY	6.1	6.43	0.33	SMH09	D	0	GRMZM2G378547
3	89,609,331	GY	5.15	-6.39	0.38	SMH09	D	0	GRMZM2G503156
3	114,352,093	GY	5.34	7.9	0.19	GLOBAL	D	0	GRMZM2G105991
3	158,971,477	FFD	5.31	1.49	0.4	SMH10	A	6890	GRMZM2G171622

Chr.	Pos .	Trait	Log (p)	Effect	MAF	Env.	Genetic effect	Dist. to closest gene	Closest gene
3	170,863,207	GY	5.11	-8.55	0.22	SMH0910	D	4693	GRMZM2G144602
3	171,445,575	PH	5	-10.7	0.26	SMH10	D	0	GRMZM2G110398
3	171,469,580	PH	5.68	-13	0.24	SMH10	D	4588	GRMZM2G354711
3	171,745,377	GY	5.04	-5.93	0.42	SMH10	A	7090	GRMZM2G039365
3	180,742,905	FFD	5.3	-1.39	0.14	SMH09; SMH0910	A	4260	GRMZM2G157679
3	208,658,481	PH	5.7	-9.51	0.49	SMH10	A	32934	GRMZM2G022052
3	210,821,486	GY	6.63	-7.66	0.49	SMH09	A	0	GRMZM2G150631
3	214,822,666	PH	5.16	-6.85	0.43	GLOBAL	A	0	GRMZM2G067583
3	131149180-131298807	FFD	6.84	1.57	0.49	SMH09; SMH10; SMH0910	A	65503	GRMZM2G016598
3	3404168-3406301	GY	5.38	6.62	0.43	SMH10	D	4832	GRMZM2G400954
3	41329831-41400782	GY	5.42	11.13	0.17	SMH10	D	35302	GRMZM2G416708
3	92695574-92865146	GY	5.61	8.37	0.24	SMH09	D	13660	GRMZM2G148300
4	1,184,594	FFD	5.02	-0.93	0.4	GLOBAL	D	5128	GRMZM2G534826
4	12,975,337	GY	6.1	-11.3	0.16	SMH09	A	0	GRMZM2G124593
4	26,298,094	GY	7.49	8.7	0.16	GLOBAL	D	4643	GRMZM2G004997
4	27,418,417	PH	5.55	-8.46	0.35	SMH09	D	3598	GRMZM2G003662
4	28,115,626	PH	6.79	9.25	0.22	GLOBAL	D	9817	GRMZM2G020801
4	31,713,938	GY	5.56	2.94	0.22	GLOBAL	D	0	GRMZM2G120592
4	36,733,566	FFD	7.73	-1.85	0.28	GLOBAL; SMH10; SMH0910	A	9325	GRMZM2G050118
4	37,022,456	GY	7.85	11.47	0.21	SMH10	D	0	GRMZM2G302259
4	39,327,424	GY	5.34	-5.57	0.24	GLOBAL	D	0	GRMZM2G005583
4	102,711,237	PH	6.35	-15.7	0.13	SMH10; SMH0910	A	64434	AC194924.3_FG002
4	123,994,655	PH	5.07	5.44	0.29	GLOBAL	D	0	GRMZM2G403915
4	126,000,516	GY	5.99	2.79	0.3	GLOBAL	D	9305	GRMZM2G133568
4	146,060,989	GY	6.49	6.53	0.13	GLOBAL, SMH09	D	2904	GRMZM2G146786
4	155,489,111	GY	5.76	5.27	0.21	GLOBAL	D	41950	GRMZM2G301585
4	172,085,151	PH	7.21	7.53	0.17	GLOBAL	D	4520	GRMZM2G075124
4	179,710,419	PH	5.29	4.28	0.41	GLOBAL	D	0	GRMZM2G035704
4	201,477,709	GY	5.05	2.79	0.39	GLOBAL	D	3749	GRMZM2G017411
4	206,745,673	GY	5.2	5.79	0.5	SMH09	D	81620	GRMZM2G035068
4	210,047,921	GY	6.17	5.16	0.24	GLOBAL	D	1391	AC212343.3_FG002
4	210,137,466	GY	5.18	6.86	0.28	SMH09	D	0	GRMZM2G109071
4	229,557,310	GY	5.24	-4.33	0.27	GLOBAL; SMH09	A	20056	GRMZM2G040452
4	234,052,342	GY	5.04	3.49	0.47	GLOBAL	D	60209	GRMZM2G026151
4	234,501,811	GY	5.17	4.41	0.23	GLOBAL	D	0	GRMZM2G570768
4	236,172,599	GY	5.58	7.23	0.25	GLOBAL	D	0	GRMZM2G041159

Chr.	Pos .	Trait	Log (p)	Effect	MAF	Env.	Genetic effect	Dist. to closest gene	Closest gene
4	236,733,130	PH	5.25	12.34	0.18	SMH10	D	2149	GRMZM2G101916
4	106687232-108061392	GY	7.75	4.19	0.29	GLOBAL	D	60042	AC210035.3_FG002
4	126475219-126646939	GY	7.99	3.98	0.25	GLOBAL	D	0	GRMZM2G159049
4	199956398-199958785	FFD	5.02	-0.93	0.38	SMH0910	D	0	GRMZM2G098557
4	223075763-223075982	PH	5.06	-7.08	0.41	GLOBAL	A	24933	AC205834.3_FG006
4	225879843-225955977	GY	5.48	4.48	0.24	GLOBAL; SMH09	D	24565	GRMZM2G153602
5	8,618,375	GY	5.02	9.57	0.14	GLOBAL	D	0	GRMZM5G839349
5	15,701,584	PH	5.26	4.73	0.45	GLOBAL	D	0	GRMZM2G007810
5	69,321,839	FFD	5.56	-1.05	0.45	SMH10; SMH0910	D	0	GRMZM2G088114
5	94,872,205	FFD	5.07	-1.15	0.31	SMH10	D	21040	AC204775.3_FG006
5	141,256,034	PH	6.68	-8.52	0.35	GLOBAL; SMH09; SMH10; SMH0910	A	0	AC214771.3_FG003
5	146,375,848	GY	5	-8.26	0.25	SMH09	A	3096	GRMZM2G064558
5	166,540,056	PH	5.19	-15.4	0.18	SMH10	D	25163	GRMZM2G047999
5	168,885,676	PH	5.96	-9.52	0.32	SMH10	A	0	GRMZM2G115504
5	173,806,622	PH	5.52	-7.31	0.36	SMH0910	D	0	GRMZM2G702026
5	193,536,597	GY	5.83	7.64	0.19	GLOBAL	D	2228	GRMZM2G465953
5	206,909,375	GY	5.02	6.71	0.39	SMH10; SMH0910	D	5835	GRMZM2G105539
5	211,200,023	FFD	5.33	-1.59	0.22	SMH10	D	1879	AC193374.2_FG008
5	105553383-107932483	PH	5.02	10.02	0.08	GLOBAL	D	7819	GRMZM2G374263
5	179270601-179288758	GY	5.37	6.12	0.37	SMH09	D	10758	GRMZM2G113064
5	61794547-61801008	GY	7.29	5	0.4	GLOBAL	D	2248	GRMZM5G811851
6	3,986,632	FFD	5.17	-2.06	0.16	GLOBAL	A	0	GRMZM2G476040
6	3,986,632	GY	5.88	-11.1	0.16	GLOBAL	A	0	GRMZM2G476040
6	86,627,246	PH	5.19	11.4	0.17	SMH09	D	0	GRMZM5G899378
6	94,844,621	GY	5.07	4.61	0.36	GLOBAL	D	0	GRMZM5G815846
6	109,292,534	GY	5.27	5.55	0.46	SMH09	D	3843	GRMZM2G048194
6	112,589,871	PH	5.21	4.83	0.49	GLOBAL	D	0	GRMZM2G142751
6	120,593,736	GY	5.31	7.96	0.16	GLOBAL	D	2254	GRMZM2G397889
6	122,336,190	GY	5.5	5	0.27	GLOBAL	D	47157	AC194592.2_FG001
6	128,629,690	PH	5.97	8.52	0.48	SMH10	D	12617	GRMZM2G137849
6	130,132,330	FFD	5.06	1.45	0.27	SMH0910	A	0	GRMZM5G886257
6	133,743,129	PH	5.24	4.64	0.43	GLOBAL	D	63960	GRMZM2G176655
6	138,432,576	PH	5.19	11.14	0.23	SMH10	D	2675	GRMZM2G117439

Chr.	Pos .	Trait	Log (p)	Effect	MAF	Env.	Genetic effect	Dist. to closest gene	Closest gene
6	143,306,584	GY	5.6	7.72	0.15	GLOBAL; SMH09; SMH0910	D	3445	GRMZM2G014300
6	150,742,852	PH	5.3	4.63	0.33	GLOBAL	D	0	GRMZM2G066189
6	150,802,335	PH	5.91	7.51	0.38	SMH09	D	3865	GRMZM2G067156
6	158,481,116	FFD	5.25	1.71	0.17	GLOBAL	D	0	GRMZM2G435475
6	162,147,163	PH	5.48	-7.7	0.27	SMH0910	A	5451	AC204050.4_FG006
6	167,576,720	PH	6.21	-22.9	0.11	SMH10	D	0	GRMZM2G070659
6	150837664-150940465	PH	5.85	14.45	0.11	SMH09; SMH09	D	8124	GRMZM2G178797
6	39703185-39867984	PH	5.41	5.07	0.35	GLOBAL; SMH09	D	44084	GRMZM2G174716
6	96689281-96882712	GY	5.79	11.98	0.13	SMH09	D	0	AC203752.5_FG002
7	38,899,761	PH	5.36	-10.8	0.13	GLOBAL	A	3672	GRMZM5G816799
7	47,970,597	PH	5.33	-14.7	0.19	SMH10	D	45062	GRMZM2G026218
7	92,258,579	PH	5.09	-10.7	0.26	SMH09	A	51210	GRMZM2G571942
7	115,724,716	PH	5.34	-14.8	0.2	SMH10	D	0	GRMZM2G412601
7	129,310,531	PH	5.56	-9.15	0.34	SMH09	A	40778	GRMZM2G115304
7	136,176,460	GY	5.18	-4.43	0.42	GLOBAL	D	0	GRMZM5G867125
7	138,728,692	FFD	5.05	-1.27	0.48	SMH10	A	0	AC234154.1_FG007
7	157,434,910	PH	5.04	-9.1	0.39	SMH10	D	6435	GRMZM5G801949
7	165,511,148	PH	5.42	-13.5	0.15	SMH09	A	11567	GRMZM2G414043
7	19419871-19442742	GY	5.68	-8.37	0.3	SMH09	A	23977	GRMZM2G104559
7	43364019-43405815	PH	6.05	-9.72	0.16	SMH09	A	9619	GRMZM2G366803
8	79,070,583	PH	5.08	-8.02	0.39	SMH0910	A	3981	GRMZM2G126026
8	123,055,339	PH	5.58	-8.14	0.41	SMH0910	A	0	AC209819.3_FG006
8	123,104,112	FFD	5.01	-1.26	0.49	GLOBAL	A	4091	GRMZM2G102638
8	123,148,008	FFD	5.4	-1.38	0.43	GLOBAL	A	2589	GRMZM2G096115
8	124,937,035	GY	5.82	7.26	0.21	GLOBAL	D	0	GRMZM2G319747
8	131,967,089	GY	5.54	6.59	0.46	SMH10	D	1483	AC219006.2_FG006
8	152,769,397	GY	5.16	6.92	0.28	SMH09	D	0	GRMZM2G395672
8	166,410,663	GY	5.13	5.5	0.46	SMH0910	D	0	GRMZM2G046025
8	169,139,234	GY	5.93	-8.03	0.17	SMH09	A	0	GRMZM2G370048
8	123103962-123104112	PH	5.09	-8.18	0.48	GLOBAL; SMH09; SMH0910	A	4174	GRMZM2G102638
8	123200729-123272484	FFD	6.65	-1.56	0.5	GLOBAL; SMH09; SMH0910	A	32426	GRMZM2G091592
8	123411505-123473596	FFD	6.58	-1.5	0.46	GLOBAL	A	59232	GRMZM2G073044
8	123504353-123504889	FFD	6.54	-1.59	0.49	GLOBAL; SMH09; SMH0910	A	0	GRMZM2G179257

Chr.	Pos .	Trait	Log (p)	Effect	MAF	Env.	Genetic effect	Dist. to closest gene	Closest gene
8	133560305-133562520	FFD	5.62	-1.41	0.44	SMH10; GLOBAL	A	1064	GRMZM2G124755
8	139345470-139345499	PH	5.2	-11.2	0.42	GLOBAL; SMH10	A	0	GRMZM2G027333
8	168175579-168175667	GY	5.06	10.08	0.15	SMH09	D	5488	GRMZM2G028768
8	79143441-79148716	PH	6.7	8.4	0.46	GLOBAL; SMH0910	A	1731	GRMZM2G417164
9	102,939,820	PH	5.33	-11.5	0.17	SMH09	A	0	GRMZM2G152120
9	114,964,366	GY	5.3	8.85	0.17	SMH10	A	0	GRMZM5G813105
9	126,722,676	GY	5.78	6.25	0.4	SMH0910	D	0	GRMZM2G012140
9	144,880,648	PH	5.03	8.09	0.34	SMH10	D	0	GRMZM2G021573
9	146,081,357	PH	5.05	5.42	0.24	GLOBAL; SMH0910	D	5827	GRMZM2G046729
9	149,042,976	PH	5.74	15.7	0.19	SMH10	D	0	GRMZM2G096348
9	150,207,102	GY	5.88	10.57	0.2	SMH09	D	3085	GRMZM5G860976
9	110489960-110491284	GY	5.94	-8.49	0.27	SMH09; SMH0910	A	5702	GRMZM2G451746
9	125127223-125128492	PH	5.12	3.69	0.49	GLOBAL; SMH09	D	0	GRMZM2G127141
9	145900003-145903569	PH	5.61	-10.5	0.1	GLOBAL	A	20161	GRMZM2G126834
9	152186846-152189673	GY	5.33	5.71	0.34	SMH09	D	0	GRMZM2G178072
9	9801258-9801319	PH	5.8	10.18	0.32	SMH10	D	12885	GRMZM2G067743
10	594,777	GY	5.83	5.48	0.21	GLOBAL	D	40548	GRMZM2G109674
10	1,945,767	PH	6.22	5.28	0.42	GLOBAL	D	1683	GRMZM2G043119
10	4,911,409	GY	6.27	6.17	0.17	GLOBAL; SMH10	D	0	GRMZM2G014282
10	22,418,621	GY	5.03	1.95	0.22	GLOBAL	D	27530	GRMZM2G335694
10	38,530,675	GY	6.14	7.13	0.21	GLOBAL	D	0	GRMZM2G010406
10	80,053,613	FFD	5.4	1.57	0.17	GLOBAL	D	0	GRMZM2G385989
10	104,186,232	PH	5.41	5.72	0.27	GLOBAL	D	0	GRMZM2G178537
10	130,145,383	FFD	5.19	1.07	0.3	GLOBAL	D	0	GRMZM2G139797
10	148,607,529	GY	5.56	-5.58	0.22	GLOBAL	A	0	GRMZM2G074787
10	148,636,360	GY	5.24	-5.11	0.2	GLOBAL	A	0	GRMZM2G074754
10	149,269,932	FFD	5	-1.98	0.13	SMH09	A	0	GRMZM2G110932
10	100099929-100227228	GY	5.47	8.21	0.2	SMH09	D	0	GRMZM2G017164
10	100229025-100247073	GY	5.03	9.38	0.15	SMH09	D	6071	GRMZM2G035282
10	1026785-1026792	GY	5.21	3.78	0.28	GLOBAL	D	53424	GRMZM5G854762
10	36775171-36777827	GY	5.46	8.72	0.21	GLOBAL	D	2050	GRMZM2G049915
10	79271295-79275300	GY	5.61	6.48	0.32	SMH09	D	0	GRMZM2G032564

Discussion générale

Un seul génome de référence ne permet pas de capturer l'ensemble des régions génomiques

Des études précédentes avaient déjà identifié qu'un seul génome de référence ne permettait pas de capturer l'ensemble de la diversité génétique d'une espèce avec par exemple 30% du génome du maïs qui n'était pas présent dans le génome de référence B73 (Gore et al., 2009; Lu et al., 2015). Le reséquençage de nombreux génomes a permis d'identifier également, chez plusieurs espèces végétales, un nombre important de gènes qui ne sont pas partagés par l'ensemble des individus. Un dixième des gènes d'Arabidopsis sont absents d'au moins une des 80 lignées reséquencées (Tan et al., 2012) et 18% des gènes identifiés parmi 10 lignées de chou ne sont pas partagés par ces lignées (Golicz et al., 2016). Chez le riz, grâce au reséquençage de plus de 3000 lignées, plus de 10 000 gènes ont été identifiés comme absents du génome de référence (Sun et al., 2017; Wang et al., 2018). Sur 140 500 gènes identifiés à partir de 18 lignées de blé, environ 59 430 sont absents d'au moins une de ces lignées (Montenegro et al., 2017). Plus récemment, le reséquençage de 287 lignées cultivées de tournesol, 17 populations et 189 apparentés sauvages a montré que 27% des gènes n'étaient pas partagés entre ces lignées (Hübner et al., 2019). Ces études montrent que les séquences non partagées peuvent contenir des gènes, et donc que leur présence / absence pourrait impacter le phénotype des individus. Ces résultats illustrent la nécessité d'utiliser comme référence un pangénome, définit comme un génome comportant l'ensemble des gènes d'une espèce, composé à la fois de gènes partagés par l'ensemble des individus et de gènes qui peuvent être absents dans un sous-ensemble d'individus (Tettelin et al., 2005; Li et al., 2010).

Nous avons donc développé une puce afin de génotyper et tester l'effet de 61 942 séquences non partagées entre 4 lignées (B73, F2, C103 et PH207) utilisées pour la découverte de nos InDel. Parmi ces séquences, 22 987 ont été annotées et 7 988 sont situées dans des gènes.

Outre le fait de pouvoir découvrir des InDel, l'utilisation d'un pangénome comme référence permet de découvrir des polymorphismes dans les séquences non partagées par l'ensemble des individus et d'identifier des nouvelles régions impliquées dans l'architecture des caractères et/ou l'adaptation. Dans notre étude, les SNP génotypés avec les puces 50K et 600K ainsi que le GBS, ont été découverts à partir de l'alignement de séquences seulement sur le génome de référence B73. Les SNP présents dans les régions absentes du génome de référence n'ont donc pas pu être découverts et par conséquent n'ont jamais été directement utilisés dans les analyses génétiques comme la génétique d'association ou la recherche de signatures de sélection (Hurgobin and Edwards, 2017). Par ailleurs, j'ai identifié une moins forte densité de SNP génotypés aves les puces 50K et 600K et le GBS dans les séquences présentes chez B73 mais absentes d'une des 3 lignées (F2, C103 et PH207) par rapport à la densité de SNP sur le génome entier (1 SNP/3669bp vs 1 SNP/1913bp). Ce résultat suggère que lors de la découverte des SNP et/ou du design des puces, les SNP présents dans ces régions ont été contre sélectionnés, probablement dû à un taux de données manquantes important à cause de l'absence des séquences non partagées chez certains individus (Didion et al., 2012). Grâce à notre puce de génotypage des InDel, j'ai génotypé 37 737 SNP dans

des séquences non partagées entre 4 lignées de maïs (B73, F2, C103 et PH207). J'ai ainsi pu identifier 39 associations significatives dans le panel de lignées entre le génotype au SNP considéré et des caractères de phénologie, d'architecture et des composantes du rendement. Ces résultats indiquent que les SNP dans les InDel peuvent révéler de nouveaux QTL. De façon similaire, Yao et al. (2015) ont réalisé une étude d'association à partir de 584 lignées de riz reséquencées. Ils ont montré que 41,6% des SNP associés à la largeur du grain et aux 840 caractères métaboliques analysés ont été découverts dans des régions non partagées entre les lignées.

Ce biais de découverte des SNP à partir d'une seule référence chez le maïs a certainement biaisé l'étude de la complémentarité entre les deux groupes hétérotiques majoritairement utilisés dans les programmes de sélection nord-européens (corné et denté). Leur étude a permis précédemment d'identifier des allèles en fréquence différentiée entre les groupes, notamment pour des gènes impliqués dans la floraison par une approche de détection de signatures de sélection (Unterseer et al., 2016). Ces travaux, basés sur les SNP de la puce 600K, n'ont certainement permis d'étudier qu'une partie de la complémentarité entre les groupes génétiques, car les SNP ont été découverts à partir d'un alignement de séquences sur le génome de référence B73 qui est une lignée dentée américaine (Unterseer et al., 2014). Les séquences présentes uniquement chez les cornés (et absentes chez les dentés) ne peuvent donc pas être détectées en utilisant comme référence un génome denté. L'utilisation des polymorphismes découverts à partir d'un pangénome regroupant l'ensemble de l'information génétique contenue dans les lignées dentées et cornées permettrait de mieux comprendre la complémentation entre les séquences dans les hybrides issus de croisement entre les lignées ces deux groupes majeurs.

Les InDel révèlent de nouvelles régions impliquées dans la variation des caractères d'intérêt

Plusieurs études ont montré chez l'humain que les polymorphismes présent/absent étaient majoritairement en déséquilibre de liaison (DL) élevé avec des SNP adjacents (Redon et al., 2006; Conrad et al., 2010; Mills et al., 2011; Sudmant et al., 2015). Chez le maïs, Chia et al. (2012) ont identifié que 88% des séquences présentant une variation de nombre de copies étaient significativement associées aux SNP par DL. Plus récemment, Lu et al. (2015) ont montré que 228 620 InDel sur 1,1 million (~20%) étaient significativement associées à des SNP (p-value < 1x10-6). Ces résultats suggèrent que les SNP pourraient capturer en grande partie l'effet des polymorphismes présent/absent dans les analyses génétiques et donc qu'il n'y aurait pas ou peu d'intérêt à génotyper les InDel en plus des SNP. J'ai pu montrer dans cette étude que 51% des InDel positionnés n'étaient pas en DL supérieur à 0.8 avec au moins un SNP. Par conséquent, l'effet de la présence et de l'absence des séquences de ces InDel sur le phénotype n'est pas capturé facilement par les SNP que nous avons utilisés. Néanmoins, 23 786 des 38 259 InDel (62%) non positionnées sont en DL supérieur à 0.5 avec au moins un SNP ce qui suggère que l'effet de ces InDel pourrait être partiellement capturé par des SNP à condition que l'effet sur le phénotype soit fort.

Chez le maïs, l'effet de la présence et de l'absence des séquences sur le phénotype n'a pas été directement testé. En effet, les études précédentes ont testé indirectement l'effet des InDel en testant

l'effet des SNP, en DL ou non avec ces InDel (Chia et al., 2012; Lu et al., 2015). Grâce à notre puce de génotypage, j'ai pu tester directement l'effet de la présence/absence de la séquence de 61 942 InDel sur la variation phénotypique de 23 caractères mesurés pour 362 lignées de maïs. J'ai également testé l'effet d'un million de SNP pour étudier la complémentarité entre les deux types de polymorphismes. J'ai ainsi détecté 294 QTL pour ces 23 caractères dont 6 grâce aux InDel et aux SNP et 13 uniquement grâce aux InDel. L'effet de ces 13 QTL n'était donc jusqu'à maintenant pas capturé avec cette densité de SNP.

Globalement, nous nous attendions à trouver un effet plus fort des InDel par rapport aux SNP dans la mesure où les InDel peuvent correspondre à l'absence totale d'une séquence génique et entrainer la perte de leur fonction chez l'individu (Redon et al., 2006). Cependant, j'ai noté une plus faible proportion de QTL détectés avec des InDel par rapport aux SNP. Plusieurs hypothèses pourraient expliquer cette différence. Premièrement, la majorité des InDel que l'on a découvertes sont probablement neutres, comme Cooper et al. (2007) l'ont montré chez l'humain. Deuxièmement, les InDel représentent des séquences qui ne sont pas indispensables pour la survie de la plante (Swanson-Wagner et al., 2010). Cela suggère que leur effet est faible, en particulier pour des caractères essentiels à la survie de la plante, car leur absence n'impacte que peu la fitness des individus. Enfin, les gènes qui se trouvent dans les InDel semblent faire partie de familles multi-géniques (Swanson-Wagner et al., 2010) : l'absence d'un gène peut alors être compensée par les gènes issus de la même famille.

Fu et Dooner (2002), ont fait l'hypothèse que la complémentation des gènes chez l'hybride présents dans un parent et absents dans l'autre pouvait expliquer une part de l'hétérosis : l'effet délétère de l'absence d'un gène chez un parent serait compensé par la présence de ce gène chez le deuxième parent conduisant à un effet de dominance. Pour tester cette hypothèse, nous avons évalué les effets d'additivité et de dominance du polymorphisme présent / absent de 51 844 InDel par génétique d'association sur un panel de 287 hybrides, ce qui n'avait encore jamais été fait chez le maïs à ma connaissance. La proportion de la variance des phénotypes expliquée par la dominance ou les interactions dominance x environnement est plus ou moins forte selon le phénotype (38% pour le rendement, 14% pour la floraison et 10% pour la hauteur avec le modèle complet). Nous avons détecté 21 QTL grâce aux InDel dont 15 avec un effet de dominance significatif. Nous avons également testé l'effet de 469 267 SNP sur les performances de ces hybrides, et détecté 193 QTL dont 122 avec un effet de dominance significatif. D'après l'hypothèse de Fu et Dooner (2002), nous nous attendions à détecter un enrichissement en InDel avec un effet de dominance significatif par rapport aux SNP, ce qui n'a pas été vérifié par mes résultats. Cela étant dit, la complémentation des séquences dans les hybrides sera d'autant plus forte si une séquence génique est présente chez un parent et totalement absente chez l'autre parent (les PAV). Or, seuls 15% des InDel que nous avons analysées en valeur hybride sont totalement spécifiques, *i.e.* des PAV. Parmi les 75% des InDel partiellement spécifiques, seules 9% ont une séquence spécifique supérieure à 1kbp qui pourrait donc contenir des gènes et être complémentées chez les hybrides. La faible proportion de PAV génotypés avec notre puce pourrait expliquer que, bien qu'important, le nombre de QTL détectés avec un effet de dominance significatif ne soit pas plus important pour les InDel.

Bien que nous ayons trouvé des QTL en valeur hybride exclusivement grâce aux InDel, l'intégration des InDel dans les modèles de prédiction génomique basés sur l'approche GBLUP n'a pas amélioré les prédictions par rapport aux SNP seuls, quels que soient le caractère et l'environnement étudiés. Néanmoins, l'intégration d'un vecteur de consanguinité calculé à partir des InDel comme covariable dans mon modèle de prédiction a permis de sensiblement améliorer les prédictions pour le rendement. Etant donné que les matrices d'apparentement estimées avec les InDel ou les SNP sont très corrélées, il semble

cohérent que les InDel n'améliorent pas la précision des prédictions par rapport aux SNP. Le peu de différence entre les valeurs d'apparentement calculées avec les InDel ou les SNP suggère que ces deux polymorphismes auraient suivi une histoire évolutive similaire. Plusieurs études ont montré des résultats similaires chez les humains (Jakobsson et al., 2008; Conrad et al., 2010), mais aussi chez les bovins (Xu et al., 2016).

En comparant les résultats des études d'associations menées sur les panels des lignées et de leurs hybrides pour les 2 caractères mesurés sur les deux panels (floraison et hauteur de plante), je n'ai identifié qu'un seul SNP associé à la fois en valeur propre et en valeur hybride avec la hauteur de plante. Dans l'étude du panel de lignées par génétique d'association, seul l'effet additif du marqueur était testé. Par conséquent, les QTL avec un effet de dominance n'ont pas pu être détectés avec le panel de lignées. De plus, les sélectionneurs ont remarqué que les valeurs phénotypiques des lignées étaient de mauvais prédicteurs des valeurs hybrides (Crow, 1998; Schrag et al., 2009, 2010), ce qui peut expliquer que quasiment aucun QTL n'est commun aux deux panels. Enfin, la diversité génétique et le nombre d'individus sont différents entre les deux panels : le premier panel est composé de lignées cornées, dentées et tropicales alors que les hybrides sont issus uniquement du croisement entre des lignées cornées et dentées. Ceci peut expliquer une différence de puissance de détection et ainsi la différence entre les QTL détectés entre les deux panels (Rincent et al., 2014).

La différence entre les QTL détectés entre les deux panels pourrait aussi avoir une autre origine. Pour l'analyse des lignées, nous avons analysé chaque caractère avec sa moyenne ajustée, calculée à partir de 9 environnements différents. Nous avons donc tamponné les interactions génotype x environnement (GxE) dans cette analyse. Or les effets génétiques peuvent être différents en fonctions des contraintes environnementales (Des Marais et al., 2013). A partir d'un panel de 288 lignées dentées de maïs, Millet et al. (2016) ont identifié une grande majorité de QTL de rendement spécifique de certains environnements très contrastés. Pour l'analyse des valeurs hybrides, nous avons utilisé 4 moyennes ajustées différentes, calculées à partir de 4 combinaisons d'environnements différentes (SMH09, SMH10, SMH0910 et GLOB). 83% des QTL identifiés sont spécifiques d'une seule combinaison d'environnements, ce qui suggère des interactions génotype x environnement significatives. Ces résultats suggèrent une forte dépendance entre l'effet génétique des QTL détectés et les conditions climatiques spécifiques de chaque environnement.

Les InDel contribueraient à l'adaptation à des environnements contrastés

Des études précédentes ont montré un enrichissement en gènes impliqués dans les tolérances aux stress dans les InDel découvertes chez le maïs (Swanson-Wagner et al., 2010; Hirsch et al., 2016; Jiao et al., 2017; Darracq et al., 2018) mais aussi chez de nombreuses autres espèces (McHale et al., 2012; Tan et al., 2012; Saxena et al., 2014; Golicz et al., 2016; Hardigan et al., 2016; Dolatabadian et al., 2017; Sun et al., 2017; Montenegro et al., 2017; Hübner et al., 2019). Les InDel peuvent permettre une adaptation rapide des plantes contraintes à des pressions sélectives fortes, notamment par leur capacité à modifier le contenu et l'expression des gènes (DeBolt, 2010).

Chez le riz, Monat et al. (2018) ont montré que le pangénome de l'espèce cultivée était plus petit que celui de l'espèce sauvage. De plus, la taille du génome non-partagé par l'ensemble des individus est plus importante pour l'espèce cultivée que pour l'espèce sauvage. Ces résultats montrent que la présence

et l'absence des séquences a probablement joué un rôle à la fois dans la domestication du riz mais aussi dans la diversification des riz cultivés.

Dans mon étude, 56 (30%) des 188 régions génomiques sous sélection entre les lignées tropicales, cornées et dentées ont été identifiées par les InDel alors qu'elles ne représentent que 7% des marqueurs, ce qui dénote un enrichissement très significatif en InDel sous sélection par rapport aux SNP. Comme les lignées de ces différents groupes génétiques sont cultivées dans des conditions agro-climatiques très contrastées, on peut s'attendre à ce que la majorité de ces régions génomiques identifiées par les InDel soient impliquées dans l'adaptation à des environnements contrastés, notamment dans les stress biotiques et abiotiques et dans la floraison.

En accord avec cette hypothèse, plusieurs gènes impliqués dans la tolérance aux stress co-localisent avec les régions sous sélection identifiées par les InDel. Deux InDel localisés sur le chromosome 2 sont proches de gènes en lien avec la tolérance au stress, l'une proche d'un gène (ZmAsr2) (7Kbp) en lien avec la tolérance au stress hydrique (Virlouvet et al., 2011) et l'autre à 10Kbp d'un gène issu d'une famille (Zinc finger) impliquée dans la floraison et la tolérance aux stress (Chen and Ni, 2006; Ma et al., 2009). Une InDel sous sélection sur le chromosome 9 est également proche (3.3Kbp) d'un gène de la famille des CAMTA (calmodulin-binding transcription activators), connue pour être impliqué dans la croissance des plantes et la réponse aux stress (Yue et al., 2015). Sur ce même chromosome, une InDel sous sélection est à 6,9Kbp d'un gène impliqué dans la production d'une protéine MAP (Mitogen-Activated Protein) jouant un rôle dans la réponse aux stress (Hua et al., 1998; Lalle et al., 2005).

Nous avons également identifié un QTL de rendement avec les InDel (par génétique d'association sur le panel hybride) sur le bras court du chromosome 6. Cette région a été préalablement identifiée par Millet et al., (2016) comme responsable d'une augmentation du rendement en conditions chaudes. Ils ont identifié un gène dans ce QTL, codant pour une protéine induite par le stress hydrique (protéine ABA). Cette région a également été identifiée comme associée à la tolérance au virus de la mosaïque de la canne à sucre (Gustafson et al., 2018).

La précocité de floraison est le caractère qui a permis au maïs de s'adapter aux climats tempérés par l'ajustement de son cycle (Tenaillon and Charcosset, 2011; Swarts et al., 2017; Romero Navarro et al., 2017). En génétique d'association sur le panel de lignées, c'est pour ce caractère de floraison que j'ai identifié le plus d'associations significatives pour les InDel. Parmi ces associations, 3 sont proches de gènes connus : 2 sur le chromosome 3, proche du gène vgt3 (Salvi et al., 2007) et une sur le chromosome 8 proche du gène vgt1 (Salvi et al., 2002; Ducrocq et al., 2008).

Mes résultats confortent le rôle que pourrait jouer les InDel dans les mécanismes liés à l'adaptation et à la tolérance aux stress. L'étude du rôle des InDel dans l'adaptation pourrait permettre d'accélérer le développement de matériel végétal tolérant aux stress.

Perspectives

Etudier l'adaptation et les interactions génotype x environnement avec les InDel

Un des enjeux de l'agriculture d'aujourd'hui est de produire en quantité une alimentation de qualité pour une population mondiale grandissante dans un contexte de changement climatique et de réduction des intrants qui conduisent à une augmentation des stress biotiques et abiotiques.

Je viens de montrer que les InDel pourraient être potentiellement impliquées dans l'adaptation et dans la tolérance aux stress. Pour confirmer cette hypothèse, il serait intéressant d'évaluer un panel d'individus pour des caractères impliqués dans la valeur sélective des individus et/ou la tolérance aux stress dans plusieurs environnements contrastés. Cela permettrait d'identifier des nouveaux QTL spécifiques de certains environnements potentiellement impliqués dans ces caractères. Sous l'hypothèse que les Indel sont plus fortement impliquées dans les interactions GxE, je m'attendrais donc à détecter un enrichissement de QTL spécifiques de certains environnements pour ces caractères en utilisant les InDel par rapport aux SNP.

Le panel et l'approche multi-environnement utilisés par Millet et al., (2016) semblent être appropriés pour vérifier l'hypothèse selon laquelle les InDel seraient plus souvent impliquées dans les caractères liés à la tolérance aux stress par rapport aux SNP. En effet, ce panel composé de 244 lignées dentées a été évalué pour le rendement dans 29 environnements différents (avec six scénarios de température et déficit en eau) ; des données de protéomique et de métabolomique sont également disponibles. Ce panel a également été génotypé avec les InDel de notre puce et 1M de SNP. Il serait donc l'outil idéal pour tester l'effet des InDel sur la tolérance aux stress hydriques et à la chaleur.

Le développement rapide du séquençage permet maintenant de découvrir des InDel à partir de nombreux génomes

Le choix des lignées pour découvrir les polymorphismes à génotyper peut entrainer un biais dans les analyses, notamment en génétique des populations lorsque la diversité de ces lignées ne représente pas celle de la population à analyser (Clark, 2005; Albrechtsen et al., 2010). Dans notre cas, le choix des lignées utilisées pour la découverte des InDel s'est porté sur 1 lignée cornée européenne (F2) et 3 lignées dentées américaines (B73, C103 et PH207). Cependant, dans le panel de lignées, nous avons analysé également des lignées tropicales qui divergent génétiquement des lignées tempérées (Doebley et al., 1986). Les 4 lignées de maïs utilisées ne représentent donc certainement pas la diversité génétique de notre panel. Comme les conditions agro-climatiques de culture des maïs tropicaux et tempérés sont très contrastées, l'utilisation de lignées tropicales pour découvrir de nouvelles InDel permettrait probablement de découvrir des séquences impliquées dans l'adaptation à des climats tempérés (Romero Navarro et al., 2017; Swarts et al., 2017).

Aujourd'hui, grâce au rapide développement des technologies de séquençage, 10 génomes assemblés de lignées de maïs sont disponibles. Ils permettraient de découvrir un nombre plus important

d'InDel. Nous pouvons espérer que dans un futur proche, le reséquençage de panels complets de lignées de maïs permettra de réaliser des analyses génétiques à partir de plusieurs millions de polymorphismes, dont les SNP et les InDel. Chez le riz, le reséquençage d'un large panel de lignées a permis de réaliser une étude d'association à partir de 584 lignées, avec notamment des SNP découverts dans des séquences non partagées entre les différentes lignées (Yao et al., 2015). Plus récemment, 287 lignées de tournesol, ainsi que 17 populations et 189 apparentés ont été reséquencés (Hübner et al., 2019). Une étude d'association réalisée à partir des SNP découverts entre les différentes lignées a permis d'identifier des QTL en lien avec la résistance au mildiou.

Validation des QTL et sélection assistée par marqueurs

Lorsque le point de cassure est connu, la technologie Kaspar nous permet de génotyper la présence et l'absence des InDel. Une prochaine étape serait alors de caractériser des populations d'introgression avec ces marqueurs pour valider l'effet des QTL identifiés avec des InDel. Ces populations pourraient permettre de caractériser finement l'effet du QTL, notamment sa composante additive et de dominance, mais aussi l'interaction de ce QTL avec le fond génétique. L'étude plus fine de ces régions, et notamment la validation fonctionnelle, permettrait de déterminer si les InDel sont les facteurs causaux ou si ces polymorphismes capturent seulement l'effet d'une région proche par DL. Les QTL dont l'effet sur le phénotype est validé pourraient ensuite être utilisés en sélection assistée par marqueurs.

Etudier la contribution d'autres variations structurales

Je me suis intéressé pendant cette thèse aux variations de type présence et absence à un locus. A l'avenir, j'aurais aimé me focaliser sur l'effet du nombre de copies des gènes sur le phénotype. De nombreux exemples dans la littérature nous montrent que la variation du nombre de copies d'un gène peut avoir un fort effet sur le phénotype (Mileyko et al., 2008; Gaines et al., 2010; Cook et al., 2012; Nitcher et al., 2013; Maron et al., 2013; Francia et al., 2016; Lyra et al., 2018).

Des outils sont aujourd'hui disponibles pour détecter des variations du nombre de copie, notamment à partir des données de génotypage des SNP issus des puces de génotypage (Wang et al., 2007; Cooper et al., 2008; Winchester et al., 2009; Dellinger et al., 2010; Wang et al., 2017). L'application de ces outils aux données de génotypage de la puce 600K que j'ai utilisé permettrait d'identifier des CNV comme l'on fait Lyra et al. (2018). Je pourrais ensuite tester leur effet sur l'ensemble des phénotypes que j'ai étudié, mais aussi leur différentiation entre les groupes génétiques à partir des deux panels que j'ai utilisé pendant ma thèse.

References

Akhunov, E.D., Akhunova, A.R., Linkiewicz, A.M., Dubcovsky, J., Hummel, D., Lazo, G., Chao, S., Anderson, O.D., David, J., Qi, L., Echalier, et al., 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc. Natl. Acad. Sci. 100, 10836. https://doi.org/10.1073/pnas.1934431100

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664. https://doi.org/10.1101/gr.094052.109

Aliloo, H., Pryce, J.E., González-Recio, O., Cocks, B.G., Hayes, B.J., 2016. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genet. Sel. Evol. 48. https://doi.org/10.1186/s12711-016-0186-0

Alkan, C., Coe, B.P., Eichler, E.E., 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376. https://doi.org/10.1038/nrg2958

AltschuP, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool 8.

Anderson, J.E., Kantar, M.B., Kono, T.Y., Fu, F., Stec, A.O., Song, Q., Cregan, P.B., Specht, J.E., Diers, B.W., Cannon, S.B., McHale, L.K., Stupar, R.M., 2014. A Roadmap for Functional Structural Variants in the Soybean Genome. G3amp58 GenesGenomesGenetics 4, 1307–1318. https://doi.org/10.1534/g3.114.011551

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C.J., Stein, N., Choulet, F., Distelfeld, Aet al, 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, eaar7191. https://doi.org/10.1126/science.aar7191

Astle, W., Balding, D.J., 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat. Sci. 24, 451–471. https://doi.org/10.1214/09-STS307

Aulchenko, Y.S., Ripke, S., Isaacs, A., van Duijn, C.M., 2007. GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294–1296. https://doi.org/10.1093/bioinformatics/btm108

Barrett, S.C.H., Charlesworth, D., 1991. Effects of a change in the level of inbreeding ont he genetic load. Nature 352.

Bateson, W., 1908. Facts limiting the theory of heridity. Science 26, 60–649.

Beló, A., Beatty, M.K., Hondred, D., Fengler, K.A., Li, B., Rafalski, A., 2010. Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor. Appl. Genet. 120, 355–367. https://doi.org/10.1007/s00122-009-1128-9

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bernardo, R., 1996. Best Linear Unbiased Prediction of Maize Single-Cross Performance. Crop Sci. 36, 50. https://doi.org/10.2135/cropsci1996.0011183X003600010009x

Bernardo, R., 1994. Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Sci. 34, 20. https://doi.org/10.2135/cropsci1994.0011183X003400010003x

Bernardo, R., 1992a. Relationship between single-cross performance and molecular marker heterozygosity. Theor. Appl. Genet. 83, 628–634. https://doi.org/10.1007/BF00226908

Bernardo, R., 1992b. Relationship between single-cross performance and molecular marker heterozygosity. Theor. Appl. Genet. 83, 628–634. https://doi.org/10.1007/BF00226908

Blanc, G., Charcosset, A., Mangin, B., Gallais, A., Moreau, L., 2006. Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. Theor. Appl. Genet. 113, 206–224. https://doi.org/10.1007/s00122-006-0287-1

Böhm, J., Schipprack, W., Utz, H.F., Melchinger, A.E., 2017. Tapping the genetic diversity of landraces in allogamous crops with doubled haploid lines: a case study from European flint maize. Theor. Appl. Genet. 130, 861–873. https://doi.org/10.1007/s00122-017-2856-x

Bouchet, S., Bertin, P., Presterl, T., Jamin, P., Coubriche, D., Gouesnard, B., Laborde, J., Charcosset, A., 2017. Association mapping for phenology and plant architecture in maize shows higher power for developmental traits compared with growth influenced traits. Heredity 118, 249–259. https://doi.org/10.1038/hdy.2016.88

Bouchet, S., Servin, B., Bertin, P., Madur, D., Combes, V., Dumas, F., Brunel, D., Laborde, J., Charcosset, A., Nicolas, S., 2013. Adaptation of maize to temperate climates: mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the Vgt2 (ZCN8) locus. PLoS One 8, e71377.

Brandenburg, J.-T., Mary-Huard, T., Rigaill, G., Hearne, S.J., Corti, H., Joets, J., Vitte, C., Charcosset, A., Nicolas, S.D., Tenaillon, M.I., 2017. Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLOS Genet. 13, e1006666. https://doi.org/10.1371/journal.pgen.1006666

Brauner, P.C., Müller, D., Schopp, P., Böhm, J., Bauer, E., Schön, C.-C., Melchinger, A.E., 2018. Genomic Prediction Within and Among Doubled-Haploid Libraries from Maize Landraces. Genetics 210, 1185–1196. https://doi.org/10.1534/genetics.118.301286

Browning, B.L., Browning, S.R., 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. Am. J. Hum. Genet. 84, 210–223. https://doi.org/10.1016/j.ajhg.2009.01.005

Browning, S.R., Browning, B.L., 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. 81, 1084–1097. https://doi.org/10.1086/521987

Bruce, A., 1910. The mendelian theory of heridity and the augmentation of vigor. Science 32, 627–628.

Brunner, S., 2005. Evolution of DNA Sequence Nonhomologies among Maize Inbreds. PLANT CELL ONLINE 17, 343–360. https://doi.org/10.1105/tpc.104.025627

Buckler, E.S., Thornsberry, J.M., 2002. Plant molecular diversity and applications to genomics. Curr. Opin. Plant Biol. 5, 107–111. https://doi.org/10.1016/S1369-5266(02)00238-8

Camus-Kulandaivelu, L., 2006. Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics 172, 2449–2463. https://doi.org/10.1534/genetics.105.048603

Camus-Kulandaivelu, L., Veyrieras, J.-B., Madur, D., Combes, V., Fourmann, M., Barrault, S., Dubreuil, P., Gouesnard, B., Manicacci, D., Charcosset, A., 2006. Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics 172, 2449–2463. https://doi.org/10.1534/genetics.105.048603

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J., Weigel, D., 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. 43, 956–963. https://doi.org/10.1038/ng.911

Castelletti, S., Tuberosa, R., Pindo, M., Salvi, S., 2014. A MITE Transposon Insertion Is Associated with Differential Methylation at the Maize Flowering Time QTL *Vgt1*. G3amp58 GenesGenomesGenetics 4, 805–812. https://doi.org/10.1534/g3.114.010686

Causse, M., Desplat, N., Pascual, L., Le Paslier, M.-C., Sauvage, C., Bauchet, G., Bérard, A., Bounon, R., Tchoumakov, M., Brunel, D., Bouchet, J.-P., 2013. Whole genome resequencing in tomato

reveals variation associated with introgression and breeding events. BMC Genomics 14, 791. https://doi.org/10.1186/1471-2164-14-791

Cavanagh, C., Morell, M., Mackay, I., Powell, W., 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr. Opin. Plant Biol. 11, 215–221. https://doi.org/10.1016/j.pbi.2008.01.002

Charcosset, A., Essioux, L., 1994. The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. Theor. Appl. Genet. 89–89, 336–343. https://doi.org/10.1007/BF00225164

Charcosset, A., Lefort-Buson, M., Gallais, A., 1991. Relationship between heterosis and heterozygosity at marker loci: a theoretical computation. Theor. Appl. Genet. 81, 571–575. https://doi.org/10.1007/BF00226720

Charlesworth, D., Willis, J.H., 2009. The genetics of inbreeding depression. Nat. Rev. Genet. 10, 783–796. https://doi.org/10.1038/nrg2664

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., Shi, X., Fulton, R.S., Ley, T.J., Wilson, R.K., Ding, L., Mardis, E.R., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods 6, 677–681. https://doi.org/10.1038/nmeth.1363

Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., Gore, M., Guill, K.E., Holland, J., Hufford, M.B., Lai, J., Li, M., Liu, X., Lu, Y., McCombie, R., Nelson, R., Poland, J., Prasanna, B.M., Pyhäjärvi, T., Rong, T., Sekhon, R.S., Sun, Q., Tenaillon, M.I., Tian, F., Wang, J., Xu, X., Zhang, Z., Kaeppler, S.M., Ross-Ibarra, J., McMullen, M.D., Buckler, E.S., Zhang, G., Xu, Y., Ware, D., 2012a. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44, 803–807. https://doi.org/10.1038/ng.2313

Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., Gore, M., Guill, K.E., Holland, J., Hufford, M.B., Lai, J., Li, M., Liu, X., Lu, Y., McCombie, R., Nelson, R., Poland, J., Prasanna, B.M., Pyhäjärvi, T., Rong, T., Sekhon, R.S., Sun, Q., Tenaillon, M.I., Tian, F., Wang, J., Xu, X., Zhang, Z., Kaeppler, S.M., Ross-Ibarra, J., McMullen, M.D., Buckler, E.S., Zhang, G., Xu, Y., Ware, D., 2012b. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44, 803–807. https://doi.org/10.1038/ng.2313

Clark, A.G., 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15, 1496–1502. https://doi.org/10.1101/gr.4107905

Conrad, D.F., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., MacArthur, D.G., MacDonald, J.R., Onyiah, I., Pang, A.W.C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N.P., Lee, C., Scherer, S.W., Hurles, M.E., 2010. Origins and functional impact of copy number variation in the human genome. Nature 464, 704–712. https://doi.org/10.1038/nature08516

Cook, D.E., Lee, T.G., Guo, X., Melito, S., Wang, K., Bayless, A.M., Wang, J., Hughes, T.J., Willis, D.K., Clemente, T.E., 2012. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science 338, 1206–1209.

Cooper, G.M., Nickerson, D.A., Eichler, E.E., 2007. Mutational and selective effects on copy-number variants in the human genome. Nat. Genet. 39, S22–S29. https://doi.org/10.1038/ng2054

Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E., Nickerson, D.A., 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat. Genet. 40, 1199–1203. https://doi.org/10.1038/ng.236

Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E., Nickerson, D.A., 2008b. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat. Genet. 40, 1199–1203. https://doi.org/10.1038/ng.236

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D., Mathews, K., 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112, 48–60. https://doi.org/10.1038/hdy.2013.16

Darracq, A., Vitte, C., Nicolas, S., Duarte, J., Pichon, J.-P., Mary-Huard, T., Chevalier, C., Bérard, A., Le Paslier, M.-C., Rogowsky, P., Charcosset, A., Joets, J., 2018. Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genomics 19. https://doi.org/10.1186/s12864-018-4490-7

Davenport, C.B., 1908. Degeneration, Albinism and Inbreeding. Science 28, 454–455.

Dell'Acqua, M., Gatti, D.M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., Hlaing, A.L., Aung, H.H., Nelissen, H., Baute, J., Frascaroli, E., Churchill, G.A., Inzé, D., Morgante, M., Pè, M.E., 2015. Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. Genome Biol. 16. https://doi.org/10.1186/s13059-015-0716-z

Dellaporta, S.L., Wood, J., Hicks, J.B., 1983. A plant DNA minipreparation: Version II. Plant Mol. Biol. Report. 1, 19–21. https://doi.org/10.1007/BF02712670

Dellinger, A.E., Saw, S.-M., Goh, L.K., Seielstad, M., Young, T.L., Li, Y.-J., 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 38, e105–e105. https://doi.org/10.1093/nar/gkq040

Didion, J.P., Yang, H., Sheppard, K., Fu, C.-P., McMillan, L., de Villena, F.P.-M., Churchill, G.A., 2012. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics 13, 34. https://doi.org/10.1186/1471-2164-13-34

Doebley, J., Wendel, J.D., Smith, J.S.C., Stuber, C.W., Goodman, M.M., 1988. The origin of cornbelt maize: The isozyme evidence. Econ. Bot. 42, 120–131. https://doi.org/10.1007/BF02859042

Doebley, J.F., Goodman, M.M., Stuber, C.W., 1986. Exceptional Genetic Divergence of Northern Flint Corn. Am. J. Bot. 73, 64. https://doi.org/10.2307/2444278

Dooner, H.K., Robbins, T.P., Jorgensen, R.A., 1991. GENETIC AND DEVELOPMENTAL CONTROL OF ANTHOCYANIN BIOSYNTHESIS 27.

dos Santos, J.P.R., Vasconcellos, R.C. de C., Pires, L.P.M., Balestre, M., Von Pinho, R.G., 2016. Inclusion of Dominance Effects in the Multivariate GBLUP Model. PLOS ONE 11, e0152045. https://doi.org/10.1371/journal.pone.0152045

Dubreuil, P., Warburton, M., Chastanet, M., Hoisington, D., Charcosset, A., 2006. More on the introduction of temperate maize into Europe: a large scalebulk SSR genotyping and new historical elements 51, 281–291.

Ducrocq, S., Madur, D., Veyrieras, J.-B., Camus-Kulandaivelu, L., Kloiber-Maitz, M., Presterl, T., Ouzunova, M., Manicacci, D., Charcosset, A., 2008. Key Impact of Vgt1 on Flowering Time Adaptation in Maize: Evidence From Association Mapping and Ecogeographical Information. Genetics 178, 2433–2437. https://doi.org/10.1534/genetics.107.084830

Duvick, D.N., 2005. The Contribution of Breeding to Yield Advances in maize (Zea mays L.), in: Advances in Agronomy. Elsevier, pp. 83–145. https://doi.org/10.1016/S0065-2113(05)86002-X

East, E., 1908. Inbreding in corn. Rep. Conn. Agric. Exp. Stn. Years 1097-1908 97–119.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6, e19379. https://doi.org/10.1371/journal.pone.0019379

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

Flint-Garcia, S.A., Thornsberry, J.M., Buckler, E.S., 2003. Structure of Linkage Disequilibrium in Plants. Annu. Rev. Plant Biol. 54, 357–374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Foll, M., Gaggiotti, O., 2008. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. Genetics 180, 977–993. https://doi.org/10.1534/genetics.108.092221

Francia, E., Morcia, C., Pasquariello, M., Mazzamurro, V., Milc, J.A., Rizza, F., Terzi, V., Pecchioni, N., 2016. Copy number variation at the HvCBF4–HvCBF2 genomic segment is a major component of frost resistance in barley. Plant Mol. Biol. 92, 161–175. https://doi.org/10.1007/s11103-016-0505-4

Frascaroli, E., Canè, M.A., Landi, P., Pea, G., Gianfranceschi, L., Villa, M., Morgante, M., Pè, M.E., 2007. Classical Genetic and Quantitative Trait Loci Analyses of Heterosis in a Maize Hybrid Between Two Elite Inbred Lines. Genetics 176, 625–644. https://doi.org/10.1534/genetics.106.064493

Frascaroli, E., Canè, M.A., Pè, M.E., Pea, G., Morgante, M., Landi, P., 2009. QTL detection in maize testcross progenies as affected by related and unrelated testers. Theor. Appl. Genet. 118, 993–1004. https://doi.org/10.1007/s00122-008-0956-3

Fu, H., Dooner, H.K., 2002. Intraspecific violation of genetic colinearity and its implications in maize. Proc. Natl. Acad. Sci. 99, 9573–9578.

Gabur, I., Chawla, H.S., Snowdon, R.J., Parkin, I.A.P., 2018. Connecting genome structural variation with complex traits in crop plants. Theor. Appl. Genet. https://doi.org/10.1007/s00122-018-3233-0

Gaines, T.A., Zhang, W., Wang, D., Bukun, B., Chisholm, S.T., Shaner, D.L., Nissen, S.J., Patzoldt, W.L., Tranel, P.J., Culpepper, A.S., Grey, T.L., Webster, T.M., Vencill, W.K., Sammons, R.D., Jiang, J., Preston, C., Leach, J.E., Westra, P., 2010. Gene amplification confers glyphosate resistance in Amaranthus palmeri. Proc. Natl. Acad. Sci. 107, 1029–1034. https://doi.org/10.1073/pnas.0906649107

Ganal, M.W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E.S., Charcosset, A., Clarke, J.D., Graner, E.-M., Hansen, M., Joets, J., Le Paslier, M.-C., McMullen, M.D., Montalent, P., Rose, M., Schön, C.-C., Sun, Q., Walter, H., Martin, O.C., Falque, M., 2011. A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. PLoS ONE 6, e28334. https://doi.org/10.1371/journal.pone.0028334

Giraud, H., Bauland, C., Falque, M., Madur, D., Combes, V., Jamin, P., Monteil, C., Laborde, J., Palaffre, C., Gaillard, A., Blanchard, P., Charcosset, A., Moreau, L., 2017a. Linkage Analysis and Association Mapping QTL Detection Models for Hybrids Between Multiparental Populations from Two Heterotic Groups: Application to Biomass Production in Maize (*Zea mays* L.). G3amp58 GenesGenomesGenetics 7, 3649–3657. https://doi.org/10.1534/g3.117.300121

Giraud, H., Bauland, C., Falque, M., Madur, D., Combes, V., Jamin, P., Monteil, C., Laborde, J., Palaffre, C., Gaillard, A., Blanchard, P., Charcosset, A., Moreau, L., 2017b. Reciprocal Genetics: Identifying QTL for General and Specific Combining Abilities in Hybrids Between Multiparental Populations from Two Maize (*Zea mays* L.) Heterotic Groups. Genetics 207, 1167–1180. https://doi.org/10.1534/genetics.117.300305

Giraud, H., Lehermeier, C., Bauer, E., Falque, M., Segura, V., Bauland, C., Camisan, C., Campo, L., Meyer, N., Ranc, N., Schipprack, W., Flament, P., Melchinger, A.E., Menz, M., Moreno-González, J., Ouzunova, M., Charcosset, A., Schön, C.-C., Moreau, L., 2014. Linkage Disequilibrium with Linkage Analysis of Multiline Crosses Reveals Different Multiallelic QTL for Hybrid Performance in the Flint and Dent Heterotic Groups of Maize. Genetics 198, 1717–1734. https://doi.org/10.1534/genetics.114.169367

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., Buckler, E.S., 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9, e90346. https://doi.org/10.1371/journal.pone.0090346

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. 108, 1513–1518. https://doi.org/10.1073/pnas.1017351108

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., Paterson, A.H., Pires, J.C., Sharpe, A.G., Tang, H., Teakle, G.R., Town, C.D., Batley, J., Edwards, D., 2016. The pangenome of an agronomically important crop plant Brassica oleracea. Nat. Commun. 7. https://doi.org/10.1038/ncomms13390

Gore, M.A., Chia, J.-M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J., 2009. A first-generation haplotype map of maize. Science 326, 1115–1117.

Gorjanc, G., Jenko, J., Hearne, S.J., Hickey, J.M., 2016. Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17. https://doi.org/10.1186/s12864-015-2345-z

Gouesnard, B., Negro, S., Laffray, A., Glaubitz, J., Melchinger, A., Revilla, P., Moreno-Gonzalez, J., Madur, D., Combes, V., Tollon-Cordet, C., Laborde, J., Kermarrec, D., Bauland, C., Moreau, L., Charcosset, A., Nicolas, S., 2017. Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. Theor. Appl. Genet. 130, 2165–2189. https://doi.org/10.1007/s00122-017-2949-6

Gouesnard, B., Rebourg, C., Welcker, C., Charcosset, A., 2012. Analysis of photoperiod sensitivity within a collection of tropical maize populations 12.

Grandbastien, M.-A., 2015. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim. Biophys. Acta BBA - Gene Regul. Mech. 1849, 403–416. https://doi.org/10.1016/j.bbagrm.2014.07.017

Gremme, G., Steinbiss, S., Kurtz, S., 2013. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. IEEE/ACM Trans. Comput. Biol. Bioinform. 10, 645–656. https://doi.org/10.1109/TCBB.2013.68

Gu, W., Zhang, F., Lupski, J.R., 2008. Mechanisms for human genomic rearrangements. PathoGenetics 1, 4. https://doi.org/10.1186/1755-8417-1-4

Gustafson, T.J., de Leon, N., Kaeppler, S.M., Tracy, W.F., 2018. Genetic Analysis of Resistance in the Wisconsin Diversity Panel of Maize. Crop Sci. 58, 1853. https://doi.org/10.2135/cropsci2017.11.0675

Hardigan, M.A., Crisovan, E., Hamilton, J.P., Kim, J., Laimbeer, P., Leisner, C.P., Manrique-Carpintero, N.C., Newton, L., Pham, G.M., Vaillancourt, B., Yang, X., Zeng, Z., Douches, D.S., Jiang, J., Veilleux, R.E., Buell, C.R., 2016. Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated *Solanum tuberosum*. Plant Cell 28, 388–405. https://doi.org/10.1105/tpc.15.00538

Hastings, P.J., Ira, G., Lupski, J.R., 2009. A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. PLoS Genet. 5, e1000327. https://doi.org/10.1371/journal.pgen.1000327

Hill, W.G., Weir, B.S., 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33, 54–78. https://doi.org/10.1016/0040-5809(88)90004-4

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Penagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., de Leon, N., Kaeppler, S.M., Buell, C.R., 2014. Insights into the Maize Pan-Genome and Pan-Transcriptome. Plant Cell 26, 121–135. https://doi.org/10.1105/tpc.113.119982

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., Fields, C.J., Wright, C.L., Koehler, K., Springer, N.M., Buckler, E., Buell, C.R., de Leon, N., Kaeppler, S.M., Childs, K.L., Mikel, M.A., 2016. Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. Plant Cell 28, 2700–2714. https://doi.org/10.1105/tpc.16.00353

Huang, X., Han, B., 2014. Natural Variations and Genome-Wide Association Studies in Crop Plants. Annu. Rev. Plant Biol. 65, 531–551. https://doi.org/10.1146/annurev-arplant-050213-035715

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., Ebert, D.P., Ostevik, K.L., Moyers, B.T., Yakimowski, S., Masalia, R.R., Gao, L., Ćalić, I., Bowers, J.E., Kane, N.C., Swanevelder, D.Z.H., Kubach, T., Muños, S., Langlade, N.B., Burke, J.M., Rieseberg, L.H., 2019. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat. Plants 5, 54–62. https://doi.org/10.1038/s41477-018-0329-0

Hull, F.H., 1946. Overdominance and Corn Breeding Where Hybrid Seed is Not Feasible. Agron. J. 38, 1100. https://doi.org/10.2134/agronj1946.00021962003800120007x

Hung, H.-Y., Shannon, L.M., Tian, F., Bradbury, P.J., Chen, C., Flint-Garcia, S.A., McMullen, M.D., Ware, D., Buckler, E.S., Doebley, J.F., Holland, J.B., 2012. ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proc. Natl. Acad. Sci. 109, E1913–E1921. https://doi.org/10.1073/pnas.1203189109

Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F., Barillot, E., 2004. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20, 3413–3422. https://doi.org/10.1093/bioinformatics/bth418

Hurgobin, B., Edwards, D., 2017. SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? Biology 6, 21. https://doi.org/10.3390/biology6010021

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.-K.K., Tirnaz, S., Dolatabadian, A., Schiessl, S.V., Samans, B., Montenegro, J.D., Parkin, I.A.P., Pires, J.C., Chalhoub, B., King, G.J., Snowdon, R., Batley, J., Edwards, D., 2018. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. Plant Biotechnol. J. 16, 1265–1274. https://doi.org/10.1111/pbi.12867

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K.L., Wolfgruber, T.K., May, M.R., Springer, N.M., Antoniou, E., McCombie, W.R., Presting, G.G., McMullen, M., Ross-Ibarra, J., Dawe, R.K., Hastie, A., Rank, D.R., Ware, D., 2017. Improved maize reference genome with single-molecule technologies. Nature. https://doi.org/10.1038/nature22971

Jones, D.F., 1917. Dominance of Linked Factors As a Means of Accounting For Heterosis. Genetics 2, 466.

Khotyleva, L.V., Kilchevsky, A.V., Shapturenko, M.N., 2017. Theoretical aspects of heterosis. Russ. J. Genet. Appl. Res. 7, 428–439. https://doi.org/10.1134/S2079059717040049

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al., 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56–64. https://doi.org/10.1038/nature06862

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C.E., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M., Snyder, M., 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318, 420. https://doi.org/10.1126/science.1149504

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., 2010. Genome-wide patterns of genetic variation among elite maize inbred lines. Nat. Genet. 42, 1027.

Lai, J., Li, Y., Messing, J., Dooner, H.K., 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc. Natl. Acad. Sci. 102, 9068–9073. https://doi.org/10.1073/pnas.0502923102

Lander, S., Botstein, D., 1989. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps 15.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A., Freeling, M., 2004. Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. Genetics 166, 935–945. https://doi.org/10.1534/genetics.166.2.935

Laporte, F., 2018. Développement de méthodes statistiques pour l'identification de gènes d'intérêt en présence d'apparentement et de dominance, application à la génétique du maïs. Université Paris-Sud.

Laporte, F., Charcosset, A., Mary-Huard, T., 2019. Efficient Reml inference in variance component mixed-model using a min-max algorithms. Manuscr. Submitt.

Larièpe, A., Mangin, B., Jasson, S., Combes, V., Dumas, F., Jamin, P., Lariagon, C., Jolivot, D., Madur, D., Fiévet, J., Gallais, A., Dubreuil, P., Charcosset, A., Moreau, L., 2012. The Genetic Basis of Heterosis: Multiparental Quantitative Trait Loci Mapping Reveals Contrasted Levels of Apparent Overdominance Among Traits of Agronomical Interest in Maize (*Zea mays* L.). Genetics 190, 795–811. https://doi.org/10.1534/genetics.111.133447

Larièpe, A., Moreau, L., Laborde, J., Bauland, C., Mezmouk, S., Décousset, L., Mary-Huard, T., Fiévet, J.B., Gallais, A., Dubreuil, P., Charcosset, A., 2017. General and specific combining abilities in a maize (Zea mays L.) test-cross hybrid panel: relative importance of population structure and genetic divergence between parents. Theor. Appl. Genet. 130, 403–417. https://doi.org/10.1007/s00122-016-2822-z

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., Liu, J., Warburton, M.L., Cheng, Y., Hao, X., Zhang, P., Zhao, J., Liu, Y., Wang, G., Li, J., Yan, J., 2013. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. 45, 43–50. https://doi.org/10.1038/ng.2484

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D., 2011. FaST linear mixed models for genome-wide association studies. Nat. Methods 8, 833–835. https://doi.org/10.1038/nmeth.1681

Lisch, D., 2013. How important are transposons for plant evolution? Nat. Rev. Genet. 14, 49–61. https://doi.org/10.1038/nrg3374

Liu, J., Qu, J., Yang, C., Tang, D., Li, J., Lan, H., Rong, T., 2015. Development of genome-wide insertion and deletion markers for maize, based on next-generation sequencing data. BMC Genomics 16. https://doi.org/10.1186/s12864-015-1797-5

Liu, S., Ying, K., Yeh, C.-T., Yang, J., Swanson-Wagner, R., Wu, W., Richmond, T., Gerhardt, D.J., Lai, J., Springer, N., Nettleton, D., Jeddeloh, J.A., Schnable, P.S., 2012. Changes in genome content generated via segregation of non-allelic homologs: *Segregation of non-allelic homologs*. Plant J. 72, 390–399. https://doi.org/10.1111/j.1365-313X.2012.05087.x

Lu, F., Romay, M.C., Glaubitz, J.C., Bradbury, P.J., Elshire, R.J., Wang, T., Li, Yu, Li, Yongxiang, Semagn, K., Zhang, X., Hernandez, A.G., Mikel, M.A., Soifer, I., Barad, O., Buckler, E.S., 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. Nat. Commun. 6. https://doi.org/10.1038/ncomms7914

Lu, H., Romero-Severson, J., Bernardo, R., 2003. Genetic basis of heterosis explored by simple sequence repeat markers in a random-mated maize population. Theor. Appl. Genet. 107, 494–502. https://doi.org/10.1007/s00122-003-1271-7

Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A.J., Li, T., Ma, H., 2012. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome Res. 22, 508–518. https://doi.org/10.1101/gr.127522.111

Lye, Z.N., Purugganan, M.D., 2019. Copy Number Variation in Domestication. Trends Plant Sci. https://doi.org/10.1016/j.tplants.2019.01.003

Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits, Sinauer Associates. ed. Sunderland MA.

Lyra, D.H., Galli, G., Alves, F.C., Granato, Í.S.C., Vidotti, M.S., Bandeira e Sousa, M., Morosini, J.S., Crossa, J., Fritsche-Neto, R., 2018. Modeling copy number variation in the genomic prediction of maize hybrids. Theor. Appl. Genet. https://doi.org/10.1007/s00122-018-3215-2

Mace, E.S., Tai, S., Gilding, E.K., Li, Y., Prentis, P.J., Bian, L., Campbell, B.C., Hu, W., Innes, D.J., Han, X., Cruickshank, A., Dai, C., Frère, C., Zhang, H., Hunt, C.H., Wang, X., Shatte, T., Wang, M., Su, Z., Li, J., Lin, X., Godwin, I.D., Jordan, D.R., Wang, J., 2013. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat. Commun. 4. https://doi.org/10.1038/ncomms3320

Maenhout, S., De Baets, B., Haesaert, G., 2010. Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. Theor. Appl. Genet. 120, 415–427. https://doi.org/10.1007/s00122-009-1200-5

Makarevitch, I., Waters, A.J., West, P.T., Stitzer, M., Hirsch, C.N., Ross-Ibarra, J., Springer, N.M., 2015. Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. PLoS Genet. 11, e1004915. https://doi.org/10.1371/journal.pgen.1004915

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., Cierco-Ayrolles, C., 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 108, 285–291. https://doi.org/10.1038/hdy.2011.73

Manicacci, D., Camus-Kulandaivelu, L., Fourmann, M., Arar, C., Barrault, S., Rousselet, A., Feminias, N., Consoli, L., Frances, L., Mechin, V., Murigneux, A., Prioul, J.-L., Charcosset, A., Damerval, C., 2009. Epistatic Interactions between Opaque2 Transcriptional Activator and Its Target Gene CyPPDK1 Control Kernel Trait Variation in Maize. PLANT Physiol. 150, 506–520. https://doi.org/10.1104/pp.108.131888

Marioni, J.C., Thorne, N.P., Tavare, S., 2006. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 22, 1144–1146. https://doi.org/10.1093/bioinformatics/btl089

Maron, L.G., Guimarães, C.T., Kirst, M., Albert, P.S., Birchler, J.A., Bradbury, P.J., Buckler, E.S., Coluccio, A.E., Danilova, T.V., Kudrna, D., 2013. Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl. Acad. Sci. 110, 5241–5246.

Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez G., J., Buckler, E., Doebley, J., 2002. A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. 99, 6080–6084. https://doi.org/10.1073/pnas.052125199

McClintock, B., 1950. The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. 36, 344–355. https://doi.org/10.1073/pnas.36.6.344

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J.A., Stupar, R.M., 2012. Structural Variants in the Soybean Genome Localize to Clusters of Biotic Stress-Response Genes. PLANT Physiol. 159, 1295–1308. https://doi.org/10.1104/pp.112.194605

McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S.E., Peterson, B., Pressoir, G., Romero, S., Rosas, M.O., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J.C., Goodman, M., Ware, D., Holland, J.B., Buckler, E.S., 2009. Genetic Properties of the Maize Nested Association Mapping Population. Science 325, 737–740. https://doi.org/10.1126/science.1174320

Millet, E., Welcker, C., Kruijer, W., Negro, S., Nicolas, S., Praud, S., Ranc, N., Presterl, T., Tuberosa, R., Bedo, Z., Draye, X., Usadel, B., Charcosset, A., van Eeuwijk, F., Tardieu, F., Coupel-Ledru, A., Bauland, C., 2016. Genome-wide analysis of yield in Europe: allelic effects as functions of drought and heat scenarios. Plant Physiol. pp.00621.2016. https://doi.org/10.1104/pp.16.00621

Mir, C., Zerjal, T., Combes, V., Dumas, F., Madur, D., Bedoya, C., Dreisigacker, S., Franco, J., Grudloyma, P., Hao, P.X., Hearne, S., Jampatong, C., Laloë, D., Muthamia, Z., Nguyen, T., Prasanna, B.M., Taba, S., Xie, C.X., Yunus, M., Zhang, S., Warburton, M.L., Charcosset, A., 2013. Out of America: tracing the genetic footprints of the global diffusion of maize. Theor. Appl. Genet. 126, 2671–2682. https://doi.org/10.1007/s00122-013-2164-z

Monat, C., Tranchand, C., Engelen, S., Labadie, K., Paradis, E., Tando, N., Sabot, F., 2018. Comparison of two African rice species through a new pan-genomic approach on massive data. bioRxiv. https://doi.org/10.1101/245431

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chan, C.-K.K., Visendi, P., Lai, K., Doležel, J., Batley, J., Edwards, D., 2017. The pangenome of hexaploid bread wheat. Plant J. 90, 1007–1013. https://doi.org/10.1111/tpj.13515

Morgante, M., Depaoli, E., Radovic, S., 2007. Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. 10, 149–155. https://doi.org/10.1016/j.pbi.2007.02.001

Muñoz-Amatriaín, M., Eichten, S.R., Wicker, T., Richmond, T.A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., Mayer, K.F., Taudien, S., Platzer, M., Jeddeloh, J.A., Springer, N.M., Muehlbauer, G.J., Stein, N., 2013. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol. 14. https://doi.org/10.1186/gb-2013-14-6-r58

Negro, S.S., Millet, E., Madur, D., Bauland, C., Combes, V., Welcker, C., Tardieu, F., Charcosset, A., Nicolas, S.D., 2018. Genotyping-by-sequencing and microarrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. bioRxiv. https://doi.org/10.1101/476598

Nei, M., 1973. Analysis of Gene Diversity in Subdivided Populations. Proc. Natl. Acad. Sci. 70, 3321–3323.

Nicolas, S.D., Péros, J.-P., Lacombe, T., Launay, A., Le Paslier, M.-C., Bérard, A., Mangin, B., Valière, S., Martins, F., Le Cunff, L., Laucou, V., Bacilieri, R., Dereeper, A., Chatelet, P., This, P., Doligez, A., 2016. Genetic diversity, linkage disequilibrium and power of a large grapevine (Vitis vinifera L) diversity panel newly designed for association studies. BMC Plant Biol. 16. https://doi.org/10.1186/s12870-016-0754-z

Nielsen, 2005. Molecular Signatures of Natural Selection. Annu. Rev. Genet. 39, 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420

Nitcher, R., Distelfeld, A., Tan, C., Yan, L., Dubcovsky, J., 2013. Increased copy number at the HvFT1 locus is associated with accelerated flowering time in barley. Mol. Genet. Genomics 288, 261–275. https://doi.org/10.1007/s00438-013-0746-8

Nordborg, M., Innan, H., 2002. Molecular population genetics. Curr. Opin. Plant Biol. 5, 69–73.

Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M., 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557–572. https://doi.org/10.1093/biostatistics/kxh008

Owens, G.L., Baute, G.J., Hubner, S., Rieseberg, L.H., 2018. Genomic sequence and copy number evolution during hybrid crop development in sunflowers. Evol. Appl. https://doi.org/10.1111/eva.12603

Pace, J., Gardner, C., Romay, C., Ganapathysubramanian, B., Lübberstedt, T., 2015. Genome-wide association analysis of seedling root development in maize (Zea mays L.). BMC Genomics 16. https://doi.org/10.1186/s12864-015-1226-9

Pascual, L., Desplat, N., Huang, B.E., Desgroux, A., Bruguier, L., Bouchet, J.-P., Le, Q.H., Chauchard, B., Verschave, P., Causse, M., 2015. Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. Plant Biotechnol. J. 13, 565–577. https://doi.org/10.1111/pbi.12282

Paterson, A.H., Lander, E.S., Hewitt, J.D., Peterson, S., Lincoln, S.E., Tanksley, S.D., 1988. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335, 721–726. https://doi.org/10.1038/335721a0

Peiffer, J.A., Romay, M.C., Gore, M.A., Flint-Garcia, S.A., Zhang, Z., Millard, M.J., Gardner, C.A.C., McMullen, M.D., Holland, J.B., Bradbury, P.J., Buckler, E.S., 2014. The Genetic Architecture Of Maize Height. Genetics 196, 1337–1356. https://doi.org/10.1534/genetics.113.159152

Piaskowski, J., Hardner, C., Cai, L., Zhao, Y., Iezzoni, A., Peace, C., 2018. Genomic heritability estimates in sweet cherry reveal non-additive genetic variance is relevant for industry-prioritized traits. BMC Genet. 19. https://doi.org/10.1186/s12863-018-0609-8

Picard, F., Robin, S., Lavielle, M., Vaisse, C., Daudin, J.-J., 2005. A statistical approach for array CGH data analysis. BMC Bioinformatics 14.

Picard, F., Robin, S., Lebarbier, E., Daudin, J.-J., 2007. A Segmentation/Clustering Model for the Analysis of Array CGH Data. Biometrics 63, 758–766. https://doi.org/10.1111/j.1541-0420.2006.00729.x

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B., Gray, J.W., Albertson, D.G., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. 20, 207–211. https://doi.org/10.1038/2524

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M.C., Zaina, G., Bastien, C., Cattonaro, F., Marroni, F., Morgante, M., 2016. Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. Mol. Biol. Evol. 33, 2706–2719. https://doi.org/10.1093/molbev/msw161

Pires, J.C., Zhao, J., Schranz, M.E., Leon, E.J., Quijada, P.A., Lukens, L.N., Osborn, T.C., 2004. Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (Brassicaceae): NOVEL VARIATION IN RESYNTHESIZED BRASSICA POLYPLOIDS. Biol. J. Linn. Soc. 82, 675–688. https://doi.org/10.1111/j.1095-8312.2004.00350.x

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

Rafalski, A., 2002. Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. 5, 94–100. https://doi.org/10.1016/S1369-5266(02)00240-6

Rebourg, C., Chastanet, M., Gouesnard, B., Welcker, C., Dubreuil, P., Charcosset, A., 2003. Maize introduction into Europe: the history reviewed in the light of molecular data. Theor. Appl. Genet. 106, 895–903. https://doi.org/10.1007/s00122-002-1140-9

Rieseberg, L.H., 2001. Chromosomal rearrangements and speciation. Trends Ecol. Evol. 16, 351–358.

Rincent, R., Moreau, L., Monod, H., Kuhn, E., Melchinger, A.E., Malvar, R.A., Moreno-Gonzalez, J., Nicolas, S., Madur, D., Combes, V., Dumas, F., Altmann, T., Brunel, D., Ouzunova, M., Flament, P., Dubreuil, P., Charcosset, A., Mary-Huard, T., 2014. Recovering Power in Association Mapping Panels with Variable Levels of Linkage Disequilibrium. Genetics 197, 375–387. https://doi.org/10.1534/genetics.113.159731

Rio, S., Mary-Huard, T., Moreau, L., Bauland, C., Carine, P., Delphine, M., Valéries, C., Charcosset, A., 2019. Disentangling group specific group QTL allele effects from genetic background epistasis using admixed individuals in GWAS: an application to maize flowering. Prep.

Romay, M.C., Millard, M.J., Glaubitz, J.C., Peiffer, J.A., Swarts, K.L., Casstevens, T.M., Elshire, R.J., Acharya, C.B., Mitchell, S.E., Flint-Garcia, S.A., McMullen, M.D., Holland, J.B., Buckler, E.S., Gardner, C.A., 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14. https://doi.org/10.1186/gb-2013-14-6-r55

Romero Navarro, J.A., Willcox, M., Burgueño, J., Romay, C., Swarts, K., Trachsel, S., Preciado, E., Terron, A., Delgado, H.V., Vidal, V., Ortega, A., Banda, A.E., Montiel, N.O.G., Ortiz-Monasterio, I., Vicente, F.S., Espinoza, A.G., Atlin, G., Wenzl, P., Hearne, S., Buckler, E.S., 2017. A study of allelic diversity underlying flowering-time adaptation in maize landraces. Nat. Genet. 49, 476–480. https://doi.org/10.1038/ng.3784

Saintenac, C., Jiang, D., Akhunov, E.D., 2011. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. 12, R88.

Salvi, S., Corneti, S., Bellotti, M., Carraro, N., Sanguineti, M.C., Castelletti, S., Tuberosa, R., 2011. Genetic dissection of maize phenology using an intraspecific introgression library. BMC Plant Biol. 11, 4. https://doi.org/10.1186/1471-2229-11-4

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X., Fengler, K.A., Meeley, R., Ananiev, E.V., Svitashev, S., Bruggemann, E., Li, B., Hainey, C.F., Radovic, S., Zaina, G., Rafalski, J.-A., Tingey, S.V., Miao, G.-H., Phillips, R.L., Tuberosa, R., 2007. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. 104, 11376–11381. https://doi.org/10.1073/pnas.0704145104

Salvi, S., Tuberosa, R., Chiapparino, E., Maccaferri, M., Veillet, S., van Beuningen, L., Isaac, P., Edwards, K., Phillips, R.L., 2002. Toward positional cloning of Vgt1, a QTL controlling the transition from the vegetative to the reproductive phase in maize 13.

Saxena, R.K., Edwards, D., Varshney, R.K., 2014. Structural variations in plant genomes. Brief. Funct. Genomics 13, 296–307. https://doi.org/10.1093/bfgp/elu016

Schnable, J.C., Springer, N.M., Freeling, M., 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. 108, 4069–4074. https://doi.org/10.1073/pnas.1101368108

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al., 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326, 1112–1115. https://doi.org/10.1126/science.1178534

Schön, C.C., Dhillon, B.S., Utz, H.F., Melchinger, A.E., 2010. High congruency of QTL positions for heterosis of grain yield in three crosses of maize. Theor. Appl. Genet. 120, 321–332. https://doi.org/10.1007/s00122-009-1209-9

Shen, X., Liu, Z.-Q., Mocoeur, A., Xia, Y., Jing, H.-C., 2015. PAV markers in Sorghum bicolour: genome pattern, affected genes and pathways, and genetic linkage map construction. Theor. Appl. Genet. 128, 623–637. https://doi.org/10.1007/s00122-015-2458-4

Shull, G.H., 1914. Duplicate genes for capsule-form in B ¢rsa bursa-pastoris 1). 53.

Shull, G.H., 1908. The Composition of a Field of Maize. J. Hered. os-4, 296–301. https://doi.org/10.1093/jhered/os-4.1.296

Sprague, F., Tatum, L.A., 1942. GENERALVS. SPECIFIC COMBININGABILITY IN SINGLE CROSSES OF CORN1 o 10.

Springer, N.M., Stupar, R.M., 2007. Allelic variation and heterosis in maize: How do two halves make more than a whole? Genome Res. 17, 264–275. https://doi.org/10.1101/gr.5347007

Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A.L., Barbazuk, W.B., Jeddeloh, J.A., Nettleton, D., Schnable, P.S., 2009. Maize Inbreds

Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet. 5, e1000734. https://doi.org/10.1371/journal.pgen.1000734

Stankiewicz, P., Lupski, J.R., 2010. Structural Variation in the Human Genome and its Role in Disease. Annu. Rev. Med. 61, 437–455. https://doi.org/10.1146/annurev-med-100708-204735

Stjernqvist, S., Rydén, T., Sköld, M., Staaf, J., 2007. Continuous-index hidden Markov modelling of array CGH copy number data. Bioinformatics 23, 1006–1014. https://doi.org/10.1093/bioinformatics/btm059

Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjaris, T., Lander, E.S., 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 823–839.

Sturtevant, A.H., 1924. The Effects of unequal crossing over at the bar locus in drosophilia. Genetics 10, 117.

Su, G., Christensen, O.F., Ostersen, T., Henryon, M., Lund, M.S., 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLoS ONE 7, e45293. https://doi.org/10.1371/journal.pone.0045293

Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, Haiming, Zhao, Hainan, Song, W., Zhang, M., Cui, Y., Dong, X., Liu, H., Ma, X., Jiao, Y., Wang, B., Wei, X., Stein, J.C., Glaubitz, J.C., Lu, F., Yu, G., Liang, C., Fengler, K., Li, B., Rafalski, A., Schnable, P.S., Ware, D.H., Buckler, E.S., Lai, J., 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat. Genet. 50, 1289–1295. https://doi.org/10.1038/s41588-018-0182-0

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D., Springer, N.M., 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 20, 1689–1699. https://doi.org/10.1101/gr.109165.110

Swarts, K., Gutaker, R.M., Benz, B., Blake, M., Bukowski, R., Holland, J., Kruse-Peeples, M., Lepak, N., Prim, L., Romay, M.C., Ross-Ibarra, J., Sanchez-Gonzalez, J. de J., Schmidt, C., Schuenemann, V.J., Krause, J., Matson, R.G., Weigel, D., Buckler, E.S., Burbano, H.A., 2017. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science 357, 512–515. https://doi.org/10.1126/science.aam9425

Tai, T.H., Tanksley, S.D., 1990. A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue. Plant Mol. Biol. Report. 8, 297–303. https://doi.org/10.1007/BF02668766

Takuno, S., Ralph, P., Swarts, K., Elshire, R.J., Glaubitz, J.C., Buckler, E.S., Hufford, M.B., Ross-Ibarra, J., 2015. Independent Molecular Basis of Convergent Highland Adaptation in Maize 29.

Technow, F., Riedelsheimer, C., Schrag, T.A., Melchinger, A.E., 2012. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor. Appl. Genet. 125, 1181–1194. https://doi.org/10.1007/s00122-012-1905-8

Tenaillon, M.I., Charcosset, A., 2011. A European perspective on maize history. C. R. Biol. 334, 221–228. https://doi.org/10.1016/j.crvi.2010.12.015

Tenaillon, M.I., Hufford, M.B., Gaut, B.S., Ross-Ibarra, J., 2011. Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and Zea luxurians. Genome Biol. Evol. 3, 219–229. https://doi.org/10.1093/gbe/evr008

Thomas, B.C., 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16, 934–946. https://doi.org/10.1101/gr.4708406

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler, E.S., 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. 28, 286–289. https://doi.org/10.1038/90135

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B., Buckler, E.S., 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. 43, 159–162. https://doi.org/10.1038/ng.746

Toro, M.A., Varona, L., 2010. A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol. 42. https://doi.org/10.1186/1297-9686-42-33

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M.V., Eichler, E.E., 2005. Fine-scale structural variation of the human genome. Nat. Genet. 37, 727–732. https://doi.org/10.1038/ng1562

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T., Strom, T.M., Fries, R., Pausch, H., Bertani, C., Davassi, A., Mayer, K.F., Schön, C.-C., 2014. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15, 823. https://doi.org/10.1186/1471-2164-15-823

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T., Strom, T.M., Fries, R., Pausch, H., Bertani, C., Davassi, A., Mayer, K.F., Schön, C.-C., 2014b. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15, 823. https://doi.org/10.1186/1471-2164-15-823

Unterseer, S., Pophaly, S.D., Peis, R., Westermeier, P., Mayer, M., Seidel, M.A., Haberer, G., Mayer, K.F.X., Ordas, B., Pausch, H., Tellier, A., Bauer, E., Schön, C.-C., 2016. A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. Genome Biol. 17. https://doi.org/10.1186/s13059-016-1009-x

Unterseer, S., Seidel, M.A., Bauer, E., Haberer, G., Hochholdinger, F., Opitz, N., Marcon, C., Baruch, K., Spannagl, M., Mayer, K.F., 2017. European Flint reference sequences complement the maize pan-genome. bioRxiv 103747.

van Heerwaarden, J., Hufford, M.B., Ross-Ibarra, J., 2012. Historical genomics of North American maize. Proc. Natl. Acad. Sci. 109, 12420–12425. https://doi.org/10.1073/pnas.1209275109

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Varona, L., Legarra, A., Toro, M.A., Vitezica, Z.G., 2018. Non-additive Effects in Genomic Selection. Front. Genet. 9. https://doi.org/10.3389/fgene.2018.00078

Varshney, R.K., Saxena, R.K., Upadhyaya, H.D., Khan, A.W., Yu, Y., Kim, C., Rathore, A., Kim, D., Kim, J., An, S., Kumar, V., Anuradha, G., Yamini, K.N., Zhang, W., Muniswamy, S., Kim, J.-S., Penmetsa, R.V., von Wettberg, E., Datta, S.K., 2017. Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. Nat. Genet. 49, 1082–1088. https://doi.org/10.1038/ng.3872

Vitezica, Z.G., Varona, L., Elsen, J.-M., Misztal, I., Herring, W., Legarra, A., 2016. Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genet. Sel. Evol. 48. https://doi.org/10.1186/s12711-016-0185-1

Vitezica, Z.G., Varona, L., Legarra, A., 2013. On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics 195, 1223–1230. https://doi.org/10.1534/genetics.113.155176

Wallace, J.G., Bradbury, P.J., Zhang, N., Gibon, Y., Stitt, M., Buckler, E.S., 2014. Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize. PLoS Genet. 10, e1004845. https://doi.org/10.1371/journal.pgen.1004845

Wang, X., Lebarbier, E., Aubert, J., Robin, S., 2017. Variational inference for coupled Hidden Markov Models applied to the joint detection of copy number variations. ArXiv170606742 Stat.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M., 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35, 543–548. https://doi.org/10.1093/molbev/msx319

Xiang, T., Christensen, O.F., Vitezica, Z.G., Legarra, A., 2016. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet. Sel. Evol. 48. https://doi.org/10.1186/s12711-016-0271-4

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., Li, J., He, W., Zhang, G., Zheng, X., Zhang, F., Li, Y., Yu, C., Kristiansen, K., Zhang, X., Wang, Jian, Wright, M., McCouch, S., Nielsen, R., Wang, Jun, Wang, W., 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. 30, 105–111. https://doi.org/10.1038/nbt.2050

Yang, J., Mezmouk, S., Baumgarten, A., Buckler, E.S., Guill, K.E., McMullen, M.D., Mumm, R.H., Ross-Ibarra, J., 2017. Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLOS Genet. 13, e1007019. https://doi.org/10.1371/journal.pgen.1007019

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z., 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871. https://doi.org/10.1093/bioinformatics/btp394

Yu, J., Holland, J.B., McMullen, M.D., Buckler, E.S., 2008. Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics 178, 539–551. https://doi.org/10.1534/genetics.107.074245

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S., 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203–208. https://doi.org/10.1038/ng1702

Zerjal, T., Rousselet, A., Mhiri, C., Combes, V., Madur, D., Grandbastien, M.-A., Charcosset, A., Tenaillon, M.I., 2012. Maize genetic diversity and association mapping using transposable element insertion polymorphisms. Theor. Appl. Genet. 124, 1521–1537. https://doi.org/10.1007/s00122-012-1807-9

Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M.A., San Vicente, F., Olsen, M., Buckler, E., Jannink, J.-L., Prasanna, B.M., Crossa, J., 2015. Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 114, 291–299. https://doi.org/10.1038/hdy.2014.99

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., Xu, Q., Wang, Z.-X., Wei, X., Han, B., Huang, X., 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. https://doi.org/10.1038/s41588-018-0041-z

Zhou, P., Silverstein, K.A.T., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., Farmer, A.D., Steele, K.P., Stupar, R.M., Miller, J.R., Tiffin, P., Mudge, J., Young, N.D., 2017. Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. BMC Genomics 18. https://doi.org/10.1186/s12864-017-3654-1

Zirkle, 1952. Early ideas on inbreeding and cross-breeding, Iowa State College Press. ed.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Title: Contribution of Insertions / Deletions-type Structural Variations to Adaptation, Phenotypic Variation and Hybrid Performances in Maize

Keywords: Maize, Structural Variation, Insertions Deletions, GWAS, genotyping

Abstract: In the last decades, the rapid development of genome sequencing allowed to identify structural variations in many species. In maize, thousands of large insertions and deletions (InDels) from few bp to hundreds of Kbp were discovered by comparing the reference genome B73 and many other resequenced genomes. These InDel sequences can carry genes and therefore be involved in phenotypic variation by changing the gene composition between individuals, but their effect on the phenotype was not well studied. The aim of this thesis was to study the contribution of InDels to adaptation, phenotypic variations and hybrid performances in maize.

We developed an Affymetrix® Axiom® genotyping array that allowed to genotype 105,947 InDels sequences ranging from 35bp to 129,7Kbp of size. 79,969 sequences of these InDels were not present in B73 reference genome and have been discovered by assembling three genomes (F2, C103, and PH207).

We selected 61,492 polymorphic InDels to genotype a 362 maize inbred lines panel representing a broad range of diversity to study the contribution of InDels to genetic diversity, adaptation and trait variation. We also assembled one million of SNPs from two genotyping arrays and genotyping by sequencing to study the complementarity between InDels and SNPs. Genetic structuration and relatedness between inbred lines displayed by SNPs or by InDels were highly similar suggesting that almost all InDels and SNPs followed a similar evolutionary trajectory. 51% of InDels were not in high linkage disequilibrium (LD>0.8) with any nearby SNP suggesting that the effect of these InDels was not be well captured using this density of SNP. Thanks to InDels, we detected 13 new quantitative trait loci (QTLs) among 294 QTLs identified for 23 traits by a genome wide association studies (GWAS). Similarly, 56 out 188 regions under selection between tropical, dent and flint maize lines were identified by InDels leading to an enrichment of genomic regions under selection detected by InDels compared to SNPs. These InDels include genes involved in tolerance to biotic and abiotic stress and/or adaptive traits as flowering time. Accordingly, the highest number of associated InDels was found for flowering time. These results suggest that InDels were often involved in adaptation and stress tolerance.

In order to study the effect of InDels on hybrid performances, we analyzed a panel of 287 hybrids derived from the crossing of 210 maize temperate inbred lines from the previous panel. We decomposed the variance of female flowering (FF), plant height (PH) and grain yield (GY) by distinguishing the additive and dominant genetic effects. We observed the highest dominance and genotype by environment effects for GY and the lowest for FF. We performed GWAS on this panel by testing additive and dominance effects of 51,844 InDels and 469,267 SNPs on these three traits in 4 different environment combinations. We identified 78 and 133 QTLs with an additive and dominance effect, respectively including 6 and 11 QTLs discovered only by InDels. 83% of all QTLs were found with only one environment combination. One QTL for GY detected with InDels was located in a large cluster of InDels on chromosome 6, previously identified to have a strong effect on GY in heat conditions. We finally used InDels and/or SNPs genotyping to predict hybrid performances. Whereas including a dominance effect in genomic prediction models increased by 1.5 to 5.6% predictive abilities (PA) for GY, including InDels genotyping did not increased PA.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Contribution des variations structurales de type insertions/délétions sur l'adaptation, la variation des caractères et les performances hybrides chez le maïs

Mots clés : Maïs, Variations Structurales, Insertions Délétions, génétique d'association, génotypage

Résumé : Le récent développement des méthodes de séquençage permet aujourd'hui d'identifier des variations structurales chez de nombreuses espèces. Chez le maïs, des milliers de grandes insertions et délétions (InDel) de quelques pb à plusieurs centaines de Kbp ont été découvertes entre le génome de référence B73 et de nombreux autres génomes reséquencés. Ces InDel peuvent changer la composition des gènes entre les individus et donc être impliquées dans la variation du phénotype, mais cet effet sur le phénotype reste mal connu. L'objectif de cette thèse était d'étudier la contribution des InDel à l'adaptation, aux variations phénotypiques et aux performances hybrides chez le maïs.

Nous avons développé une puce de génotypage des InDel Affymetrix® Axiom® capable de génotyper 105 927 InDel de 35bp à 129,7Kbp. Parmi ces InDel, 79 969 ont leur séquences absentes du génome de référence B73 et ont été identifiées par l'assemblage 3 génomes (F2, C103, and PH207).

Nous avons sélectionné 61 492 InDel polymorphiques pour génotyper 362 lignées de maïs représentant une large gamme de diversité pour étudier la contribution des InDel à la diversité génétique, l'adaptation et la variation des caractères. Nous avons également génotypé 1 million de SNP à partir de deux puces de génotypage et du génotypage par séquençage pour étudier la complémentarité entre les InDel et les SNP. Qu'ils soient calculés avec les InDel ou les SNP la structuration génétique et les valeurs d'apparentement entre les lignées sont très similaires, ce qui suggère que la plupart des InDel ont suivi la même trajectoire évolutive que les SNP. 51% des InDel ne sont pas en déséquilibre de liaison élevé (>0.8) avec aucun SNP proche donc l'effet de ces InDel n'est donc a priori pas capturé pas des SNP à cette densité. Parmi les 294 régions génomiques associées au phénotype (QTL), 13 nouveaux QTL ont été détectés grâce aux InDel par rapport aux SNP par une approche de génétique d'association (GA). Nous avons détecté un enrichissement en InDel sous sélection entre les lignées tropicales, cornées et dentées par rapport aux SNP, avec 56 sur 188 régions sous sélection détectées avec les InDel. Ces régions contiennent des gènes impliqués dans l'adaptation et/ou la tolérance aux stress. De plus, le plus grand nombre d'associations a été découvert pour la floraison, caractère adaptatif chez le maïs. Ces résultats suggèrent que les InDel sont plus souvent impliquées dans l'adaptation et la tolérance aux stress.

Nous avons enfin testé l'effet des InDel sur les performances des hybrides en analysant un panel de 287 hybrides issus du croisement de 210 lignées tempérées du panel précédent. Nous avons décomposé la variance des performances hybrides en distinguant les effets de dominance et d'additivité pour la floraison femelle (FF), la hauteur (PH) et le rendement (GY). La plus forte part de dominance et d'interaction génotype-environnement a été observée pour le GY et la plus faible pour la FF. Les effets d'additivité et de dominance de 51,844 InDel et 469 267 SNP a été testés pour 4 combinaisons d'environnements par une approche de GA. 78 et 133 QTL avec un effet additif et dominant respectivement ont été identifiés, dont 6 et 11 avec des InDel. 83% de ces QTL ont été identifiés dans une seule combinaison d'environnements. Un des QTL de rendement identifié avec des InDel est situé dans un large cluster d'InDel sur le chromosome 6 et colocalise avec un QTL déjà identifié avec des SNP avec un effet fort dans l'augmentation du rendement sous des températures élevées. L'ajout de l'effet de dominance en plus de l'effet additif permet d'augmenter la précision des prédictions génomiques jusqu'à 5,6% pour le rendement. Cependant, l'ajout du génotypage des InDel en plus de celui des SNP n'a pas permis d'améliorer les prédictions des phénotypes hybrides.