Characterizing the genomic determinants and phenotypic responses to altitudinal adaptation in teosintes (Zea mays ssp. parviglumis and ssp. mexicana)

Natalia Elena Martínez Ainsworth

université PARIS-SACLAY

NNT : 2019SACLS376

Thèse de doctorat

Characterizing the genomic determinants and phenotypic responses to altitudinal adaptation in teosintes (*Zea mays* ssp. *parviglumis* and ssp. *mexicana*)

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud
UMR de Génétique Quantitative et Evolution – Le Moulon

École doctorale n°567 Sciences du Végétal : du gène à l'écosystème

Spécialité de doctorat : Biologie

Thèse présentée et soutenue le 25 octobre 2019, à Gif-sur-Yvette, par

Natalia Elena MARTINEZ AINSWORTH

Composition du Jury :

Jacqui SHYKOFF Directrice de recherche, CNRS, UPS (ESE)	Présidente du Jury
Joëlle RONFORT Directrice de recherche, INRA (UMR AGAP)	Rapportrice
Cristina VIEIRA Directrice de recherche, CNRS (UCBL1, UMR 1118)	Rapportrice
Valérie LE CORRE Chargée de recherche, INRA (Dijon)	Examinatrice
Maud TENAILLON Directrice de recherche, CNRS (UMR GQE-Le Moulon)	Directrice de thèse
Domenica MANICACCI (co-encadrante) Maître de conférences, UPS (UMR GQE-Le Moulon)	Invitée
Luis E. EGUIARTE Directeur de recherche, UNAM (Instituto de Ecología)	Invité

ACKNOWLEDGEMENTS

I am immensely thankful to Maud Tenaillon and Domenica Manicacci for the passion and involvement that they devoted to this work. Thank you Maud for being there and helping me reflect throughout this academic adventure. I have learnt a lot from your dynamism and broad interest in evolutionary questions. Your constant encouragements for meeting stunning scientists and attending awesome conferences has greatly enriched my researcher mind. Evolmol sessions surely kept me amazed each time. I am very grateful Domenica for your patient explanations and rigorous corrections. I felt great support and dedication from you. Also, thank you both for surviving my deadline difficulties.

I would like to strongly thank Joëlle Ronfort and Christina Vieira for having accepted to be reporters of this thesis. I would also like to thank Valerie Le Corre and Jacqui Shykoff for being my examiners, and Luis Eguiarte for being invited examiner too.

I received great support and guidance from my PhD committee members, Anna-Sophie Fiston-Lavier, Nicolas Bierne, Aleksandra Walczak and Yves Vigouroux who took precious time and attention in evaluating my work. I thank you very much. In particular, the time spent closely working with Anna-Sophie was very motivating and allowed me to take a deep dive into transposable element bioinformatic analyzes. The welcome I received at the Bethune's in Montpellier was also a heartwarming experience.

I have greatly appreciated the good mood and quick help from Jacqui Shykoff and Marianne Delarue, as well as the excellent work of the Ecole Doctorale "Sciences du végétal : du gène à l'écosystème". I would also like to congratulate the PhD students association Doc' en Herbe for their fantastic work in organizing events and great conferences. I'd like to thank the IDEEV for helping me with conference fees and allowing me to present my work in the IDEEV day. I would like to thank the Université Paris-Saclay for this fantastic experience. I am also very thankful to the national research council of Mexico, CONACYT, for my four year scholarship No. 57916/310738 that gave me the opportunity to develop this research project abroad.

I had a very gratifying experience in participating in La Fête de la Science. Working together with my colleagues and interacting with visitors of all ages was fun and enriching.

The excellent work performed by Valérie Lespinas, Sandrine Le Bihan and Rozenn Le Guyader has made all paperwork smooth and easy. Thank you for your kindness and willingness to help. Also the administrative team, Khalid Ouskou, Frédéric Huynh, Djibril Pouye and Marwa Sultan, had always a big smile that made everyday nicer.

I am very grateful to all the members of the team Dygap, each one special in their own way. Your enthusiasm to work together made this an enriching experience. I especially enjoyed the scientific discussions set by Pierre Gerard, Amandine Cornille, Clementine Vitte and Karine Alix. I

would like to give an very large thank you to Hélène Corti for her hard work at the laboratory and her cheerful manner in results discussions. I also thank masters student Juliette Aubert for helping in laboratory experiments.

I have had the enormous good luck to work with Margaux-Alison Fustier and build a beautiful friendship. Your open arms welcome to France and your efforts in teaching me French have been beyond my expectations. Your amazing high spirits capacity have been a great motivation for me. I also am very grateful to Jean-Tristan Brandenburg for your help, your patience and particular humor. Thank you for sharing your views and helping me when I needed it. And dancing.

I am also glad to have met Héloïse Giraud, who was very confronting and welcoming to this new place. I would like to give a huge thank you to Zeineb Achour with whom I shared so many long nights at work and who showed me always an open heart. I am very proud of you and I appreciate you for teaching me through your determination at work and your care as friend.

As for the second office generation I am very happy to have met Arnaud Desbiez-Piat who has always the peach and the banana. You bring a very good energy and I like discussing all sorts of topics with you. I also enjoy the tranquility of Maèva Mollion and her happy nature.

Meeting my fantastic friend Marianyela Petrizzelli has been undoubtedly a magical experience. I am very happy to have shared this part of my life with you. Your support has made it possible for me to stand here today. I also am very thankful to Adrien Falce for explaining to me computer stuff and life stuff. Sharing coffee, beers and ideas. I would like to thank Amandine Cornille for her never-ending energy, her friendliness and her outing proposals. Recently arrived Maud Fagny has been a great encounter.

I would like to especially thank Anthony Venon for his outstanding good humor, and for welcoming me to Le Moulon since day 1. Your excellent work and will to learn more have been a great motivation for me. Daily encounters with forever smiling Carine Remoué and Valérie Combes have made my days here bright. I thank Christine Dillman for urging me to take her statistics course. I enjoy your endless curiosity and scientific impulse as well as your patience.

I would like to thank Luis Eguiarte for guiding me in my bachelors years, encouraging me to take the leap and introducing me to Maud. I would also like to thank Jonás Aguirre for his help and enthusiastic attitude in our struggles with our paper’s revisor #4.

As previously noted, I have a great deal to say about the family-like working environment at Le Moulon. Also, among the lovely memories I have collected, I specially enjoyed the SemiIDEEV and the barbecues, with a special thanks to Harry for promoting them. The gym sessions with the “wheat girls” were the occasion to burn out some stress and have a laugh mid week-through. The basketball team was an superb experience for mind and soul. I especially enjoyed the positive vibes of Pierre, Christophe, Bubar, Yannick, Yu-Ming, Marianyela and Adama, on the courtyard.

I would like to thank Sharon Ainsworth and Bernard Ainsworth for their support and protection during my first ‘real’ winter. I thank Amalfi Martínez for showing me the value of joy. I thank Margarita Bernal for being welcoming and attentive to me. I sincerely thank my French family that took me in and was warm and very interesting. I feel as if I had known you since all times. I specially enjoyed sharing meals and talking with motherly Yannick and François, generous James and Danielle, kind and lively Marc and Sylvia, energetic André, caring Nathalie and Pierre, thoughtful Anne and Tom and all my supercool cousins! I would like to thank Romain Filzot, Raul Velasco and Kristin Meller for opening their workshops and homes and making everything flow. I am super grateful to have met Irene Iodice, the best roomie in the world. You have filled my days with laughter and friendship that I treasure. I thank my friends from cité, Paula Manassero, Isui Aguilar and Keiko Kitagawa, I have enjoyed every minute of your company. I thank the crazy Italian gang, Federica Maschietto, Flavia Corsi, Alice Romiti, Chiara Mazzocconi and quasi-italian Paula Adamczyk who make me feel at home and with whom I love to share picnics. I would like to especially thank Luz María Lazcano for believing in me and giving me the strength of doing it my way.

I have no words to explain how immensely grateful I am to my family. Thanks to my dad Gustavo Martínez, who was always there to cheer me up and came all the way here to share cheese with me. Thank you for your good humor and your will to enjoy life. I thank my mum, Shirley Ainsworth for being brave and happy. I look up to you and I am overjoyed to feel your constant support. A tremendous thank you to my sisters, Laura Martínez and Camila Martínez. I never feel alone because I know you´re with me. Thank you Laura for your meme therapy and Camila for your excellent humor and numerous visits. Love you ardillitas. None of this would have been possible without you.

I thank Olmo Uribe and our ocean.

I am forever grateful to the people of Mexico who payed for my scholarship with their hard work. I am indebted to you and I hope to apply the knowledge I have developed to help take care of our megadiversity for the benefit of all.

Table of content

Synthèse en français

Les deux sous-espèces annuelles de téosinte qui sont les plus proches parents sauvages du maïs sont d'excellents systèmes pour étudier l'adaptation locale car leurs distributions couvrent un large éventail de conditions environnementales (Hufford, Bilinski et al. 2012). *Zea mays* ssp. *parviglumis* (ci-après *parviglumis*) est distribuée dans un habitat chaud et mésique en dessous de 1800 m d'altitude, tandis que *Zea mays* ssp. *mexicana* (ci-après *mexicana*) prospère dans des conditions sèches et fraîches à des altitudes plus élevées (Hufford, Martínez-Meyer et al. 2012). Des études sur le processus de spéciation écologique entre *parviglumis* et *mexicana* ont mis en évidence l'existence de flux de gènes récurrents entre les deux sous-espèces (Aguirre-Liguori, Gaut et al. 2019). Malgré ces flux, les téosintes présentent une structuration génétique à l'échelle spatiale (Fukunaga, Hill et al. 2005 ; van Heerwaarden, Ross-Ibarra et al. 2010). Par ailleurs, des introgressions adaptatives ont été rapportées depuis *mexicana* vers le maïs, lui-même domestiqué à partir de populations de la sous-espèce *parviglumis* (Wang, Stec et al. 1999 ; Piperno and Flannery 2001 ; Matsuoka, Vigouroux et al. 2002 ; van Heerwaarden, Doebley et al. 2011 ; Hufford, Lubinksy et al. 2013). Nous nous sommes intéressés ici à caractériser les déterminants phénotypiques et génétiques de l'adaptation locale des téosintes *parviglumis* et *mexicana* le long de gradients altitudinaux.

Nous avons travaillé sur un panel d'association constitué de 1664 plantes provenant de graines de 11 populations de *parviglumis* (8) et *mexicana* (3). Ces populations ont été échantillonnées le long de deux gradients d'altitude relativement éloignés l'un de l'autre. Ce panel a été évalué pour 18 caractères phénotypiques durant deux années consécutives dans deux jardins communs situés au Mexique à une altitude intermédiaire. Par ailleurs, des données de séquençage haut débit de six populations comprenant des populations de basse et haute altitude, étaient disponibles. Ces données ont permis d'identifier un sous-ensemble de 171 polymorphismes nucléotidiques (SNP candidats) présentant des signaux de sélection compatibles avec leur implication dans des processus d'adaptation à l'altitude. Les SNP candidats ainsi que 38 marqueurs microsatellites ont été génotypés sur le panel d'association. En parallèle, nous avions également à notre disposition un panel de 28 populations (panel étendu contenant 10 des 11 populations du panel d'association), échantillonnées le long des mêmes gradients, mais caractérisés uniquement d'un point de vue génétique par les SNP candidats et 1000 SNP neutres.

Dans le premier chapitre, nous avons utilisé les données phénotypiques du panel d'association pour réaliser une analyse en composantes principales. Nous avons ainsi pu démontrer l'existence d'un syndrome phénotypique multivarié qui est corrélé avec l'altitude de la population d'origine. Pour chaque caractère pris indépendamment, nous avons ensuite mis en évidence des effets significatifs de l'altitude de la population d'origine sur leur variance phénotypique. Enfin, nous nous sommes basés sur la comparaison entre le niveau de divergence mesuré par des marqueurs neutres (SNP neutres et microsatellites) et le niveau de divergence phénotypique pour identifier un sous-ensemble de dix caractères évoluant sous sélection spatialisée. Ces dix caractères constituent un syndrome d'adaptation à l'altitude caractérisé par une augmentation de la précocité de floraison, une diminution de la production de talles et de la densité en stomates des feuilles ainsi qu'une augmentation de la taille des plantes, et de la longueur et du poids des grains. De façon intéressante, ce syndrome a évolué malgré la présence de flux de gènes. Nous avons en effet détecté, par l'analyse des polymorphismes neutres, des flux de gènes à longue distance entre sous-espèces et aussi entre populations d'une même sous-espèce.

Nous avons poursuivi notre étude en testant l'association entre les SNP candidats et la variation génotypique pour chacun des 18 caractères. Pour contrôler la structure génétique neutre de nos échantillons, nous avons utilisé le génotypage des marqueurs microsatellites afin de réaliser une assignation Bayésienne en groupes génétiques et de reconstruire une matrice d'apparentement. En recherchant les déterminismes génétiques sous-tendant ce syndrome, nous avons montré que le pourcentage de SNP candidats associés aux différents caractères dépendait de la prise en compte de la structure neutre soit en cinq groupes génétiques (K=5, 73.7%), soit en onze populations (POP=11, 13.5%), indiquant une stratification complexe du panel d'association. Nous avons réalisé plusieurs observations intéressantes concernant l'association des SNP candidats : 1) mis à part un SNP, tous les SNP candidats associées avec la correction à onze populations (POP=11) sont contenus dans l'ensemble de ceux détectés avec cinq groupes (K=5) ; 2) les SNP sont le plus souvent associés à plus d'un caractère ; 3) réciproquement les caractères présentent plusieurs SNP associés, et nous avons été capables de détecter dans certains cas des effets indépendants de ces SNP ; 4) globalement le déséquilibre de liaison (DL) est assez faible, même si les SNP associés présentent en général plus de DL que les autres SNP ; 5) les SNP associés sont retrouvés aussi bien dans les régions géniques qu'inter-géniques.

Afin d'étudier la correspondance entre les SNP associés à la variation phénotypique des caractères, et ceux corrélés avec la variation environnementale, nous avons testé cette dernière sur le panel étendu de 28 populations. Pour cela, nous avons « résumé » l'information contenue dans 19 variables abiotiques déterminées pour chacune des 28 populations par deux composantes

principales. Après la prise en compte d'une matrice de covariance des fréquences alléliques calculée sur les SNP neutres, nous avons établi une liste de SNP candidats associés à l'environnement. Une large proportion (50.88 %) de SNP sont associés à la première composante principale, elle-même fortement corrélée à l'altitude des populations. L'un des résultats majeurs de cette étude est la détection d'un enrichissement de SNP candidats associés aux phénotypes et à l'environnement dans trois larges inversions chromosomiques, indiquant leur rôle clé dans l'adaptation locale des populations à l'altitude.

Dans le deuxième chapitre de la thèse, nous nous sommes focalisés sur une autre source de variation génétique que les SNP, celle des éléments transposables. Ces éléments peuvent en effet jouer un rôle fonctionnel important dans les processus adaptatifs. Il s'agit d'éléments qui ont, ou ont eu, la capacité de se déplacer (transposer) dans le génome, soit *via* l'intermédiaire d'un ARN dont une copie s'insère à un autre endroit du génome (mécanisme copier-coller), soit *via* l'intermédiaire d'un ADN excisé qui s'insère à un autre endroit du génome (mécanisme couper-coller). Ces éléments sont classés en ordres, superfamilles et familles selon leur mécanisme de transposition, leurs caractéristiques et leur homologie entre eux. Des effets phénotypiques importants liés aux insertions des ET ont été répertoriés chez les plantes cultivées (Vitte et al. 2014). Chez le maïs, le contenu du génome de référence (lignée B73) et l'identité des ET ont été bien décrits. L'annotation des ET a révélé que ces éléments constituent environ 85% du génome (Schnable, Ware et al. 2009 ; Stitzer, Anderson et al. 2019). Cependant ce contenu varie beaucoup d'une lignée à l'autre, provoquant de très nombreux polymorphismes d'insertions-délétions entre lignées (Springer, Anderson et al. 2018). Nous avons ici exploré la contribution de la variation des ET à l'adaptation locale chez les téosintes. Nous nous sommes tout d'abord concentrés sur l'estimation du contenu en éléments transposables, offrant ainsi une première description chez la plante ancêtre du maïs cultivé. Ensuite, nous avons développé une méthodologie visant à estimer les fréquences alléliques d'insertions d'ET en utilisant les données de séquençage haut-débit de quatre populations.

Nous avons effectué la première description populationnelle des ET chez les téosintes pour deux catégories d'insertions : celles présentes à une position donnée dans le génome de référence de la lignée B73 mais polymorphes dans les quatre populations de téosintes (insertions de référence), et celles absentes à une position donnée dans le génome de référence mais présentes et polymorphes (insertions *de novo*) dans les quatre populations de téosinte. Nous avons montré que pour les deux types d'insertions, de référence et *de novo*, les quatre populations présentent des proportions similaires en termes de comptage d'éléments trouvés, au niveau des familles et superfamilles. Les

paysages d’insertions le long des chromosomes reflètent ceux connus chez la lignée B73 et varient d’une superfamille à l’autre.

Nous avons estimé les fréquences des insertions d’ET et identifié un échantillon de celles présentant des fréquences alléliques contrastées entre populations de basse et de haute altitude de façon parallèle dans les deux gradients. Nous avons ensuite étudié leurs contextes génomiques, notamment la distance aux gènes les plus proches et la fonction de ces gènes. L’objectif à court terme est d’utiliser le panel d’association pour tester le lien entre le polymorphisme génétique de ces insertions et la variation phénotypique des caractères mesurés au chapitre 1 pour certaines de ces insertions – celles dont la fonction des gènes pourrait être compatible avec leur implication dans le déterminisme des caractères étudiés. A l’inverse, nous avons aussi génotypé, dans le panel d'association, des insertions d’ET connues pour avoir contribué à l'évolution phénotypique du maïs et impliquées dans des caractères de floraison (insertion au locus *Vgt1*) ou d’architecture de la plante (insertion au locus *Tb1*). Dans le cas de l’insertion *Vgt1*, nous avons validé son rôle dans le contrôle de la floraison chez les téosintes, l’insertion étant associée à une plus grande précocité. Par contre, l’insertion *Tb1* n’est associée à aucun effet phénotypique chez les téosintes, ce qui suggère que son effet dépend intimement du fond génétique dans laquelle elle se trouve.

Notre étude apporte ainsi de nouvelles connaissances sur l’adaptation altitudinale chez les téosintes, et plus généralement chez les plantes tropicales. Elle ouvre la discussion sur les défis soulevés par l'utilisation (1) d'outils de génomique des populations pour identifier la variation adaptative, (2) de populations naturelles en génétique d’association, et (3) de ressources génétiques sauvages pour l'amélioration des espèces cultivées.

Références

Aguirre-Liguori, J. A., B. S. Gaut, et al. (2019). " Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (*Zea mays parviglumis* and *Zea mays mexicana*) " Molecular Ecology: 2814-2830.

Hufford, M. B., P. Bilinski, et al. (2012). "Teosinte as a model system for population and ecological genomics." Trends in Genetics **28**: 606-615.

Hufford, M. B., E. Martínez-Meyer, et al. (2012). "Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight." PLoS ONE **7**.

Hufford, M. B., P. Lubinksy, et al. (2013). "The Genomic Signature of Crop-Wild Introgression in Maize." PLoS Genetics **9**.

Fukunaga, K., J. Hill, et al. (2005). "Genetic diversity and population structure of teosinte." Genetics **169**: 2241-2254.

Matsuoka, Y., Y. Vigouroux, et al. (2002). "A single domestication for maize shown by multilocus microsatellite genotyping." Proceedings of the National Academy of Sciences of the United States of America **99**: 6080-6084.

Piperno, D. R. and K. V. Flannery (2001). "The earliest archaeological maize (zea mays l.) from highland mexico: new accelerator mass spectrometry dates and their implications. Proceedings of the National Academy of Sciences of the United States of America **98**: 2101-2013.

Schnable, P. S., D. Ware, et al. (2009). "The B73 Maize Genome: Complexity, Diversity, and Dynamics."

Springer, N. M., S. N. Anderson, et al. (2018). "The maize W22 genome provides a foundation for functional genomics and transposon biology." Nature Genetics **50**.

Stitzer, M. C., S. N. Anderson, et al. (2019). "The Genomic Ecosystem of Transposable Elements in Maize." bioRxiv: 559922.

van Heerwaarden, J., J. Ross-Ibarra, et al. (2010). "Fine scale genetic structure in the wild ancestor of maize (*Zea mays* ssp. *parviglumis*)." Molecular Ecology **19**: 1162-1173.

van Heerwaarden, J., J. Doebley, et al. (2011). "Genetic signals of origin, spread, and introgression in a large sample of maize landraces." Proceedings of the National Academy of Sciences of the United States of America **108**: 1088-1092.

Vitte, C., M. A. Fustier, et al. (2014). "The bright side of transposons in crop evolution." Briefings in Functional Genomics and Proteomics **13**: 276-295.

Wang, L., A. Stec, et al. (1999). "The limits of selection during maize domestication." Nature **398**: 236-239.

I. INTRODUCTION

One of the central questions in evolutionary biology concerns the processes that create and maintain genetic variation. Among them, local adaptation plays a central role in the maintenance of variation both at the phenotypic and genomic level (Mitchell-Olds, Willis et al. 2007). Evidence for rapid adaptation suggests that such variation may be determinant for population's capacity to respond and adapt to current environmental shifts (Bay, Rose et al. 2017). In this introduction, I provide a definition of local adaptation, how to detect it, and review what has been discovered about its underlying molecular mechanisms, focusing more particularly on higher plants. I subsequently review the literature on the role of transposable elements in local adaptation. Finally, I present my model system, the two closest wild relatives of maize, the teosinte subspecies *Zea mays* ssp. *parviglumis* and *Zea mays* ssp. *mexicana*.

I.1 LOCAL ADAPTATION

I.1.1 Definition and pervasiveness

Living species inhabit the globe forming populations of inter-fertile individuals that share a given space and time. Biological diversity is a product of evolution. Population genetics offers an interesting framework to study evolution. It focuses on describing the genetic composition of populations through space and/or time, and on investigating the evolutionary forces that drive those changes (Dobzhansky 1964). One of the major forces that we have focused on in our work is natural selection, which operates on phenotypic diversity. Phenotypic diversity emerges from genetic variation, environmental factors, and their interactions. Inheritance of genetic variation makes phenotypes heritable. Natural selection, acting on those heritable variants, leads to changes in the genetic composition of populations and their phenotypic adaptation.

Populations' environmental contexts can be highly heterogeneous with biotic and abiotic factors exerting differential selection across species ranges. This diversity of selection pressures may drive each population to different local phenotypic optima for adaptive traits. Hence, evolution through divergent natural selection provokes shifts in allele frequencies in response to local selective pressures that maximize individual's fitness – their survival and reproductive success. Because natural selection modulates allelic frequencies of each population deriving from an ancestral population, ideally one could compare ancestral and evolved populations to seek evidence

for local adaptation. Unfortunately, access to ancestral populations is often impossible, so in practice, it is easier to perform comparisons across present day populations that have evolved under different environmental conditions as a way to test for local adaptation (Kawecki and Ebert 2004).

When observing different populations at a given time point, local adaptation may be evidenced when a native population has higher fitness in its native environment than any other non-native population, and conversely its fitness is diminished in a non-native environment (Figure 1). Empirical approaches for the study of local adaptation thus include the measuring of fitness differences in reciprocal transplantations (Savolainen, Lascoux et al. 2013). These approaches are very insightful but also labor intensive. The amount of populations and the species biology can render them inappropriate in some cases. Reciprocal transplant studies have been mainly carried out in plants (Savolainen, Lascoux et al. 2013). Because of their sedentary nature, plants are more likely prone to local adaptation. Indeed, in herbaceous temperate plants, a meta-analysis on 1032 population pairs found that in ~70% of studies the native population outperformed the other populations in its native environment. Yet this figure descended to 45% when considering population pairs for which strict local adaptation – in two directions – was recorded (Leimu and Fischer 2008).

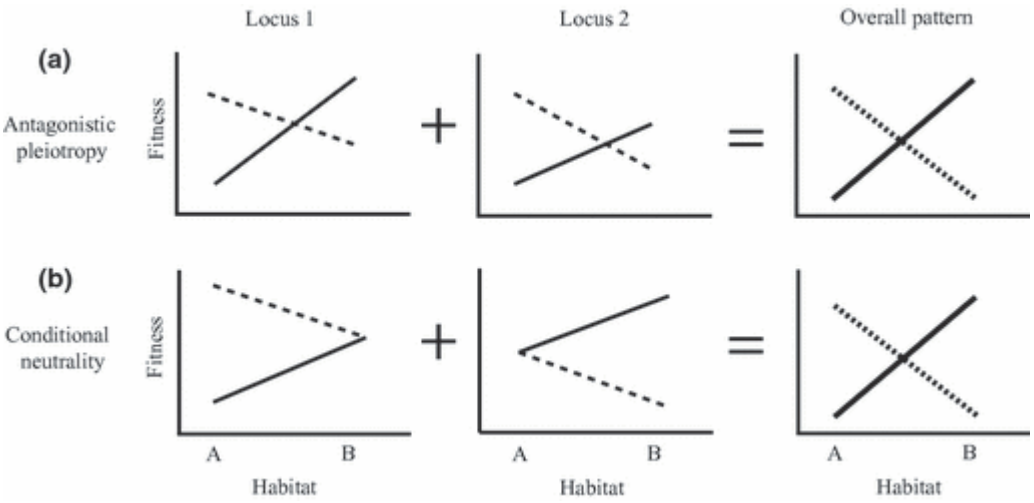


Fig.1: Two competing genetic models for local adaptation, resulting in reciprocal home site advantage. Fitness is compared between individuals bearing the allele from population A (dashed line) and population B (solid line), in both A and B habitats. In antagonistic pleiotropy (a), local alleles confer higher fitness in both habitats. In conditional neutrality (b), local alleles confer fitness advantage in only one habitat (habitat A for locus 1; habitat B for locus 2), while neutral in the other habitat. Adapted from Lowry (2012)

I.1.2 Competing models for local adaptation

There are currently two models that have been proposed to describe the genetic bases of local adaptation: conditional neutrality and antagonistic pleiotropy. On one hand, antagonistic

pleiotropy – or genetic trade-offs – occurs when at a given locus, one allele confers a fitness advantage over the other in one environment, while the opposite applies to another environment (Schnee and Thompson and Jr 1984) (Figure 1a). On the other hand, in conditional neutrality two alleles differ in their fitness effects in only one environment, so that the advantageous allele may become fixed at the species scale (Kawecki 1997) (Figure b). These two models differ in their outcomes regarding the maintenance of diversity at the species scale. In antagonistic pleiotropy, disruptive selection across populations maintains polymorphism at the species scale, whereas conditional neutrality does not necessarily predict maintenance of diversity.

A method proposed by Anderson *et al.* (2011) aids in distinguishing which of these models applies. Reciprocal transplants are used to measure fitness-related traits such as flowering, and changes in allele frequency are monitored across the genome from one generation to the next in contrasted environments. Permutation of genotypes and phenotypes are used to compute a null distribution of allele frequency changes in each environment. Outlier loci for which changes in allele frequency exceed the neutral expectation in one environment are further tested in the other environment. Application of this method in recombinant inbred lines of *Boechera stricta* grown in Montana and Colorado, has provided evidence for conditional neutrality at 8% of the loci and for antagonistic pleiotropy at 2.8% of the loci. Interestingly, the latter model concerned one major flowering quantitative trait loci (QTL).

Antagonistic pleiotropy results from trade-offs between multiple fitness components such as resource allocation to growth (survival) and reproduction (fecundity). Correlations between fitness and phenology have been reported in *Arabidopsis*. Reciprocal transplants of *A. thaliana* in two contrasting locations revealed that Single Nucleotide Polymorphisms (SNPs) whose frequencies were correlated with environmental variables, were found more often in genetic trade-off QTLs (detrimental in the opposite population) than in conditional neutrality ones (Price, Moyers et al. 2018). For instance, the gene *FRIGIDA* exhibits two categories of alleles, early flowering alleles conferring drought escape, and late flowering alleles conferring increased water use efficiency in line with a drought avoidance strategy (Lovell, Juenger et al. 2013). Likewise, in monkey flowers (*Mimulus guttatus*), several studies have reported genetic correlations between flowering phenology, viability and fecundity. Through an intra-population field experiment, authors found a genetic trade-off that probably responded to the yearly and short spatial fluctuating magnitude and direction of selection on *M. guttatus* corolla width, rendering QTL alleles unfit to increase flower size and fertility at the same time as viability (Mojica, Lee et al. 2012).

I.1.3 Conditions of emergence and maintenance of local adaptation

Local adaptation results from the interplay of local selective environmental pressures, genetic drift and gene flow. Early theoretical work has shown that for local adaptation to occur, selection should be sufficiently strong to overcome migration of maladapted alleles and prevent the loss of locally advantageous alleles (Haldane 1930; Bulmer 1972). Hence, under environmentally antagonistic selection, if gene flow counteracts natural selection, it translates into a loss of polymorphism and a migration load. Interestingly, this migration load may be reduced if fewer loci are controlling divergent phenotypes. Simulations are indeed suggesting that migration favors a genetic architecture with few alleles of large effect encoding adaptive phenotypes (Yeaman and Whitlock 2011). In addition, mutation load may also indirectly trigger selection for mechanisms reducing gene flow between habitats such as reduced dispersal, increased plasticity, and reduced recombination, i.e. linkage and/or chromosomal rearrangements (Lenormand 2002).

Local adaptation with gene flow may proceed under three main scenarios (Tigano and Friesen 2016): (1) environmentally-driven divergence of populations despite gene flow; (2) gene flow occurrence after secondary contact of diverged locally adapted populations; (3) chromosomal rearrangements that maintain adaptive morphs between (resp. within) a population despite gene flow (resp. free interbreeding). Examples of these scenarios in plants include: evidence of isolation by environment with gene flow between teosinte subspecies (Aguirre-Liguori, Tenaillon et al. 2017; Aguirre-Liguori, Gaut et al. 2019) and a chromosomal inversion in the yellow monkey flower (*Mimulus guttatus*) that allows locally adapted loci to maintain divergent annual and perennial ecotypes in the face of gene flow (Twyford and Friedman 2015). (Figure 2).

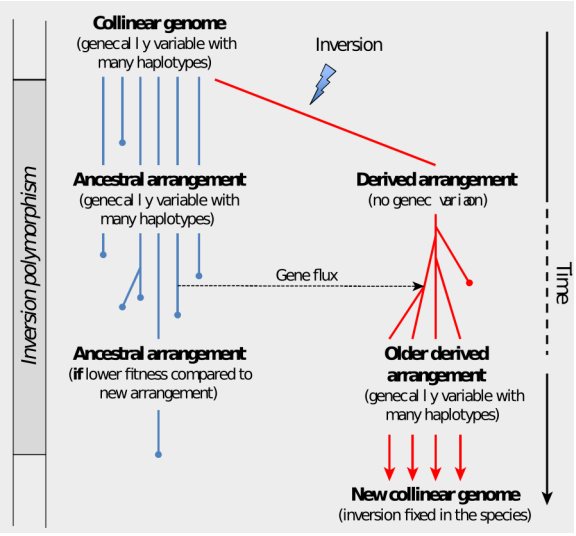


Fig.2: Simplified Life History of an Inversion. A mutation generating a new inversion results in one derived and one ancestral arrangement; the former initially without variation. Over time, point mutations and gene flow add new variation, and selection and drift reduce variation in both arrangements. Eventually, one of the arrangements (in this illustration the ancestral one) might be lost and the remaining arrangement (here the derived one) becomes the new collinear genome in this genomic position. Image and legend taken from Faria, *et al*. (2019)

I.2 GENOMIC SIGNATURES OF LOCAL ADAPTATION AND ENVIRONMENTAL DRIVERS

I.2.1 Genome-wide scans for allele differentiation

In what is known as a bottom-up or reverse ecology approach, genome scans can be utilized to find genomic regions that have been under selection without *a priori* information (that could easily be biased according to the current state of knowledge); furthermore, such pinpointed genomic regions can sometimes suggest which is the phenotypic trait that is being targeted by natural selection (Ross-ibarra, Morrell et al. 2007; Li, Costello et al. 2008). This means that relevant ecological traits and their genetic determinants can be deduced from the genomic data.

At the genomic level, the most obvious signature of local adaptation is increased allele differentiation between populations as originally proposed by (Lewontin and Krakauer 1973). Besides increased allele differentiation (classically measured by *F*st), loci targeted by local adaptation may display a loss of genetic diversity and increased Linkage Disequilibrium (LD) within-populations. The LD signature tends to dissipate quickly once the selected mutation has reached fixation, its power being therefore limited to a narrow window of time (McVean 2007). Allele differentiation can be detected by integrating a spatial component to the decomposition of allelic variance (Beaumont and Balding 2004). A popular software that implements such method is Bayescan (Foll and Gaggiotti 2008). It considers an island model, where multiple subpopulations are derived from an ancestral population. Subpopulations may have been subjected to different amounts of genetic drift and, therefore, their allele frequencies will display various degrees of differentiation from the ancestral allele frequency. This demographic component, specific to each population, is accounted for in the detection of loci that display signals of selection.

While methods based on allele differentiation are appealing, they are not without caveats. It is often difficult to distinguish locus-specific signals from genome-wide patterns generated by population demography. The use of a simple island model to describe population structure in such situation may cause a high rate of false positives (Excoffier, Hofer et al. 2009). High F_{ST} values may indeed be caused by allele surfing during range of expansions, such that differentiation at some random loci may be high between populations in the periphery of a species range (Hallatschek, Hersen et al. 2007). In addition to demography, the effects of background selection may also be misleading (Pool, Hellmann et al. 2010) (Bank, Ewing et al. 2014). Attention has also been called to avoid jointly analyzing markers with different modes of inheritance (located on sexual chromosomes versus autosomes, chloroplast or mitochondrial versus nuclear markers) since their

effective size differences could translate to overestimation of extreme F_{ST} values (Pool and Nielsen 2007).

Predictions of higher allele differentiation hold under a 'hard' sweep scenario, where adaptation proceeds through the rapid fixation of a beneficial mutations occurring after the onset of selection (Maynard Smith and Haigh 2008) (Figure 3). But, adaptation is thought to often proceed either through fixation of a mutation segregating in the population before the onset of selection (the so-called standing genetic variation) or even through recurrent beneficial mutations (Hermisson and Pennings 2017). In these 'soft' sweep scenarios where multiple alleles may be sweeping, footprints of selection are much more difficult to detect at the genome-wide scale because soft sweeps have weaker effects on linked sites. The frequency of "hard" versus "soft" sweeps ultimately depends on the effective population size and the mutation rate (Messer 2013).

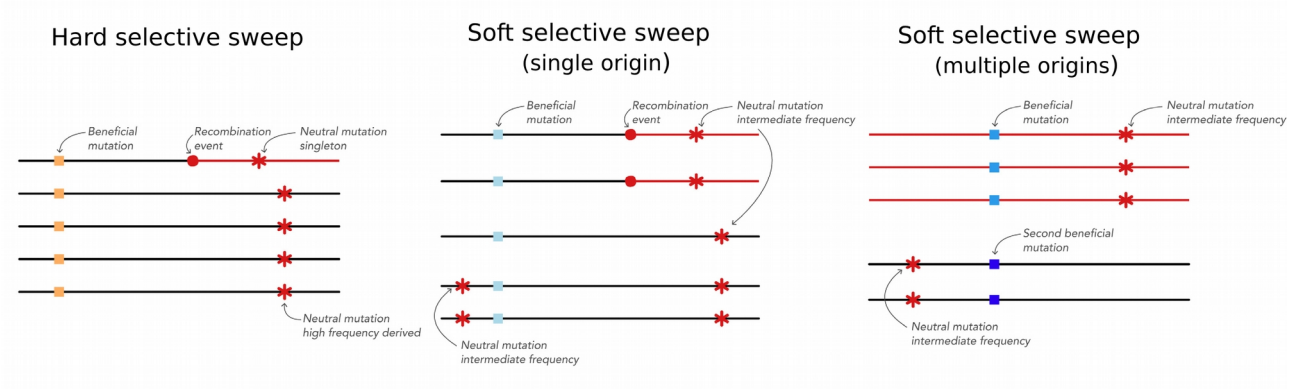


Fig. 3. Hard and soft selective sweeps. Mutations and recombination events are shown on haplotypes of the five sampled individuals. Squares indicate the beneficial mutation, circles recombination events and asterisks neutral mutations. Left panel: In a hard sweep, all ancestral variation at tightly linked sites is eliminated, and recombination leads to low-frequency and high-frequency derived variants in flanking regions. Middle panel: For a single-origin soft sweep from standing genetic variation, early recombination introduces ancestral haplotypes at intermediate frequencies. Right panel: The beneficial allele traces back to multiple origins. Each origin introduces an ancestral haplotype, typically at intermediate frequency. Figure adapted from Hermisson and Pennings (2017).

Genomic investigations in human have led to emblematic discoveries for physiological adaptations with varying degrees of complexity in their genetic architecture, from a few genes as in lactase persistence (Tishkoff, Reed et al. 2007) to a great many as in height (Turchin, Chiang et al. 2012) (Figure 4). The polygenic model of adaptation complexifies further those predictions. Polygenic adaptation from standing variation occurs when traits are encoded by a very large number of genes with small effects. Adaptation of highly polygenic traits indeed involves a myriad of subtle correlated changes of allele frequencies at the interacting loci, leaving no clear footprints at the genomic level. For example, for human height no clear signatures of strong recent selection have has been found in the genome (Figure 4), which is why this trait has benefited of the research

on polygenic scores from genome-wide association studies (GWAS), not without a series of backlashes due to misleading underlying structure (Sohail, Maier et al. 2019).

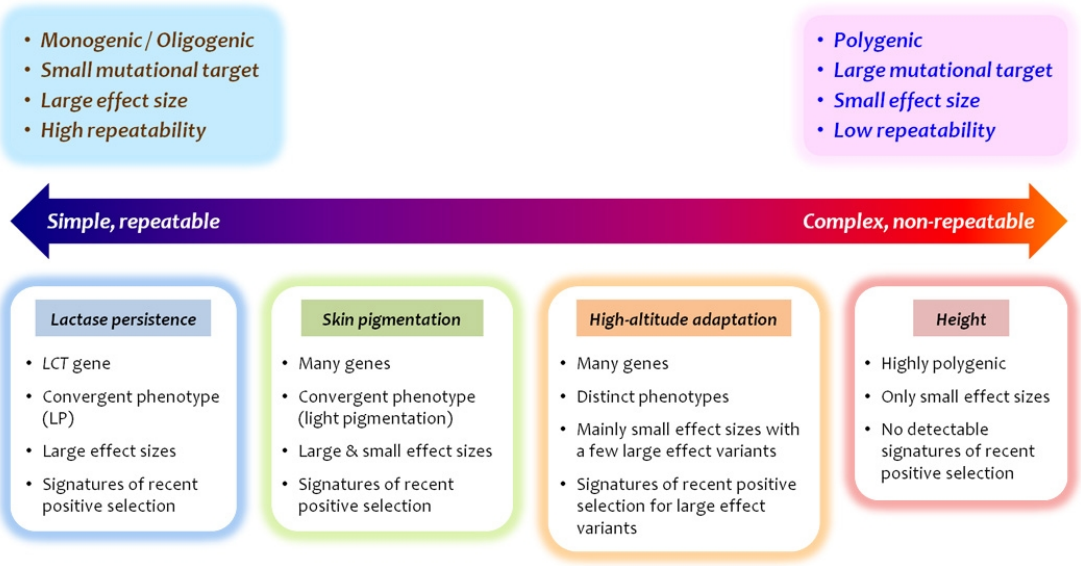


Fig. 4: A schematic view of the genetic architecture of adaptive traits across its complexity spectrum. Figure taken from Jeong and Di Rienzo (2014)

I.2.2 Correlations with environmental variables

Genome scan methods that rely on differentiation among populations to detect outliers to neutral expectations are designed to detect positive selection, yet they only assume that selection pressures vary between populations without singling out which selective pressures are at play. Approaches that incorporate environmental data as a driving force can therefore complement differentiation-based tests (Rellstab, Gugerli et al. 2015). Bayesian frameworks such as that employed in Bayenv2.0 software (Coop, Witonsky et al. 2010; Günther and Coop 2016) directly evaluate the impact of environmental factors on polymorphic genetic marker distribution while accounting for co-variation of allele frequencies that may be caused by underlying demographic processes.

An alternative approach that operates under a similar logic is that of partial Mantel tests, where the comparison between two pairwise distance matrices is controlled for the effect of a third matrix, as for example the neutral population structure estimated by genome-wide pairwise F_{ST} values. In any case, it is important to consider that environmental correlation methods that assume independence between populations may produce false positives when this assumption is flawed (Hoban, Kelley et al. 2016). An example of application of this latter method is illustrated in *Arabidopsis halleri*, an outcrosser known to grow on diverse soil types along the Alps, for which

Fischer et al., (Fischer, Rellstab et al. 2013) took population pooled high-throughput sequencing data for geographically close localities that experience steep environmental and biotic differences and employed partial Mantel tests to associate non redundant environmental variables on a set of highly differentiated SNPs. The authors posit that the footprints of selection they recovered can be explained by a reduced set of topo-climatic factors, namely site water balance, precipitation, radiation, temperature and slope.

Generally, genotype-environment correlation methods are more powerful than differentiation based methods, with the downside of a higher false positives rate (De Mita, Thuillet et al. 2013). But because environmental factors are often correlated, the causative factor is not always easy to establish (Bradburd, Ralph et al. 2013) for example when biotic factors are not directly measured but rather are reflected by an abiotic factor that varies throughout the sampling design (Hoban, Kelley et al. 2016). Besides spatial correlation of environmental factors, neutral structuring of genetic data is an important confounding factor as it can produce patterns similar to those expected for local adaptation. When neutral structure produced by population history fully overlaps that of natural selection, it is difficult to distinguish them. Control for neutral structure may also completely erase signals of natural selection. It is also important that the environmental variable has had 'enough' time to leave signatures at the genetic level (Anderson, Epperson et al. 2010). Indeed, there can be considerable time lags between the onset of selection in response to environmental pressures and its observable impact on genetic variation. And stressing the point, the spatial scale considered must be biologically meaningful for the organisms' fitness in order to find relevant signals of local adaptation (Hoban, Kelley et al. 2016).

I.3 GENETIC BASES OF LOCAL ADAPTATION

I.3.1 Spatially-varying traits

Spatially-varying selection triggers local adaptation whose traces include increased differentiation of quantitative traits among populations. Common garden data allow to compare inter-population quantitative genetic divergence for a trait, measured by Q_{ST} (Spitze 1993), with the neutral genetic differentiation measured by F_{ST} (Wright 1951; Edelaar, Burraco et al. 2011). The null hypothesis being that neutrality cannot be ruled out as the cause of the observed phenotypic patterns. Under the assumption that all genetic variation is additive and the mutation rate that contributes to the trait is equal to that found in neutral loci, then Q_{ST} is expected to be equal to the mean F_{ST} value when the trait is selectively neutral (Holsinger and Weir 2009). When Q_{ST} is

significantly smaller than F_{ST}, it can be assumed that the trait has been modeled by stabilizing selection since it would be acting on the quantitative trait in the same way in each deme (Holsinger and Weir 2009). Alternatively, if Q_{ST} is significantly larger than *F*st, this would be indicative of spatially-varying (Gilbert and Whitlock 2015) or diversifying selection (Holsinger and Weir 2009). Care must be taken in choosing the correct neutral markers, since cases in which the mutation rate of such markers is much higher than gene flow, F_{ST} estimations would be underestimated thus biasing its comparison to Q_{ST} (Edelaar, Burraco et al. 2011). Also, if selection varies spatially but fluctuates in time at a fast rate, its signature may not be recovered (Pujol, Blanchet et al. 2018). Caution is recommended when assuming that trait variability is due exclusively to additive genetic variance, since non-additive variance can cause Q_{ST} to differ from F_{ST} even for neutral traits (Leinonen, McCairns et al. 2013) Other confounding effects include dominance effects and maternal environments, which should also be formally addressed (Leinonen, McCairns et al. 2013).

A rich literature of examples has found compelling evidence of traits evolving under spatially varying selection. In plants, interesting examples have been reported for sunflowers, where flowering time and growth rate have been found to display signatures of spatially varying selection in *Helianthus maximiliani* (Kawakami, Morgan et al. 2011). *Helianthus agrophyllus* also displays two main life history syndromes driven by spatially varying selection (Moyers and Rieseberg 2016). Even when environmental heterogeneity is presented at fine spatial scales, Q_{ST}-F_{ST} comparisons have been effective in identifying local adaptation within 1,100 km^2 across sugar pine (*Pinus lambertiana*) populations (Eckert, Maloney et al. 2015). Approaches tailored for common garden data that further test if divergent selection out-competes neutral drift in explaining observed phenotypic differentiation without assuming that populations are equally related, i.e. Q_{ST} not a constant (Ovaskainen, Karhunen et al. 2011) such as Q_{ST}-F_{ST} *Comp* (Gilbert and Whitlock 2015) or DRIFTSEL package which additionally can handle multiple traits (Karhunen, Merilä et al. 2013) are evidently praised for strong structure is often present in natural settings. This last method found strong evidence that local adaptation modulates length of pelvic girdle and dorsal spine in the three-spined stickleback (*Gasterosteus aculeatus*) populations, with reduction in fresh-water *vs* marine habitats (Karhunen, Merilä et al. 2013).

I.3.2 Association mapping

Ecologically relevant traits are likely to have a complex determination. Even if gene flow may reduce the effects and number of loci to some extent (see above), local adaptation most likely occurs through shifts in allele frequencies at many loci. It is therefore challenging to identify its

determinants. A common approach to seek them is genome-wide association mapping (GWAS). This is widely used in human genetics with however limited detecting ability, where the use of panels of thousands of individuals still reveals only a small part of the phenotypic variation. Two caveats have been pointed out in the recent literature. First the missing heritability is largely due to rare variants with small effects that are simply undetectable with current sample sizes (Simons, Bullaughey et al. 2018). Second the problem of controlling for population stratification remains central (Berg, Harpak et al. 2019). In plants, however, the possibility of replicating strictly identical individuals in some systems, of generating offspring from controlled crosses that can be used in conjunction with GWAS, and the fact that selection may have been stronger, particularly in crops, increases the power of GWAS. Ultimately, control for population structure largely depends on the extent of population structure which varies widely in plants as well as on the trait itself which may or may not co-vary with the structure. Possible patterns of overlap between and adaptive structure, along environmental gradients, are depicted in Figure 5. Loci that determine adaptation along environmental gradients (Figure 5-a, the green allele confers an advantage in the darkening gray environmental area) will be detected by genetic association methods with different degrees of difficulty depending on the pattern of genome-wide neutral genetic structure. If neutral genetic structure is minimal (Figure 5-b) or independent to the environmental gradient (Figure 5-d), detection of adaptive alleles will have no confounding information. As opposed to, when the strong neutral structure covaries with the environmental gradient (Figure 5-c), correcting for the neutral structure will produce false negatives.

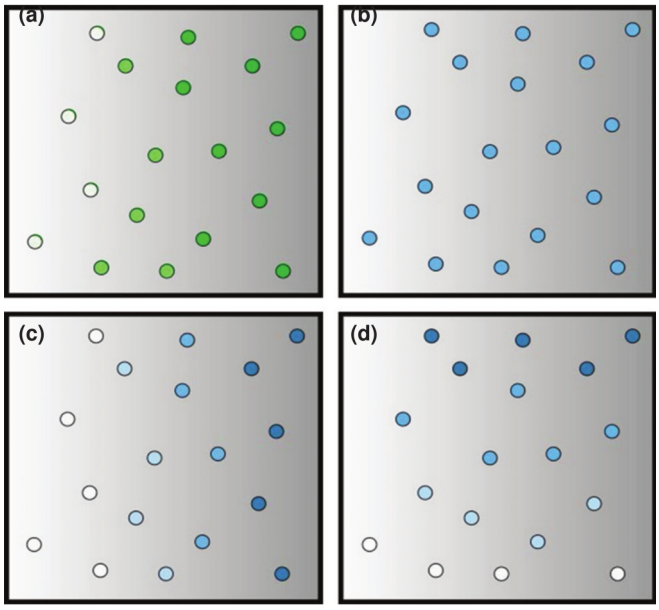


Fig. 5: Scenarios of adaptive locus overlap with neutral genetic structure along an environmental gradient. Organisms (represented by dots) are distributed in an idealized landscape with an environmental gradient (represented by the intensity of background gray shading). Dots colors intensity of shading represent genetic relatedness. In (a), variation at an 'adaptive' locus is illustrated. In this case, a green allele confers an advantage in the dark gray environments. In (b), (c) and (d), different possible patterns of genome-wide neutral variation are illustrated. In (b), gene flow is extensive, and there is little genetic differentiation across space. (c) and (d) both exhibit substantial genetic structure. In (c), neutral variation is strongly concordant with the pattern of adaptive variation illustrated in (a), possibly due to isolation by environment. Conversely, in (d), the major axes of environmental and genetic variation are orthogonal, and could be controlled for effectively when testing links between genotype and environment. Figure and legend adapted from Bragg *et al*. (2015)

In terms of overlap, flowering time can be particularly challenging, even if a number of candidate loci isolated from GWAS have been functionally validated. For instance, an association mapping study in *Arabidopsis* populations from strong altitudinal (climatic) clines has pointed to the *FRIGIDA* gene that we had presented above as a good candidate, whose different alleles affected up to 16% of the variation in climate-varying traits (Mendez-Vigo, Pico et al. 2011). It had previously been shown that this gene displays various alleles in nature, such that those found in early-flowering ecotypes are the deletions ones that disrupt the open reading frame (Johanson, West et al. 2000). An excellent example of landscape genomic approach concerns a study of geographic and climatic associations of fitness-related loci in *Arabidopsis thaliana* (Fournier-Level, Korte et al. 2011). Genotypes from accessions throughout the species range were planted into four common gardens covering a range of climate conditions. Using GWAS, the authors found SNPs significantly associated with fitness traits and demonstrated their association with climate variables while controlling for geography. They further verified that the alleles associated with higher fitness were more abundant in the planting sites closer to their population of origin (see Figure 1 in Fournier-Level et al. (2011)). They also modeled the distribution of specific alleles on the landscape. They thereby illustrate that selection across environments contributes to spatial variation in genotypes. This kind of information also has obvious utility for species that require management or conservation, such as the forest tree species discussed above.

In recent years, several authors have proposed to use the outcome of GWAS to extend our understanding of local adaptation. Because underlying structure is often an important confounding issue for GWAS analyzes and polygenic adaptation outliers recovered from such analyzes are prone to false positive results, two methods that require background genomic data have been set forward, one by Berg and Coop (2013) and one by Josephs *et al*. (2019). The first one uses GWAS outliers in Q_{ST}-F_{ST} comparisons through the statistic Q_X that takes into account background structure and can identify the populations or groups of populations that mostly contribute to the over-dispersion of genetic values (Berg and Coop 2014). Josephs *et al*. (2019) (Josephs, Berg et al. 2019) employ Berg and Coop's polygenic scores and test for excess of divergence with respect to expectations driven by population structure. Structure is here summarized by the principal components of a relatedness matrix, which is used to compute the additive genetic variance as a Q_{PC} index, itself a Q_{ST}-F_{ST} extension that can thus uncover traits that have been modeled by local adaptation. Using the Q_{PC} method on European maize landraces, the authors found that flowering time behaves as a locally adaptive trait among populations, but also found interesting evidence that this trait varies adaptively

within subpopulations (Josephs, Berg et al. 2019). Although polygenic scores calculations to define traits under spatially varying selection have been extensively used in human genetics, recent research on human height using a much less structured data set found reduced latitudinal effects and indicates that the GWAS outputs on which polygenic scores rely on tend to be loaded with false positive associations due to an insufficient correction for the underlying genetic structure (Berg, Harpak et al. 2019; Sohail, Maier et al. 2019).

I.3.3 Local adaptation along altitudinal gradients

The evolution of a combination of traits may be studied as a response to environmental gradients, Adaptation to altitude has been particularly well depicted in humans and dogs where independent mutations on the same gene (*EPAS1*) have helped them adapt to life on Tibetan highlands (Yi, Liang et al. 2010; Wang, Huang et al. 2014). Interestingly, genome scans indicate that the same metabolic pathways seem to have been selected independently in Andean highlanders (Foll, Gaggiotti et al. 2014), yet not all the physiological strategies are shared between these human groups (Petousi and Robbins 2013). Environmental changes linked to altitude include conditions that are physically linked to metres above sea level, decreasing atmospheric pressure and partial pressure of all atmospheric gases, decrease in atmospheric temperature, reduced clear-sky turbidity and higher UV-B radiation fractions. Other environmental changes that are specific to altitude but not encountered in all mountains include changes in soil composition. Finally, there are other variables that arem't specific to altitude, yet they may accompany altitudinal changes. For instance, increased altitude may be positively or negatively correlated to moisture, hours of sunshine and wind velocity, and seasonality may also change at high latitudes with increasing altitude (Körner 2007).

Elevation gradients around the globe will display rather variable clines with respect to the second and third lists of conditions, making their patterning particular depending on other factors than altitude *per se*, for example, the typologies of altitudinal trends in precipitation vary vastly for different latitudes (Körner 2007). In plants, alpine adaptation has been well documented, ranging from trees (eg. *Picea abies*) to shrubs (eg. *Arabidopsis*). As previously outlined, an attractive way to study the effects of environmental variation is through the application of common gardens to seeds collected along environmental gradients. In *Picea abies* seeds collected along eight altitudinal gradients and grown in a common garden, showed that seedlings from high-altitude populations consistently metabolized at a higher photosynthetic rate yet their performance in plant height and dry-mass diminished as a function of the altitude of origin (Oleksyn, Modrzynski et al. 1998). As

for the outcrossing *Arabidopsis lyrata,* Hamala *et al*., (Hämälä, Mattila et al. 2018) ran a series of reciprocal transplants along two altitudinal gradients. Within each of the gradients they identified gene flow, mostly from alpine populations to low altitude populations. Interestingly, fecundity promoted local superiority in fitness in the lowlands whereas in alpine populations, viability was the primary determinant of fitness differences between local and foreign populations. The same fitness traits were also selected in the alpine plant *Festuca eskia* along an altitudinal gradient (Gonzalo-Turpin and Hazard 2009). Australian alpine environments, the home of the grass *Poa hiemata,* have seemingly determined its local adaptation through the establishment of altitudinal forms. Along three gradients, several traits were favored in opposite direction between high altitude and low altitude sites (leaves were shorter and circumference size larger with increasing altitude), along with home-site advantage recorded for survival in reciprocal common garden trials (Byars, Papst et al. 2007). Along more temperate conditions, work by (Bresson, Vitasse et al. 2011) studied two tree species along two elevation gradients in the French Pyrenees and found that both the European oak (*Quercus petraea*) and beech (*Fagus slyvatica*) exhibit linear correlations of leaf traits with altitude (reduced leaf size, but increased leaf mass, stomatal conductance and leaf nitrogen content). Yet, by combining *in situ* measurements with common garden assays, they distinguished a stronger environmental over genetic effect on leaf functional traits. Studies in teosintes have described darker leaf sheaths accumulating more anthocyanin, and more abundant trichomes at higher elevations (Lauter, Gustus et al. 2004). Not so adaptively clear, maize plants also display a trend towards greater genome size with elevation (Diez, Gaut et al. 2013), a relationship that may be driven by accelerated cell division rate that in turn may confer shorter life cycles (Takuno, Ralph et al. 2015; Bilinski, Albert et al. 2018). A meta-analysis effort on plant trait differentiation and adaptation along altitude by Halbritter et al., (2018) taking into account common gardens, reciprocal transplants and genome wide studies, found that survival of genotypes was in general strongly impaired when plants were grown in foreign (different altitude) environments. Biomass unequivocally increased in lowlands for plants of any origin with plants of high elevation being shorter. The effects of altitude however revealed no preferred adaptive phenological strategy along elevation gradients, with either earlier or later seasonal development observed (Halbritter, Fior et al. 2018).

Covariates responsible for variation of plant traits along gradients are often described as a combined response or syndrome (Körner 2007). A syndrome can be defined as a suite of integrated traits that together optimize fitness (Ronce and Clobert 2012). Such patterns of covariation present different degrees of stability and may be shaped by natural selection, yet they may also carry

mechanistic constraints, and distinguishing the relative contribution of each can be complicated (Ronce and Clobert 2012). Genetic correlations between correlated traits can act as a constraint if the value of a given trait is advantageous at the cost of a detrimental effect in another, thereby impeding that both traits attain their optimal value (Shi and Lai 2015). On the other hand, genetic correlations may themselves be adaptive, that is, evolved through natural selection and thus they should be easier to decouple allowing for different correlations to be promoted under different environments/populations (Shi and Lai 2015). With respect to the emergence and maintenance of phenotypic syndromes, Legrand et al. (Legrand, Larranaga et al. 2016) studied dispersal syndromes in butterflies and concluded that the correlation between the phenotypic traits involved have a high evolutionary potential, i.e. can change rapidly when presented with different environmental conditions.

I.4 TRANSPOSABLE ELEMENTS AS FUEL FOR EVOLUTION

I.4.1 TE classification and prevalence in plant genomes

Transposable elements are selfish genetic elements that have or have had the capacity to move between different genomic locations. They were first described by Barbara McClintock in maize in 1950 (McClintock 1950). At the same time that she was analyzing maize cytology for chromosome breakage mechanisms, she noticed variegated color patterns within maize grains (Figure 6). She explained the peculiar variegated seed coat coloring as the result from the reversible alteration of color-coding gene expression governed by elements that could jump between genomic locations (McClintock 1950). Since then, transposable elements studies have come a long way with TEs being described in virtually all organisms (Wicker, Sabot et al. 2007).

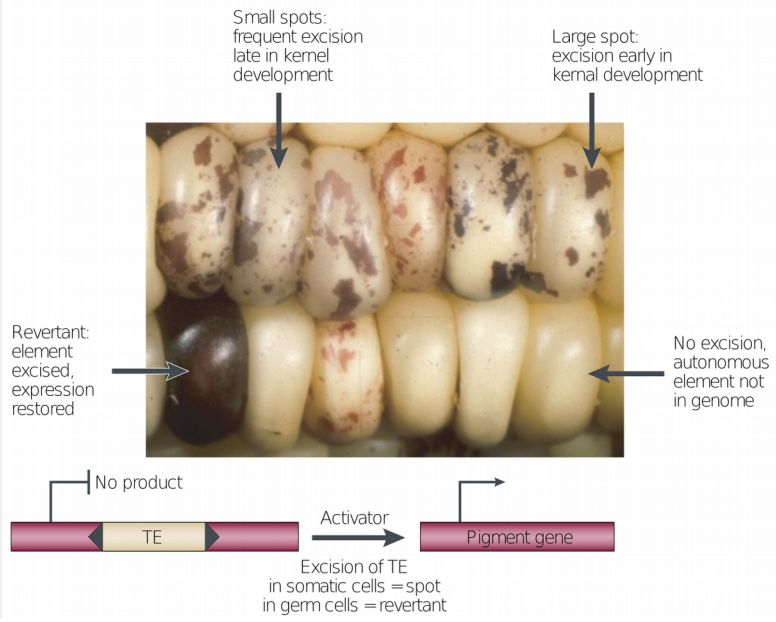
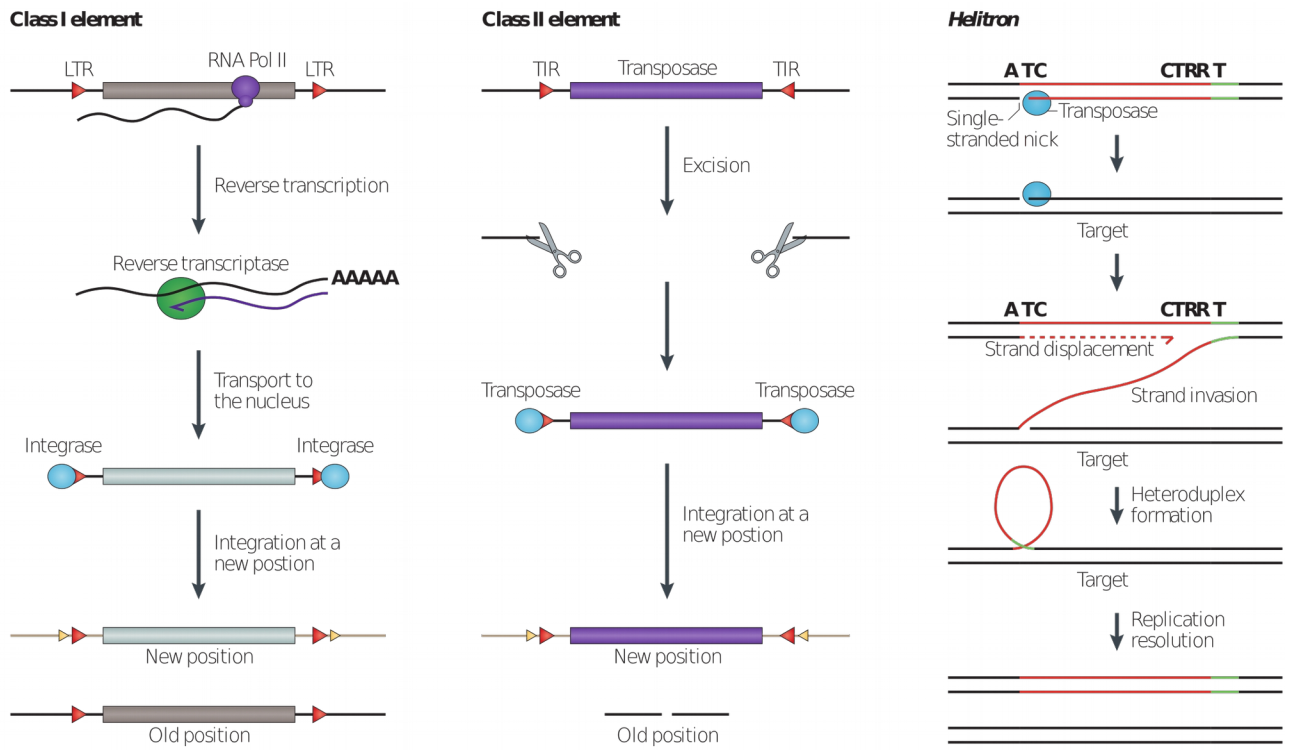


Fig. 6: Kernel phenotypes show transposon behavior. Kernels on a maize ear show unstable phenotypes due to the interplay between a transposable element (TE) and a gene that encodes an enzyme in the anthocyanin (pigment) biosynthetic pathway. Sectors of revertant (pigmented) aleurone tissue result from the excision of the TE in a single cell. The size of the sector reflects the time in kernel development at which excision occurred. Figure and description from Feschotte, *et al*. (2002).

A hierarchical classification has been proposed for TEs based on their transposition mechanisms (Figure 7), sequence similarity and structural relationships that follows the following hierarchical architecture: Class, subclass, order, superfamily, family, subfamily and taxon (Wicker, Sabot et al. 2007). Class I elements, also known as retrotransposons, require an RNA template as a transposition intermediate that is retrotranscribed into cDNA and integrated (inserted) at a new location while conserving the original copy in its location (thus their also known as 'copy and paste' mechanism) and include orders like Long Terminal Repeats (LTR)-retrotransposons, Small Interspersed Nuclear Elements (SINEs), Long Interspersed Nuclear Elements (LINEs). Class II elements or DNA transposons have a DNA intermediate and encode transposase enzyme, with which they excise themselves and reinsert in a new location ('thus moving by ‘cut and paste' mechanism), and are mainly represented by the Terminal Inverted Repeat (TIR) order (Wicker, Sabot et al. 2007). DNA transposons additionally include a subclass 2 group notable for the Helitron superfamily that has been proposed to replicate through a rolling circle mechanism similar to that of bacteria and is highly dependent on host DNA replication proteins (Kapitonov and Jurka 2001). Helitrons are not necessarily related to subclass 1 group of class I DNA transposons but their classification in this group reflects their common lack of RNA intermediate of themselves during transposition (Kapitonov and Jurka 2001).



(Caption on following page)

Fig. 7 : TEs have different transposition mechanisms. Class I elements : Retroelements that transpose via a 'copy-and-paste' mechanism. mRNA is transcribed and converted into a cDNA by reverse transcription and then integrated. Class II elements: transpose via a 'cut-and-paste' mechanism. The element is physically excised from the chromosome and reintegrated at a new location, a process that involves the transposase enzyme encoded by the TE. Helitrons: are thought to transpose via a 'rolling circle' mechanism. Figure taken from Lisch and Slotkin (2011).

In more detail, the 'copy-and-paste' mechanism RNA polymerase II transcribes mRNA from the element, which is then converted into cytosolic DNA (cDNA) through reverse transcription, then once regaining the nucleus, where it is integrated in a new position by an integrase enzyme (Figure 7, left panel). In the 'cut-and-paste' mechanism the element is excised from its current location on the chromosome and with the help of the transposase enzyme encoded in the TE. The hosts DNA double break repair mechanism ensures the process. Nonautonomous class II elements' use autonomous elements machinery, which is possible either because they are deletion derivates of autonomous elements or profit from sequence similarity at their termini to be recognized. Helitrons instead, are believed to transpose through a rolling circle mechanism in which the Helitron's terminus is nicked, and invades another region to then loop itself and obtain their partner copy by DNA synthesis (Lisch and Slotkin 2011).

TE structure according to Superfamilies is shown in (Figure 8). In retrotransposons, the protein coding genes may change order but they function in the same fashion. Upon TE integration, LTR, LINE and SINE retrotransposons as well as TIR and Maverick DNA transposons, generate Target Site Duplications (TSD), that is two short direct repeats made from the hosts code on both sides of their immediate flanking region. TSDs may present constant or variable sizes depending on their Superfamily and their presence can be used as a diagnostic feature of TE activity (Wicker, Sabot et al. 2007). Either Class I or Class II elements can include autonomous and non-autonomous elements. Autonomous elements code for all the proteins they need to effectively transpose, whereas non-autonomous elements, need the enzymes encoded by autonomous elements to be able to transpose (Lisch and Slotkin 2011).

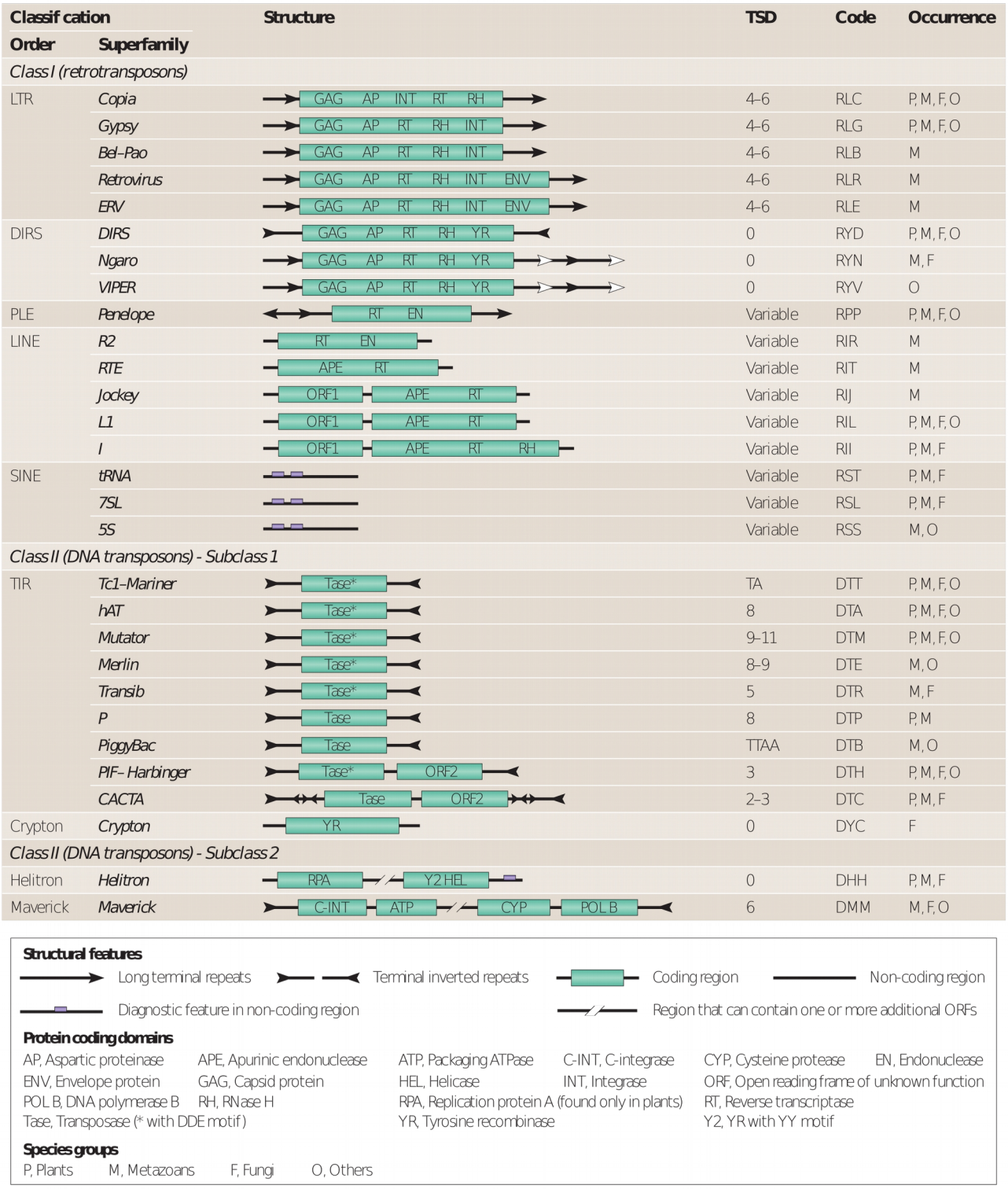


Fig. 8: Classification system for transposable elements (TEs). The classification is hierarchical and divides TEs into two main classes on the basis of the presence or absence of RNA as a transposition intermediate. They are further subdivided into subclasses, orders and superfamilies. The size of the target site duplication (TSD), which is characteristic for most superfamilies, can be used as a diagnostic feature. A three-letter code that describes all major groups and that is added to the family name of each TE. DIRS, Dictyostelium intermediate repeat sequence; LINE, long interspersed nuclear element; LTR, long terminal repeat; PLE, Penelope-like elements; SINE, short interspersed nuclear element; TIR , terminal inverted repeat. Figure and legend taken from Wicker *et al*. (2007).

In plants, while gene content varies from roughly 16,000 genes in *Chalmydomonas reinhardtii* (Merchant, Prochnik et al. 2007) to 66,000 genes in Glycine max (Schmutz, Cannon et al. 2010), genome size varies in general some 30-fold. When considering outliers, this variation reaches 2400-fold (Dodsworth et al., 2015). This huge variation is due primarily to repetitive sequences, and there is indeed a high correlation between genome size and TE content (Tenaillon, Hollister et al. 2010) (Figure 9). Most TE compartment in plants is occupied by retroelements (Figure 9).

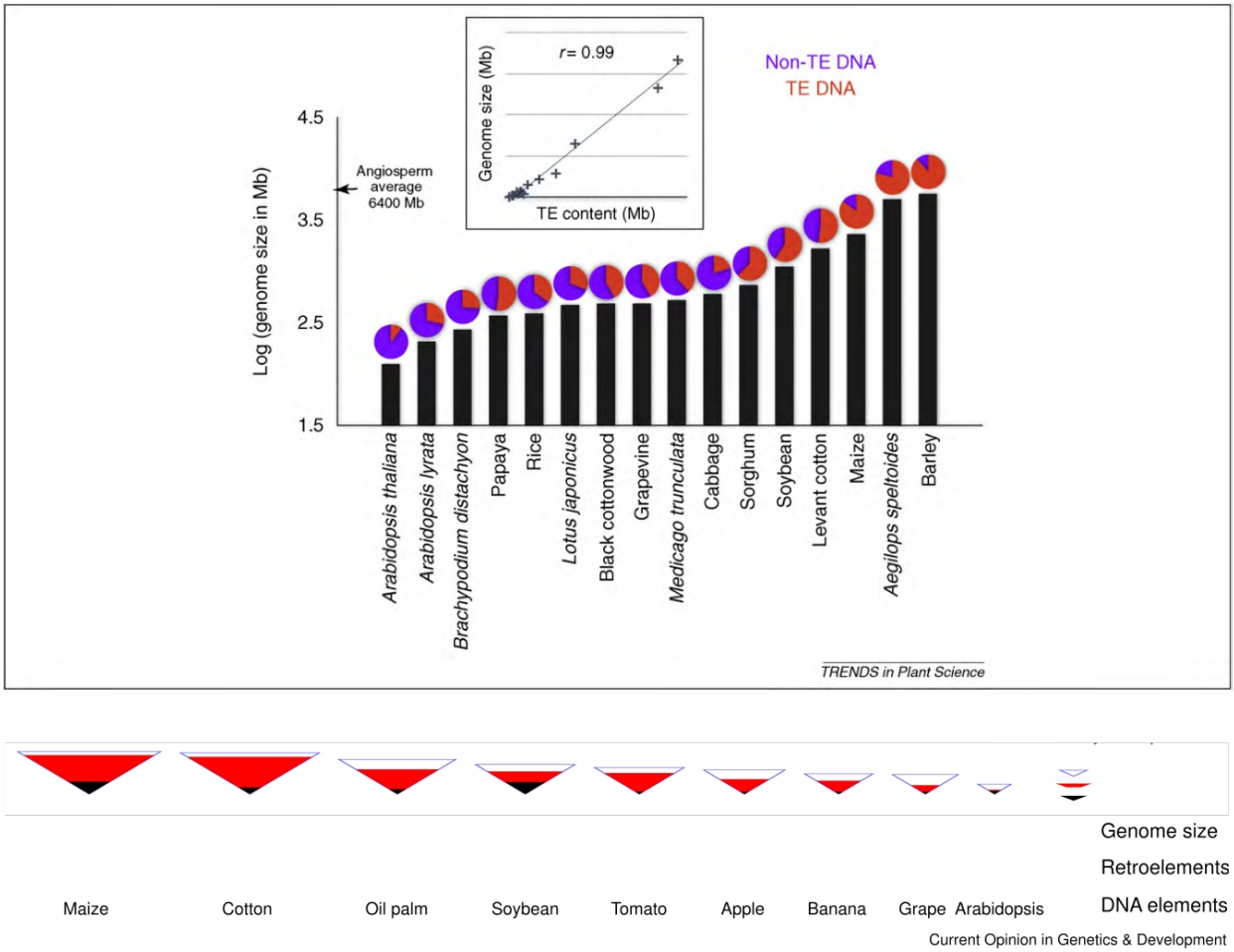


Fig. 9: Genome size and TE content in plants. Top panel: Genome size and TE content in angiosperm species are strongly correlated. All species are diploids. Image taken from (Tenaillon et al., 2010). Bottom panel: TE variation across highly annotated plant genomes. The size of the largest triangle for each species represents the entire nuclear genome size, while size of the smaller internal black triangle represents the annotated DNA TE composition and the red pigmentation represents the retroelements. The variation observation indicates major differences in TE activity and/or chromosomal DNA stability across these lineages, but is also an outcome of differences in TE annotation strategies by the genome sequencing projects involved. Figure adapted from Bennentzen *et al.*, (2018).

In plant genomes, TEs are differentially distributed across chromosomes (Figure 10), for example LTR-retrotransposons tend to be found primarily in heterochromatic regions such as pericentromeric regions, subtelomeres and knobs (Kejnovsky, Hawkins et al. 2012). Contrarily MITES (DNA transposons) can be found densely populating genes or regions close to genes, especially at their 5' position (Kejnovsky, Hawkins et al. 2012). In particular MITEs of the *Tourist* family show a preference for inserting within elements of their same family in maize and rice genome (Jiang and Wessler 2001). Also Helitrons seem to have a tendency to insert with one another in maize (Yang and Bennetzen 2009).

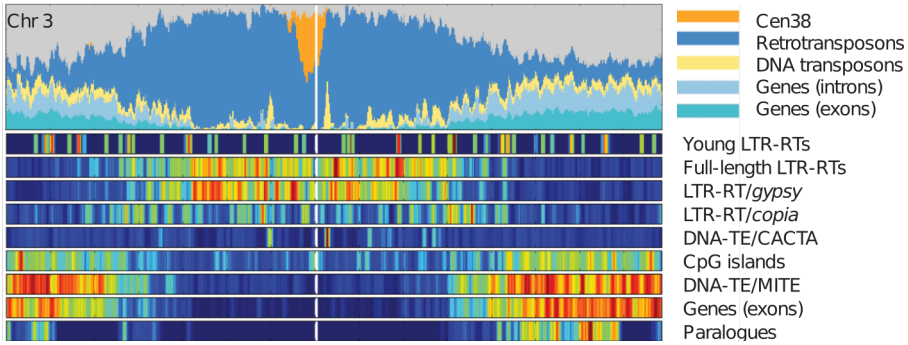


Fig. 10: Genomic landscape of sorghum for chromosomes 1. Area charts quantify retrotransposons (55%), genes (6% exons, 8% introns), DNA transposons (7%) and centromeric repeats (2%). Heat-map tracks detail the distribution of selected elements. Cen38, sorghum-specific centromeric repeat10; RTs, retrotransposons (class I); LTR-RTs, long terminal repeat retrotransposons; DNA-TEs, DNA transposons (class II). Figure taken from Paterson, *et al*. (2009).

I.4.2 TE dynamics and evolution

Self-replication of TEs leads to increase in copy number, which comes with a fitness cost. TE may disrupt gene function by landing in protein-coding or regulatory regions (Feschotte 2008). They may induce large-scale chromosomal rearrangements (insertions and deletions) through non-homologous recombination (Bailey, Liu et al. 2003). TE content therefore depends on the balance between transposition rates, host control mechanisms and elimination of TE DNA via epigenetic regulation and recombination, and population processes (Tenaillon, Hollister et al. 2010).

I.4.2.1 Transposition

Once a TE has been inserted, it may follow different fates. If it is a DNA transposon it may excise and move again to a novel site. If it an RNA transposon, a copy is kept at the site but may eventually produce additional copies that insert elsewhere in the genome. As time goes by,

mutations accumulate and TEs become increasingly fragmented, affecting their activity and recognition by bioinformatic tools (Maumus and Quesneville 2016). TE lineages that become trapped in this way will eventually become extinct (Bennetzen and Park 2018). Alternatively TEs may be 'domesticated' or co-opted by the host to perform certain services related to their DNA binding abilities, such as acting as the DNA-binding domain of plant transcription factors (Yamasaki et al., 2012). Only a few 'master' copies are transcriptionally active at any one time in *Arabidopsis* (Becker et al., 2011), as also suggested from comparison of genomic sequences and expression sequence tags data in maize (Vicient 2010). Furthermore, using mutation accumulation lines in *Arabidopsis thaliana*, (Weng, Becker et al. 2019) calculated single nucleotide mutation rates and observed more mutations falling inside TEs and pericentromeric regions, at an estimated rate of 1.36×10^{-8} inside the TEs, which doubled the genome-wide average. Theoretical models of TE life-cycles have proposed that high point mutation rates inside TEs would be a disadvantageous for them, because it could lead to sequence degradation that could compromise their further movement (Le Rouzic, Boutin et al. 2007). Weng *et al*., suggest that perhaps the high point mutation rate inside *Arabidopsis* TEs could help explain the comparatively low TE content in this small genome. Although most TEs in plant genomes are inactive, there are some interesting examples of active elements. Certainly, the discovery of transposons was lead by an active pair in maize, that in fact generated chromosome breakage upon activity. When the *Activator* (*Ac*) element was present in one chromosome, the *Dissociation* (*Ds*) generated chromosome breakage at another, with locations varying between generations (McClintock 1950; McClintock 1956). *Ac* encodes the transposase that transposes both *Ac* and *Ds*. In rice, a MITE hopscotch TE, named *mPing* was discovered to be an active element, thanks to its visible phenotypic effect (slender glumes), but also from observations on the sequenced genome, where large amount of copies of *mPing* were identical, leading to suggest they were the result of recent transposition activity (Jiang, Bao et al. 2003). *Pong* elements were discovered for their similitude to *mPing*. They are the autonomous partner, producing the transposase used by the latter. Transposon activity has been suggested to augment under stress conditions. In line with this proposal (Jiang, Bao et al. 2003) noted that *mPing* had a tendency to amplify more strongly in rice cultivars adapted to environmental extremes. More recently, approaches have been developed to estimate TE transcriptional activity - the mobilome characterization - using TEs RNAseq data. Yet TEs can also be post-transcriptionally inactivated (Quadrana, Silveira et al. 2016). Experiments on *Escherichea coli* using plasmids with TEs and fluorescent reporters have allowed for real-time TE transposition to be appreciated, as well as observing that TE activity varies greatly throughout the cell cycle (Kim, Leea et al. 2016). As for plant TEs caught 'in the jumping act', recent results in maize *bz* locus have identified large LTR

transposons moving without cues from environmental stress or epigenetic factors, yet only in pollen tissue and not in female germline or somatic tissue (Dooner, Wang et al. 2019).

I.4.2.2 Horizontal TE transfer

Besides the mentioned jumping acquisition mechanisms of TEs, they have also been known to execute larger leaps, that is horizontal TE transfers (HTTs) across reproductive barriers. HHTs are inferred adding up three criteria: sequence similarity, phylogenetic incongruence, and patchy phylogenetic distribution, each of which entails unresolved methodological and statistical issues (Loreto, Carareto et al. 2008). Although mainly described in animals and less often in plants, recently a comparative genomic survey has found that HTTs have occurred in numerous occasions in angiosperms, sometimes amplifying on arrival to the new host (Panaud 2016). Although the precise mechanism that allows for HTTs between species has yet to be elucidated, good bets are set on host-parasite interactions in view of their close relationships at the physical and chemical levels but also on virus functioning as vectors that encapsidate TEs (Panaud 2016). A recent compilation by (Gilbert and Feschotte 2018) showed that HTT seems to have a lot in common with host endogenization of viral sequences.

I.4.2.3 Epigenetic control

The vast majority TEs are not actively producing mRNAs or transposing, either because they are truncated or because they are transcriptionally silenced. Plant hosts have indeed evolved mechanisms to silence TEs thereby limiting their mutagenic potential and preventing possible damages. TE activity is suppressed, maintained or even reinforced by various processes of epigenetic silencing. Those involved chromatin condensation through histone modifications or RNA-directed DNA methylation (RdDM) which can either maintain existing methylation (where TEs are silenced by 24nt small interfering RNAs produced by inactive homologous TEs present in the genome) or by *de novo* methylation employing interfering RNAs (these 21/22 nt siRNAs are produced at the post-transcriptional level upon TE mRNA degradation) (see (Sigman and Slotkin 2016) for a review). This latter mechanism is thought to be recruited for insertions of TE not yet present in the genome. TEs that are inserted in gene-rich regions are preferentially controlled by RdDM whereas heterochromatic regions are methylated by DDM1 (Zemach, Kim et al. 2013).

A TE insertion’s genomic localization context is highly determinant of the type of epigenetic control it will be affected by. In heterochromatic regions where most TEs are inserted, all

mechanisms of TE repression are acting. Close to genes, the RdDM pathway is primarily recruited (Zemach, Kim et al. 2013).

Interestingly, methylation silencing of TEs can spread to nearby genes and repress their expression, a phenomenon known as methylation spreading. Three key observations in *Arabidopsis thaliana* have lead to the proposal that methylation spreading comes at a cost on host fitness (Hollister and Gaut 2009). First, gene expression correlates negatively with methylated TE density, second, only methylated insertions close to genes show clear signatures of purifying selection, and third, older methylated insertions tend to be located further away from genes (Hollister and Gaut 2009).

I.4.2.4 TE Removal by recombination

Mechanisms inherent to TEs structural characteristics prevent plant genomes from ever growing larger. Hence, TE removal may be achieved through intra-strand homologous recombination, which can occur either between similar TE copies in different genomic regions or between LTR motifs of the same TE originating a looped TE fragment which is excised from the genome (Kejnovsky, Hawkins et al. 2012). Given LTR-retrotransposon structure, they may engage in entanglements that lead to LTR sequences to be found on their own yet flanked by target site duplications. Two mechanisms produce solo-LTRs by unequal recombination (Devos, Brown et al. 2010): (1) intra-element homologous recombination between LTRs of the same TE forming solo-LTRs surrounded by Target Site Duplications (TSDs) (Figure 11), and (2) inter-element homologous recombination between LTR of different TEs producing solo-LTRs with no TSDs (i.e., different insertion sequences) on each side. Similarly, illegitimate recombination between non-homologous elements produces truncated elements with a single LTR and such breaks can originate from class II TE excision or through stress (Vitte and Panaud 2003).

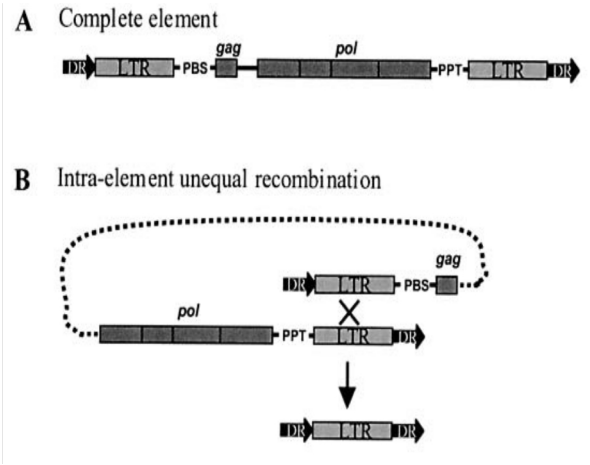


Fig. 11: Unequal intrastrand recombination between LTR retrotransposons. (A) Structure of a complete element, with a direct repeat (DR) of flanking target-site DNA, two long terminal repeats (LTRs), a primer-binding site (PBS), and polypurine tract (PPT) needed for element replication and encoded gene products (gag, pol). (B) Solo LTR resulting from intra-element recombination. The dotted line is presented to facilitate depiction of the folding needed to accomplish this recombination and does not represent any significant stretch of DNA. Figure and taken from Devos, *et al*. (2010).

TE removal by ectopic recombination has been commonly appointed as a major regulatory mechanism in Drosophila (Barrón, Fiston-Lavier et al. 2014), which implies that uncondensed genomic regions and regions of high recombination should be prone to TE loss (Tian et al. 2009), while TEs accumulate in low recombining pericentromeric regions (Petrov, Aminetzach et al. 2003). Those predictions have been verified in species such as Arabidopsis, (Pereira 2004) and tomato (Xu and Du 2014)

I.4.2.5 Purifying selection at the population level

At the population level, one important mechanism involved in TE elimination is purifying selection that purges them from genomes. Purifying selection is supposed to be acting in gene-rich regions where TE insertions have a greater probability of being detrimental, with consistent patterns being found in *Arabidopsis thaliana* (Hollister and Gaut 2009). In contrast to recombination, this population process is based on natural selection of the transposon-free allele. Its footprints can be observed on the frequency spectrum of insertions, that would tend to display an excess of rare insertions upon their removal by selection. (Figure 12). Note however that it may vary a lot among different families as shown for two species of *Arabidopsis* (Lockton and Gaut 2010).

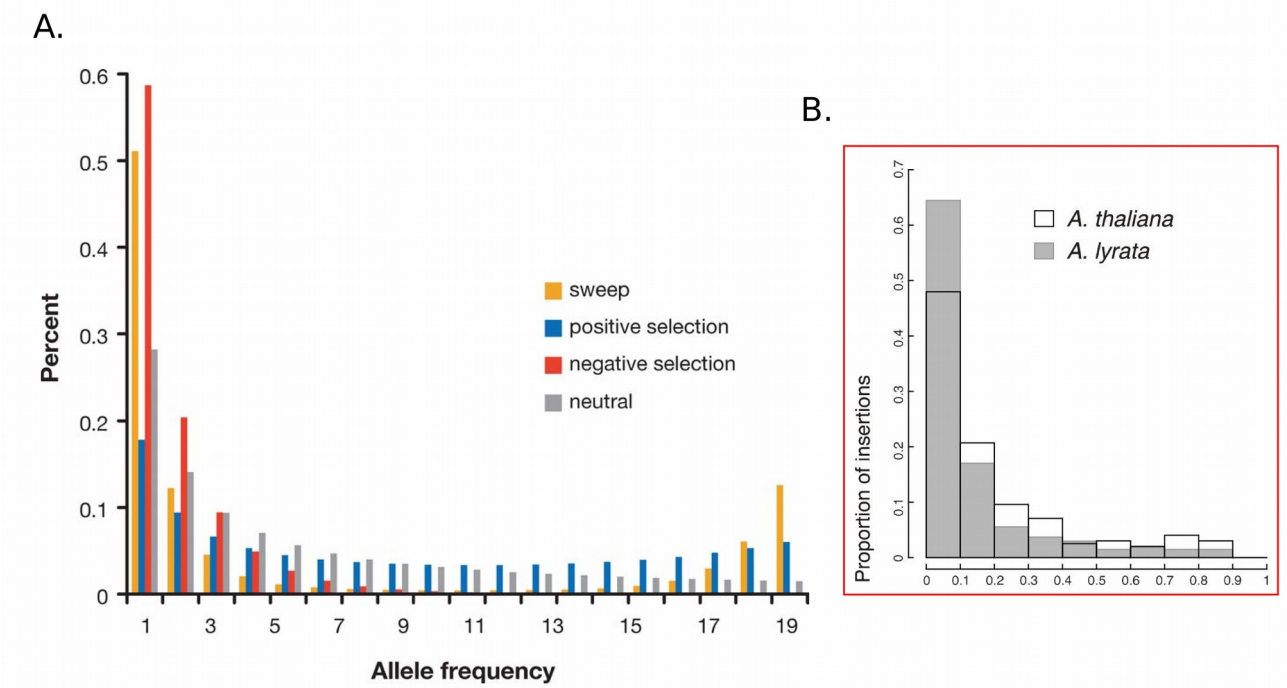


Fig. 12: The frequency spectrum under theoretical values for a selective sweep, negative selection, neutrality, and positive selection. Observed values for TE insertions in *Arabidopsis thaliana* and *Arabidopsis lyrata* Theoretical models show that, under purifying (negative) selection, an excess of rare allele frequency is expected (shown in red). The frequency spectra were calculated by Nielsen (2005) on theoretical selection models and considering a demographic model of a population of constant size with no population subdivision (A). Figure A from Nielsen (2005). In the inset, the site frequency spectrum calculated on *A. thaliana* (black outline bars with 'transparent' filling) and *A. lyrata* (gray outlined and gray filled bars) TE insertions (B). Figure B from Lockton and Gaut (2010).

Comparisons between the selfer *A. thaliana* (125Mb) and outcrosser *A. lyrata* (>200Mb) support the action of purifying selection. The stronger skew of the TE insertion site frequency spectrum in *A. lyrata* with respect to *A. thaliana* (Figure 12-B), is consistent with selection against insertions being more effective in species with outcrossing reproductive strategies (Lockton and Gaut 2010). In *Brachypodium distachyon* populations, the demographic bottleneck they have encountered is insufficient to explain the skew of TIP transposons towards rare insertions with respect to SNPs, perhaps better explained by purifying selection (Stritt, Gordon et al. 2017). Other studies have also reported that at population scale, 216 *Arabidopsis thaliana* accessions present mostly rare TE insertions, yet at those TE insertions that are common, altered expression of neighboring genes was observed as well as methylation differences (Stuart, Eichten et al. 2016). In *Arabidopsis thaliana*, a survey among 211 accessions described that their mobilome (the set of TE families with transposition activity) were mostly shared at pericentromeric regions and were mostly specific at the chromosome arms, leading the authors to propose that TE content is the result of TEs inserting indistinctly along the genome, that are then purged from gene-rich regions by purifying

natural selection making them accumulate in pericentromeric regions. (Quadrana, Silveira et al. 2016). Support for these explanations could profit from insertion age calculation as well as population frequencies estimations. Typically TE age has been calculated for LTR retrotransposons by comparing their LTRs, since they are identical copies upon insertion that can accumulate mutations independently allowing their divergence to serve as proxy of their insertion age (Sanmiguel, Gaut et al. 1998). Other methods in course of development aim to overcome the LTR restriction by calculating the divergence between consensus sequences and different copies of a TE as a proxy of age (Maumus and Quesneville 2016).

The strength of purifying selection depends on the effective population size, the expectation being a faster accumulation of TEs in species with small population sizes where TE purging is less efficient. Along the same line, mating system has been proposed to affect TE dynamics, but the outcome for highly homozygous selfing species is still hard to predict for they could have higher TE copy numbers due to a lower probability of ectopic recombination when homologous partner alleles are present or they could have lower copy numbers due to the deleterious effects of recessive TE insertions (Lockton and Gaut 2010). In a comparative study, selfer *A. thaliana* has evolved TE families with higher allele frequencies and lower selection coefficients relative to outcrossing *A. lyrata*, suggesting a reduced efficacy of natural selection on the selfer which could be partly explained by the fact that selfing diminishes the effective population size and inbreeding reduces the effective recombination rate thus reducing the efficacy of natural selection (Lockton and Gaut 2010).

I.4.3 Phenotypic impact of TEs

TEs impact genome dynamics and phenotypic changes in various ways (see (Oliver, McComb et al. 2013) for a list of TEs responsible phenotypic changes in domesticated angiosperms). This will depend on where the TE inserts itself, accordingly it can alter gene function when inserting in exons, gene expression when it falls in the 5' region of a gene (over-expressing it when it carries a promoter or when it disrupts an inhibitory sequence, or inactivating it when falling in an activating region), while insertions in introns can produce exonization, premature ends, alternative splicing, anti-sense transcription and gene silencing via methylation spreading (Casacuberta and González 2013). Since TEs may mobilize regulatory sequences such as transcription factor binding sites and promoters, their insertions may act in *cis* and create or expand gene regulatory networks, as has been recorded under stress conditions (Makarevitch, Waters et al. 2015). Recently, an interesting hypothesis proposes that larger genomes should present more functional space to produce variations

that have phenotypic effects, such that phenotypically associated loci drawn from the intergenic regions far from genes will be more abundant for teosinte's large genome with respect to, for example, *Capsella grandiflora* a smaller one (Mei, Stetter et al. 2018). Also, in view that the mutation rate indeed increases with genome size, teosintes are predicted to present lower chances of showing a hard sweep signal with respect to *C. grandiflora* (Mei, Stetter et al. 2018).

Since TE insertions can disrupt genes and reprogram gene expression, they have often participated as the underlying factors of artificially selected traits in domesticated plants. For instance, a bibliographic survey shows a 50% of domestication and diversification TE insertions were involved in gene disruption (Oliver, McComb et al. 2013). In fact, sometimes convergence has been observed between different grass species acquiring alternative TE insertions in the same gene to obtain low amylose, sticky and waxy grain traits (Varagona, Purugganan et al. 1992; Kawase, Fukunaga et al. 2005) (Hori, Fujimoto et al. 2007). As for elements located at regulatory regions, some examples show striking phenotypic impacts. In maize plant architecture, the emblematic TE insertion at the *tb1* promoter region gives a sole stalk plant, a phenotype strongly selected during maize domestication (Studer, Zhao et al. 2011).

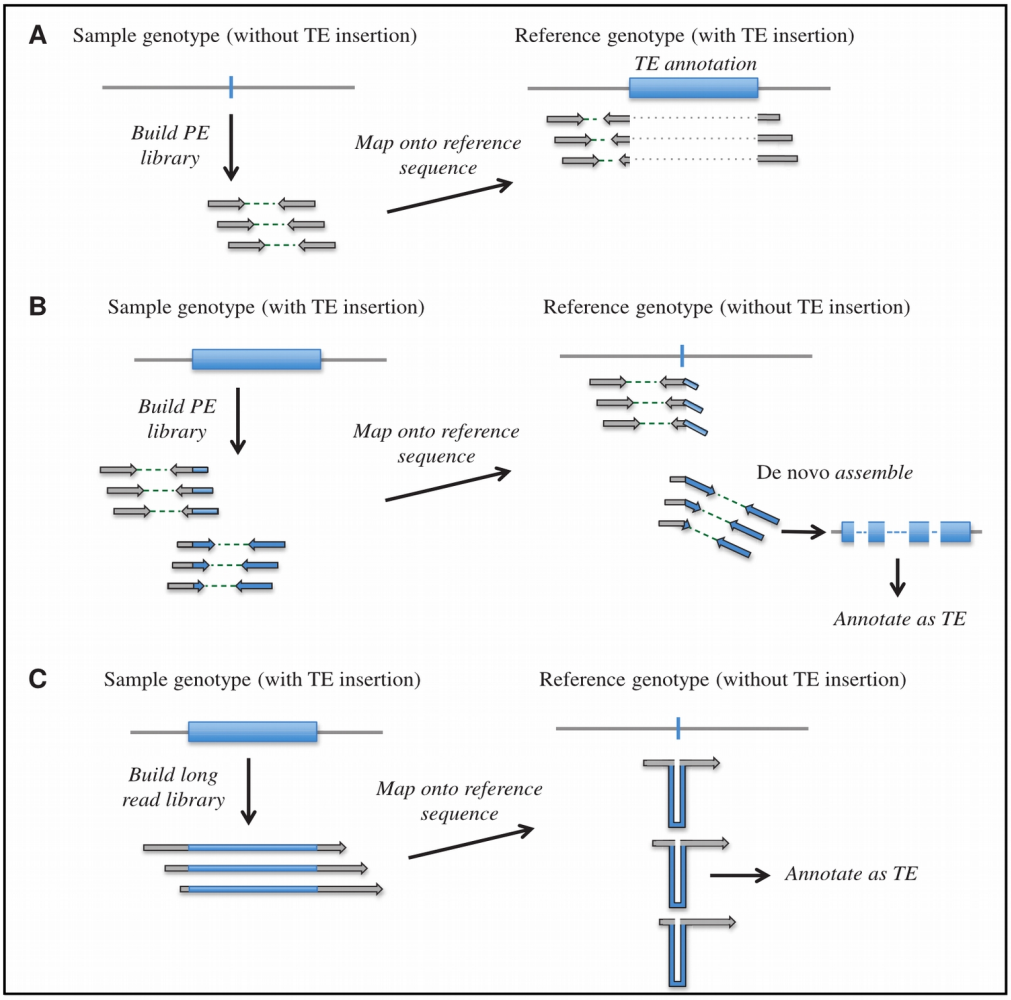
Fruit color has also shown drastic modifications, as seen for example in grapes, apples and oranges. In the case of grapes (*Vitis vinifera*), drastic modifications are linked to the insertion of a retrotransposon in *VvmybA1* – a Myb-related gene that regulates anthocyanin biosynthesis – that associates with color loss in grape fruit skin color (Kobayashi, Goto-Yamamoto et al. 2004). Similarly, a recent assembly of apple genome (*Malus domestica*) on the red colored Hanfu variety (HFTH1) has been compared to the Golden Delicious (GDDH11) reference genome and among the insertions, deletions and inversions identified, authors have discovered a gypsy-like LTR retrotransposon that is likely at the origin of red color (Zhang, Hu et al. 2019). In this case the insertion is located ~3 kb upstream of its target, a transcriptional activator (MdMYB1) of anthocyanin biosynthesis, and at least twelve apple variety trials consistently found the insertion only in red skinned varieties (Zhang, Hu et al. 2019). Sicilian blood oranges (*Citrus sinensis*) are an interesting case because not only has a copia-like retrotransposon insertion close to *Ruby* gene been found to cause their unusual red fruit color, but, in order to do so, it must be environmentally activated through exposure to low temperatures (Butelli, Licciardello et al. 2012). This active retrotransposon is released from its repression when cold stress is perceived by the plant host, thus affecting the transcriptional activator of anthocyanin production where this TE lays (Butelli, Licciardello et al. 2012). Curiously, Chinese Jinxian blood oranges have also attained their red color

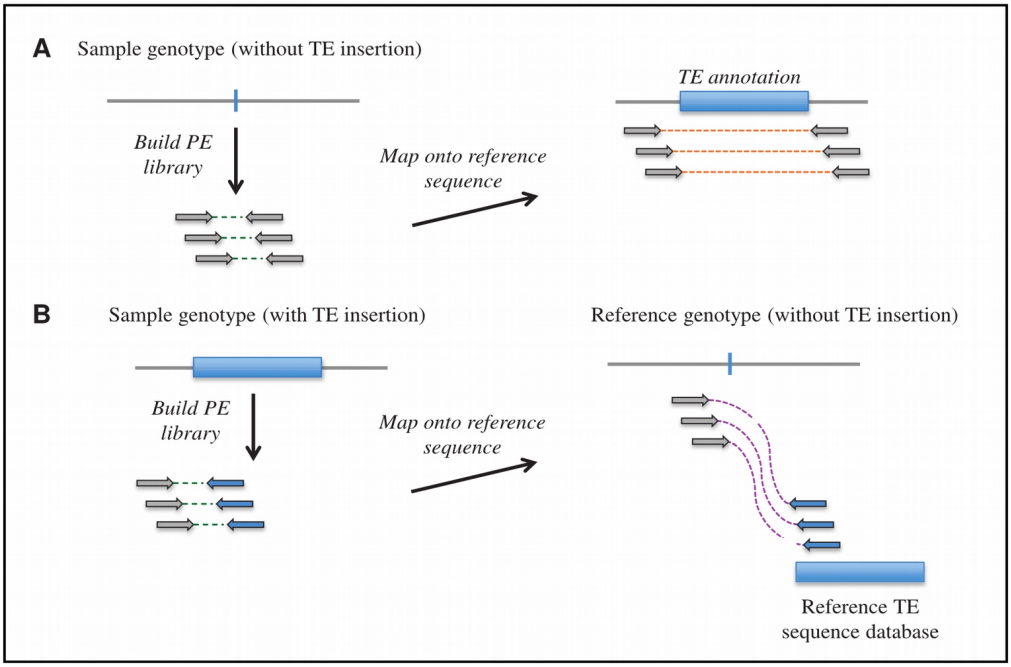
through an independent insertion of a very similar 5Kb long retroelement at 450 base pairs upstream from the same promoter gene (Butelli, Licciardello et al. 2012).

Transposable elements have also been known to affect flowering time. A multi-accession *A. thaliana* mobilome typification study reported seven COPIA family TE insertions at the locus of the FLC gene which in turn generates natural variation in flowering onset, apparently linked to the insertions as major effect alleles that lead to reduced FLC expression making plants flower earlier (Quadrana, Silveira et al. 2016). One such insertion family ATCOPIA78, was additionally observed to have copy numbers that varied with the annual temperature range adding to its mobilization adaptive potential (Quadrana, Silveira et al. 2016). Likewise, in *Arabidopsis* the ONSEN retrotransposon (the homeologue of ATCOPIA78) is known to regulate nearby genes and is activated by heat stress (Cavrak, Lettner et al. 2014). As an example of a TE insertion in a regulatory region with phenotypic effects that are driven by environmental cues there is the case of maize, where a CACTA-like TE insertion in the promoter region of the *ZmCCT* gene has been found to be an indispensable post-domestication acquisition that enables temperate zone plants to reduce their photoperiod sensitivity for flowering time onset (Yang, Li et al. 2013).

I.4.4 TE detection

TE annotation in large genomes (>1Gb) are challenging for two reasons, firstly a vast part of the unsequenced/unassembled material of most genome assemblies are repetitive sequences; and secondly rapid TE degradation often impedes their identification (Bennetzen and Park 2018). In order to investigate population processes involved in TE evolution, one needs to characterize TE insertion-deletion polymorphism. Methods most commonly rely on High-Throughput Sequencing (HTS) data and an available annotated reference genome as well as TE annotation databases (Goerner-Potvin and Bourque 2018). A difficulty encountered when mapping short reads onto TEs is that read length can be completely contained in the TE, and since TEs can be found in multiple copies throughout the genome, this may provoke reads to record multiple hits or alternatively, reads belonging to different copies of the same TE may cluster together onto one specific copy, leading in either case to spurious mapping. This is why TE flanking regions provide highly valuable information when searching for TE insertions. A number of methods take advantage of these flanking regions (Figure 13).





(Caption on following page)

Fig. 13 : Detection of TE presence/absence polymorphisms from HTS data. Upper panel : Split-read-based methodologies. TE sequence is represented as a rectangle, and empty site of insertion is shown as a vertical bar. (A) Detection of a TE insertion that is absent from the sample genome using 'split read' signature. (B) Detection of a new TE insertion in the sample genome using 'split-hanging reads'. (C) Detection of a new TE insertion in the sample genome with the use of long reads. (PE stands for paired-end). Lower panel: Complete read methodologies. (A) Detection of a TE insertion that is absent from the sample genome using long inner distance mapped reads (B) Detection of a new TE insertion in the sample genome using one end anchored reads. Figure and legend adapted from Vitte, *et al*. (2014)

When sequencing data is available in the form of short paired-end reads the following options may apply. If an insertion is present in the reference genotype but absent in the sample, a read that spans the insertion point will be split, with both fragments mapping separately at an interval of the TEs size (see Figure 13 upper panel A). The non-split read counterpart of this method is found in Figure 13 lower panel A, where pairs of reads are both mapped, but at a larger distance than that determined though their design. When this happens at a location annotated as a TE, the TE insertion is declared absent from the sample. Now, for the case where the TE insertion is present in our sample but not in the reference genome, if one of the reads is split with a fraction of it mapping and a fraction is not, the non mapped part may be clustered and *de-novo* assembly can lead to a new TE annotation (albeit rather restricted to small TE insertions) (Figure 13 upper panel B). Otherwise, reference TE sequence databases can be used to blast the clustered 'hanging' pieces and/or, as in Figure 13 lower panel C, this can be performed for cases where one read maps and its mate does not but in fact maps to a TE in the database.

Some of the paired-end read inconveniences can be circumvented by the use of long reads, which may more accurately map to a unique position. Nevertheless, as seen in Figure 13 (upper panel C) this procedure requires reads to be larger than TE length, which for long TEs such as LTR retrotransposons for example, is hard to accomplish.

Appealing variations of the outlined principles on short reads include bioinformatic treatment on the reference sequence before mapping the reads with the objective of targeting data informative positions. For example, masking the repeated sequences and then applying the one mate mapped the other unmapped approach as in Figure 13 lower panel B, but for the case where TEs insertion is present in the sample but also in the reference. Other alternatives concatenate reference annotation TE flanking regions and perform local mapping to take advantage of split reads to define exact insertion coordinates.

In summary, whole genome sequence data from short paired-end reads allows to detect structural variants which with additional evidences can identify a non-reference TE insertion as the

cause following three main methodologies, 1) mapped reads are discordant between pairs, yet this method does not yield exact insertion points, 2) split reads are clustered and share alignment junctions, this method can pinpoint exact insertion coordinates 3) sequence contig assemblies (Ewing 2015). As a general rule, it is usually recommendable to filter for TE insertions that map to coordinates where the same TE subfamily is recorded on the reference genome, when such information is available (Ewing 2015).

I.5 THE ZEA MAYS MODEL

The genus *Zea* (Poaceae) is a member of the grass family, and it is classified into two sections: *Luxuriantes* and *Zea*. Section *Luxuriantes* includes perennial species *Zea diploperennis* and *Zea perennis* (the only autotetraploid with 40 chromosome pairs), as well as annual, flood tolerant species *Zea luxurians* (southeastern Guatemala), (Doebley and Iltis 1980);(Iltis and Doebley 1980) and *Zea nicaraguensis* (geographically isolated to Western Nicaragua) (Iltis and Doebley 1980). Section *Zea* includes *Zea mays* species only, which encompass four subspecies: *Zea mays* ssp. *huehuetenangensis*, found only in western Guatemala, *Zea mays* ssp. *mexicana*, distributed along highlands of the Mexican Central Plateau, *Zea mays* ssp. *parvilgumis*, found along Mexican southwest lowlands and the cultivated maize *Zea mays* ssp. *mays* (Doebley and Iltis 1980); (Iltis and Doebley 1980). All *Zea* are commonly known as 'teosintes' with the exception of cultivated maize. Efforts of phenotypic and ecogeographic characterization clearly separate subspecies with some degree of additional sub-structuring in *Z. mays* spp *mexicana* into four races (Sanchez, Kato Yamakake et al. 1998),(De Jesús Sánchez González, Corral et al. 2018; Rivera-Rodríguez, de Jesús Sánchez González et al. 2019) : race Central Plateau and race Chalco with large distributions along central highlands, and small isolated northern races Nobogame and Durango. Besides their different ecological niches, scarce description of teosintes *Zea mays* ssp. *mexicana* and *Zea mays* ssp. *parvilgumis* indicate that they hardly distinguishable phenotypically. Some of the differences may include a larger area of sheath leaf pigmentation in *Zea mays* ssp. *mexicana* (Lauter, Gustus et al. 2004). Previous estimates indicate that *Zea mays* ssp *mexicana* and *Zea mays* ssp *parviglumis* were separated from each other for ~ 60,000 y BP (Ross-Ibarra, Tenaillon et al. 2009). As for the relationships among these two lineages, recent work on population SNP data tested different demographic inference scenarios and complemented with environmental data found stronger support for an early branching from *parviglumis* Balsas populations with an

ongoing ecological speciation process between subspecies, with recurrent geneflow and/or secondary contact. (Aguirre-Liguori, Gaut et al. 2019).

I.5.1 Maize evolutionary history

Maize was domesticated from teosinte *Zea mays* ssp. *parviglumis* in southwest Mexican lowlands about 9,000 years ago according to genetic (Wang, Stec et al. 1999 ; Matsuoka, Vigouroux et al. 2002) and archaeological findings (Piperno and Flannery, 2001). Evidence of maize starch on grinding tools dating from 8700 y BP was recovered at the Xihatoxtla cave in the Balsas Valley, southwestern Mexico (Piperno, Ranere et al. 2009). The oldest maize cob fossils were found in Guilá Naquitz cave in the state of Oaxaca and dated 3,200 y BP (Benz 2001), as well as in the San Marcos cave in Tehuacán in the state of Puebla were they dated back to approximately 5000 y BP. Sequencing of the Tehuacán ancient genomes revealed evidence of a yet incomplete domestication with inbreeding traces, for instance quasi-absent nucleotide variability with respect to Balsas teosinte at domestication loci *Teosinte branched1* (*Tb1*) and *Brittle endosperm 2* (*Bt2*), yet only partially reduced diversity at *Teosinte glume architecture* (*Tga1*) and *Sugary1*(*Su1*) (Vallebueno-Estrada, Rodríguez-Arévalo et al. 2016). Microsatellite phylogenetic analyses on more than 250 plants that include maize landraces from all around the American continent as well as *parviglumis* and *mexicana* teosintes point to a single domestication of maize located in the Balsas basin in southwest Mexico in the state of Guerrero where *parviglumis* populations share the highest genetic resemblance to maize and are basal to this crop in the phylogeny (Matsuoka, Vigouroux et al. 2002) (Figure 13). Scenarios of stratified maize domestication have been recently proposed, where after initial domestication stages in the Balsas basin, the maize plants that were taken to the southwest Amazon basin about 6500 yr B.P. were a partially domesticated crop that was further anthropogenically improved before the divergence of two South American groups (Kistler, Maezumi et al. 2018). After its initial domestication, maize has undergone a rapid diffusion throughout the American continent (Vigouroux, Glaubitz et al. 2008; Da Fonseca, Smith et al. 2015). Clearly, maize diffusion was accompanied by local adaptation to day-length and cooler temperatures.

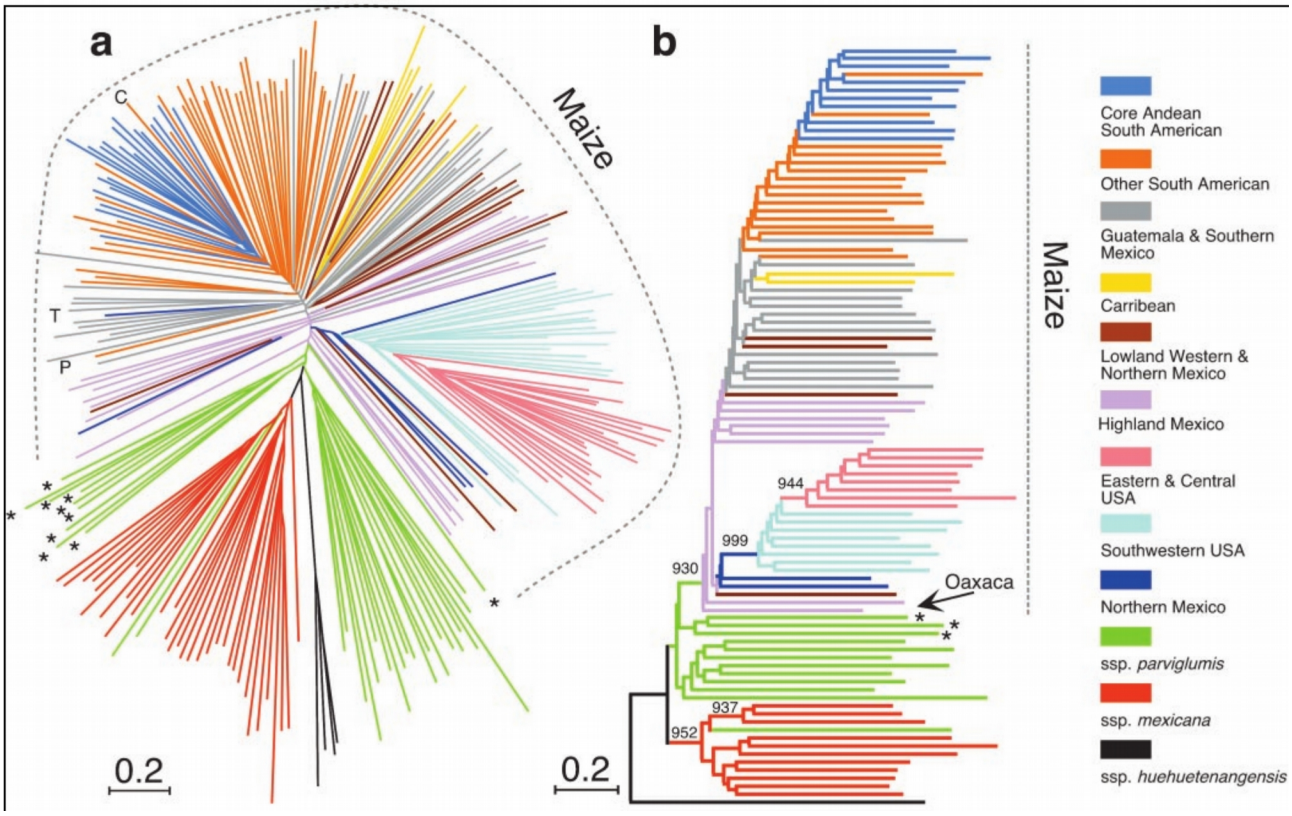


Fig. 13: Phylogenies (genetic distance trees) of maize and teosinte rooted with ssp. *huehuetenangensis* based on 99 microsatellites. Dashed gray line circumscribes the monophyletic maize lineage. Asterisks identify those populations of ssp. *parviglumis* basal to maize, all of which are from the central Balsas River drainage. (a) Individual plant tree based on 193 maize and 71 teosinte. (b) Tree based on 95 ecogeographically defined groups. The numbers on the branches indicate the number of times a clade appeared among 1,000 bootstrap samples. Only bootstrap values greater than 900 are shown. The arrow indicates the position of Oaxacan highland maize that is basal to all of the other maize. Figure and legend taken from Matsuoka, *et al*. (2002)

Maize domestication has resulted in important morphological changes (Figure 15). Those include a strong apical dominance in maize with the emergence of a single tiller as opposed to many tillers and lateral branches in teosintes. The tiller of maize is terminated by the male inflorescence, the panicle. Teosinte lateral inflorescences are numerous and often composed of male and female inflorescences while in maize lateral branches are condensed and terminated by a single female inflorescence producing the ear. In maize, the ear is composed of rigid and polystichous rachis bearing multiple rows with hundreds of large naked grains. In teosintes, the ear is constituted by a single row of a handful of grains covered by a hard cupulate fruitcase. In maize, grains have lost their dormancy and do not shatter at maturity.

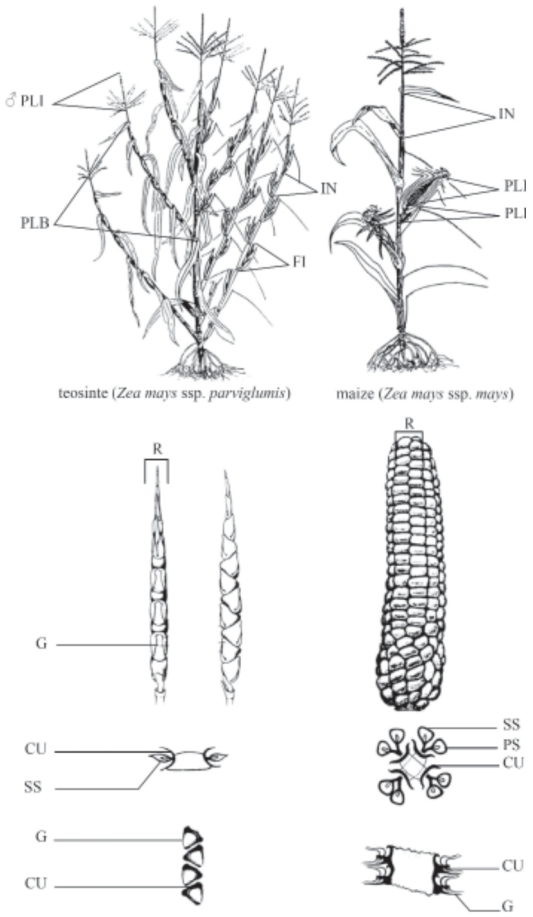


Fig. 15: Major morphological differences between teosinte (left) and maize (right). Plant architecture (top) shows number of primary lateral branches (PLB) or tillers, length of internodes (IN) of the PLBs, number of female inflorescences (FI) on the PLBs and sex of the inflorescence that terminates the PLB (PLI). Female inflorescence characteristics (middle) shows number of ranks (R) and cupules (CU), presence/absence of pedicellate spikelet (PS) and sessile spiklelet (SS) in each cuple, the kernels glume is also depicted (G). Teosinte kernels disarticulate and have a hard glume. Taken from Doebley (1992 and 1995) adaptations presented by Tenaillon and Manicacci (2011).

Maize has undergone a series of genetic bottlenecks potentially accompanied by a rapid exponential growth (Tenaillon, U'Ren et al. 2004; Wright, Bi et al. 2005; Beissinger, Wang et al. 2015). The reduction in genetic variability found in maize has been estimated as 20% of nucleotide diversity at genome level (Wright, Bi et al. 2005) perhaps losing important wild adaptive genetic variability. In total, it is considered that around 2% of the genome has contributed to domestication (Wright, Bi et al. 2005)

I.5.1 *Zea mays* genomes

The APGv4 annotated genome carries 22,048 orthologous gene sets to the grass common ancestor distributed in 10 chromosomes (Jiao, Peluso et al. 2017). The reference B73 maize genome measures 2.2 Gb and is the result of several rounds of genome duplications – including a recent one

that occurred after the divergence with Sorghum – that finally returned to a diploid state (Schnable, Ware et al. 2009). Using genomic fragments around 5 maize duplicated loci with orthologues in sorghum and rice, Swigonova *et al.*, (2004) estimated divergence times among the clades, and further supported the hypothesis of a tetraploid origin of maize, predating the major maize genome expansion (Figure 16) (Swigoňová, Lai et al. 2004). Some authors still differentiate two subgenomes in maize, arguing that they determine ongoing fractionation among inbred lines with one genome being systematically over-expressed (Schnable, Springer et al. 2011).

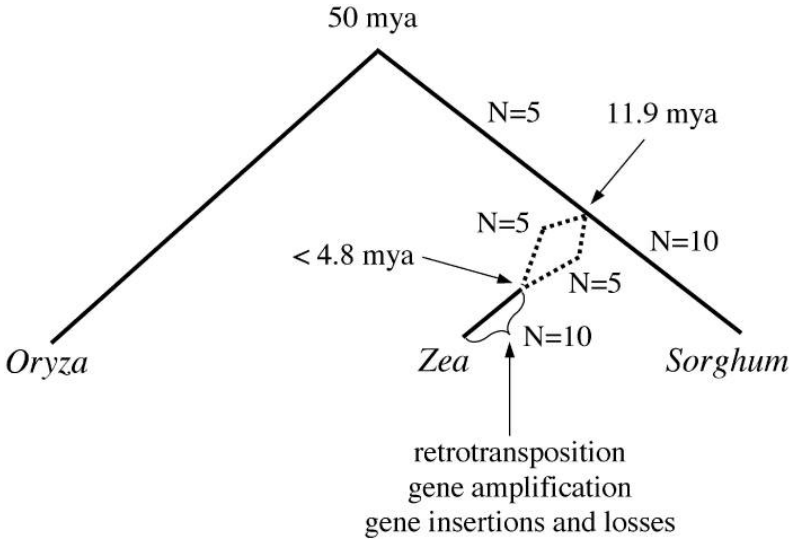


Figure 16 Hypothetical origin of maize and sorghum. *Zea* divergence from sorghum was estimated at 11.9 mya and *Zea* tetraplodization to at least 4.8 mya. Image taken from Swigoňová, *et al*. (2004)

Approximately 85% of the maize genome is composed of TEs, which are found distributed in a non-uniform fashion according to their TE family. The maize genome owes about half of its size to repeated bursts of retrotransposons in the last six million years (Sanmiguel, Gaut et al. 1998). Maize LTR retrotransposon distribution is unequal along different parts of the chromosomes (Figure 17). As mentioned, TE age estimations are possible for LTR retrotransposons by comparing the divergence between LTR pairs, since they would have been identical upon insertion (Sanmiguel, Gaut et al. 1998). Evidence from LTR divergence to calculate the insertion dates for 17 to 23 retrotransposons in maize near the *adh1* gene, has shown that all of these occurred in the last six million years, and especially during the last three million years (Sanmiguel, Gaut et al. 1998). In principle, the recent massive amplification of a few repeat families implies that many copies will be highly similar, thus allowing considerable inter-element homologous recombination, that could eliminate intervening DNA and eliminate ancient repeats at intergenic areas, leading to a potentially high TE turnover rate (Maumus and Quesneville 2016). For maize LTR retrotransposons, genome cartology methods (Estill and Bennetzen 2009) implemented by the authors that included LTR

annotation and taxonomic classification of 31,000 intact elements followed by graph-theory based clustering these TEs into families, recovered monophyletic signals for two large family clusters, *Ji/opie* and *cinful/zeon*, and found that the *huck* family encompasses two separate groups. Maize helitrons tend to carry gene fragments that they have captured (Morgante, Brunner et al. 2005). Helitron DNA transposons have been studied in detail between maize inbred lines, leading to propose that helitron's rolling circle replication mechanism more than often (approximately 10,000 events) acquires gene fragments from diverse locations and sometimes amplifies them elsewhere, as well as participating in exon shuffling and possibly the appearance of new proteins (Morgante, Brunner et al. 2005). Their impact in expressing fragments from different genes and modifying the genetic co-linearity can affect the genome's architecture, genetic networks and ultimately individual's phenotypes (Morgante, Brunner et al. 2005).

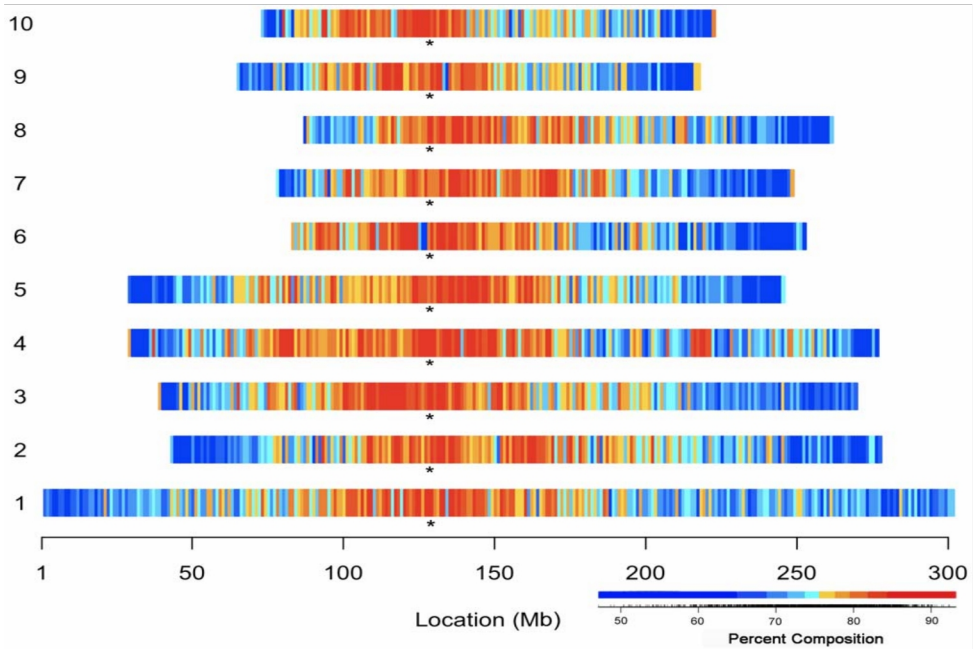


Fig. 17: The chromosomal distribution of the LTR retrotransposon (RLC, RLG, and RLX) composition of the B73 maize genome. The RepeatMasker identified LTR retrotransposons are summarized as percent composition in 1Mb bins along each of the ten chromosomes. The heatmap was derived by classifying the percent composition values into equal interval quantiles. The distribution of these classified values are illustrated as color tiles superimposed under the empirical cumulative distribution of the observed percent composition values. Asterisks indicate approximate centromere positions. Figure and legend taken from Baucom, *et al*. (2009).

The most recent TE annotation of the maize reference genome B73 (AGPv4) employed a structural identification of TEs. Its content confirms that although maize TE superfamilies show some general patterns, they nevertheless hold several families that vary in genome occupancy, frequency along the chromosome, age of insertion and tissue specificity (Stitzer, Anderson et al. 2019) (Figure 18)

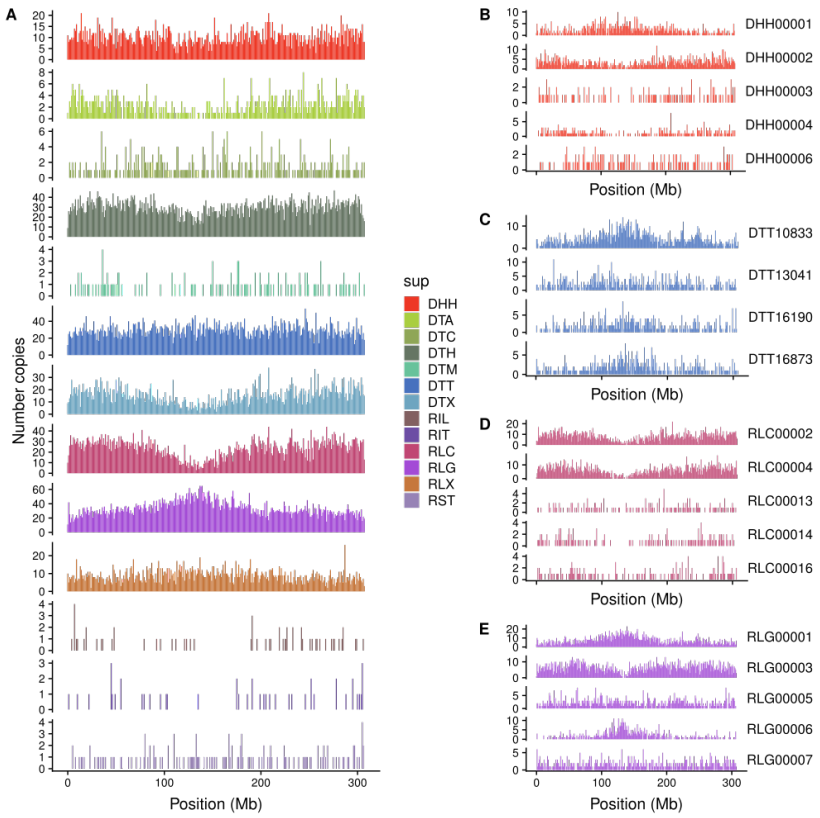


Fig. 18: Chromosomal distribution of transposable element superfamilies and example families. Counts of number of insertions in 1 Mb bins across chromosome 1 for (A) TE superfamilies and (B-D) the 5 families with highest copy number in each of four superfamilies: DHH or Helitrons (B), DTT or Tc1/Mariner (C), RLC or Ty1/Copia (D), and RLG or Ty3/Gypsy (E). Figure adapted and legend taken from Stitzer *et al*. (2019).

In maize, there are differences in genome size that have repeatedly been observed with relation to altitude (Diez, Gaut et al. 2013; Bilinski, Albert et al. 2018). Concerning causality for these differences Billinski *et al*., (Bilinski, Albert et al. 2018) measured genome size and genomic repeat abundance for 16 teosinte populations and analyzed genome size variation in correlation to cell growth phenotypic traits for one teosinte population. The authors suggest that flowering time, which was negatively correlated to the rate of cell production, is the trait being selected by natural selection and it acts on chromosomal knobs as a driver of genome downsizing with increasing altitude.

I.5.2 Gene flow across *Zea mays* species

In view of the synchronous flowering of sympatric populations of maize and teosinte in Mexican localities, it has been troubling to explain the low frequency of hybrids observed in the field (Wilkes 1977). Recently, the cross-incompatibility between maize strains linked to the *Teosinte crossing barrier-1* (*Tcb1-s*) haplotype has been proposed to be due to a gene expressed in

the pistil encoding a pectin methylesterase likely modifying the pollen tube cell wall (Lu, Hokin et al. 2019). *Tcb1-s* mainly found in wild *mexicana* teosinte populations and also registered for *parviglumis*, confers unilateral cross-incompatibility of female teosintes towards pollen from maize strains, which usually carry the *tcb1* allele. On the other hand, teosinte pollen can in fact fertilize maize plants, yet shows a competitive disadvantage with maize pollen (Evans and Kermicle 2001).

There is evidence that gene flow from teosintes to maize (Fukunaga, Hill et al. 2005) may have contributed adaptive variation to maize. For instance, as maize spread to higher grounds reciprocal introgression seemingly occurred with teosinte *Zea mays* ssp. *mexicana* conferring better adaptation to highlands (van Heerwaarden, Doebley et al. 2011; Hufford, Lubinksy et al. 2013). *Zea mays* ssp. *mexicana* indeed grows at higher altitudes than *Zea mays* ssp. *parviglumis* and is well-adapted to highlands. The direction of this gene flow is stronger from *mexicana* to maize than in the opposite direction (van Heerwaarden, Doebley et al. 2011).

I.5.3 Local adaptation of teosintes

Teosintes have been acknowledged as ideal systems on which to study local adaptation because their distributions span diverse climatic conditions, they show various degrees of population structure and phenotypic differences and profit from being the wild relatives of cultivated corn (*Zea mays* spp. *mays*) which has been extensively studied and for which reference genome annotations are available (Hufford, Bilinski et al. 2012).

Inversions are a type of chromosomic rearrangement that reduce recombination, which is why they may in turn facilitate local adaptation and speciation (Kirkpatrick 2010). If an inversion contains locally adapted alleles it can spread in a population because it avoids these alleles from recombining and diluting through gene flow with less adapted ones (Kirkpatrick and Barton 2006). In teosintes, some atractive inversion polymorphisms have been described. Interestingly they are enriched for SNPs correlating with environmental factors. Fang *et al.*, (2012) (Fang, Pyhäjärvi et al. 2012) used population genomic data from 941 SNPs on a sample of 2782 individuals encompassing domesticated maize, *mexicana* and *parvigulumis*, and evaluated LD among the markers, which pointed to a putative ~50Mb region on chromosome 1 (*Inv1n*) among teosintes, yet absent from cultivated maize. Authors described a high representation of the inversion in *parviglumis* with populations displaying frequencies as high as 90%, and a strong negative altitudinal cline among 33 populations. In addition, no recombination was recovered within the inversion when *parviglumis* and maize hybrids were formed, and the inversion was found to correlate negatively to culm diameter a trait that differentiates maize from teosinte. In a 50K (MaizeSNP50 BeadChip)

genotyping approach, Pyhäjarvi *et al*., (2013) screened 21 teosinte populations and identified many SNPs falling into genes strongly differentiated among populations many of which were associated to altitude and temperature. Furthermore, many genes were located inside the putative chromosomal inversions on chromosomes 1, 4 and 9. *Inv1n* presented clinal patterns of allele frequency across both subspecies, *Inv4m* was present in *mexicana* reflecting subspecies differentiation, *Inv9d* was enriched for SNPs associated to altitude and temperature variables yet polymorphic only in *mexicana*, and *Inv9e* associated mainly to top soil variables and precipitation seasonality also within *mexicana* (Pyhäjärvi, Hufford et al. 2013). Screening of 8,479,581 SNPs identified by whole-genome sequencing of three *parviglumis* and three *mexicana* populations recovered signatures of local adaption at 47 candidate regions along teosinte chromosomes (Fustier, Brandenburg et al. 2017). Among these, inversion *Inv1n* was recovered with 20 outlier SNPs further supporting its adaptive role. Finally, Aguirre *et al*. (2017) used MaizeSNP50 BeadChip data on 49 teosinte populations to identify SNPs potentially involved in ecological differentiation between *mexicana* and *parviglumis* (Aguirre-Liguori, Tenaillon et al. 2017). Among SNPs displaying environmental association signals, eight SNPs were located in *Inv9e* and associated to temperature in each species niche (Aguirre-Liguori, Tenaillon et al. 2017). Such inversion being identified only in *mexicana* was interpreted as a possible driver of ecological speciation between teosinte populations (Aguirre-Liguori, Tenaillon et al. 2017). In a subsequent study, authors added 9,780 DarTseq SNPs on 47 populations and found a prominent role of five putative chromosomic inversions (chromosomes 1,3,4,8 and 9) in teosinte adaptive divergence by demonstrating: 1) their overlap with F_{ST} blocks that formed islands of divergence, 2) enrichment in candidate SNPs at the inversions and 3) higher LD and signals of isolation by environment instead of isolation by distance among the SNPs contained in the inversion with respect to SNPs outside the inversions. Interestingly, phosphorous concentration of soil samples surrounding the teosinte roots was included as an environmental variable, and was found to strongly associate with candidate SNPs frequencies (Aguirre-Liguori, Gaut et al. 2019).

Some studies have shed light on a localized action of natural selection in teosintes. For instance, among six *parviglumis* populations, one of them showed evidence of a recent selective sweep on the *wip* (wound-induced serine protease inhibitor) plant immunity gene that was found to display high differentiation and low nucleotide variation as well as a substitution in the protein's active site (Moeller and Tiffin 2008). Also, soil interactions seem to have exercised a selective pressure for resistance to high soil acidity in maize and teosinte, possibly linked to a tandem

duplication of *MATE1* gene, a gene known to participate in toxic compound elimination (Maron, Guimarães et al. 2013).

Other interesting patterns recovered for teosintes at SNP level include, the elevation differences that correlate with genetic differentiation on 978 SNP loci between 61 teosinte populations, perhaps owing in part to proposed bottlenecks in some *mexicana* highland populations and populations from putative hybrid zones between these subspecies (Bradburd, Ralph et al. 2013). On the afroementioned study of Aguirre et al., 2017 on 39 teosinte populations, authors distinguished that niche border popualtions were enriched in candidate SNPs (Aguirre-Liguori, Tenaillon et al. 2017). Finally, in view that putative inversions were highly populated with candidate environmentally associated SNPs, they've been signaled as fundamental to gene-flow reduction between locally adapted populations enhancing teosintes genomic differentiation (Aguirre-Liguori, Gaut et al. 2019).

I.6 OBJECTIVES

Patterns of genomic variation in teosintes have been described in previous studies using both restricted sets of ascertained SNPs and whole genome sequencing data. Altogether these studies have revealed interesting features: they have pointed to extensive local adaptation at restricted geographical scale, they have highlighted the role of chromosomal inversions, they have contributed to establish list of potential candidate nucleotide polymorphisms, some of which are concentrated in genomic regions of particular relevance for local adaptation. The objectives of my PhD were to undertake a step further into the characterization of the ecological and genetic determinants of local adaptation in teosintes by (1) describing the extent of phenotypic variation among populations, characterizing the traits evolving under spatially-varying selection, linking variation at candidate loci to adaptive phenotypes; and by (2) performing a first population-level description of transposons in teosintes along with the detection of potentially adaptive insertions. Transposons may indeed be relevant in local adaptation processes owing to their potentially higher mutation rate than nucleotide polymorphisms and their phenotypic impacts.

I divided my PhD document in two main chapters that made use of the same datasets to address different questions. Chapter 1 is presented as a manuscript that was recently accepted for publication in PLOS genetics journal. This chapter is the continuation of a work that was initiated by Margaux-Alison Fustier, a previous PhD student. Firstly, I wished to inquire if teosinte phenotypic variation along altitude hinted to local adaptive processes. I then wished to answer how can previous work on population genomics analyzes on a subset of high-throughput sequenced teosinte populations gear the deciphering of explanatory loci underlying teosinte local adaptation when confronted with climatic and common garden phenotypic data measurements. That is to say, are past selection signatures on a small set of strategically chosen populations useful to uncover present day species-wide adaptive polymorphisms with links to the environment and/or phenotypes?

Chapter 2 is presented as a draft manuscript for which we are still gathering additional experimental data. This chapter has been particularly challenging as detection of transposable elements in complex genomes is still as its infancy. I was nevertheless able to achieve methodological improvements and to provide the first answers to the following questions: How different are contrasted altitude teosinte populations with regards to their transposable element content? How can we exploit pooled sequencing data on a few teosinte populations to extract a set of candidate adaptive transposable element insertions? Can we characterize teosinte population

frequencies of transposable elements absent from the maize reference genome? How do transposable element insertions with known phenotypic effects in maize “behave” in teosintes? This last question relates to the broader context of crop domestication. I therefore contributed to the writing of a review paper on plant domestication processes highlighting convergent and particular patterns in their genetic and phenotypic mechanisms. This paper has been published in Comptes Rendus Biologies Vol. 339 (2016) and is included as Annex II.

I.7 REFERENCES

Aguirre-Liguori, J. A., M. I. Tenaillon, et al. (2017). "Connecting genomic patterns of local adaptation and niche suitability in teosintes." Molecular Ecology **26**: 4226-4240.

Aguirre-Liguori, J. A., B. S. Gaut, et al. (2019). " Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (*Zea mays parviglumis* and *Zea mays mexicana*) " Molecular Ecology: 2814-2830.

Anderson, C. D., B. K. Epperson, et al. (2010). "Considering spatial and temporal scale in landscape-genetic studies of gene flow." Molecular Ecology **19**: 3565-3575.

Anderson, J. T., J. H. Willis, et al. (2011). "Evolutionary genetics of plant adaptation." Trends in Genetics **27**: 258-266.

Bailey, J. A., G. Liu, et al. (2003). "An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications." American Journal of Human Genetics **73**: 823-834.

Bank, C., G. B. Ewing, et al. (2014). "Thinking too positive? Revisiting current methods of population genetic selection inference." Trends in Genetics **30**: 540-546.

Barrón, M. G., A.-S. Fiston-Lavier, et al. (2014). " Population Genomics of Transposable Elements in Drosophila " Annual Review of Genetics **48**: 561-581.

Baucom, R. S., J. C. Estill, et al. (2009). "Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome." PLoS Genetics **5**.

Bay, R. A., N. Rose, et al. (2017). "Predicting Responses to Contemporary Environmental Change Using Evolutionary Response Architectures." The American Naturalist **189**: 463-473.

Beaumont, M. A. and D. J. Balding (2004). "Identifying adaptive genetic divergence among populations from genome scans." Molecular Ecology **13**: 969-980.

Beissinger, T. M., L. Wang, et al. (2015). "Recent demography drives changes in linked selection across the maize genome." bioRxiv: 31666.

Bennetzen, J. L. and M. Park (2018). "Distinguishing friends, foes, and freeloaders in giant genomes." Current Opinion in Genetics and Development **49**: 49-55.

Benz, B. F. (2001). "Archaeological evidence of teosinte domestication from Guilá Naquitz, Oaxaca." Proceedings of the National Academy of Sciences of the United States of America **98**: 2104-2106.

Berg, J. J. and G. Coop (2014). "A Population Genetic Signal of Polygenic Adaptation." PLoS Genetics **10**.

Berg, J. J., A. Harpak, et al. (2019). "Reduced signal for polygenic adaptation of height in UK Biobank." eLife **8**: 1-47.

Bilinski, P., P. S. Albert, et al. (2018). "Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays." PLoS Genetics **14**.

Bradburd, G. S., P. L. Ralph, et al. (2013). "Disentangling the effects of geographic and ecological isolation on genetic differentiation." Evolution **67**: 3258-3273.

Bragg, J. G., M. A. Supple, et al. (2015). "Genomic variation across landscapes: insights and applications." The New Phytologist **207**(4): 953-967.

Bresson, C. C., Y. Vitasse, et al. (2011). "To what extent is altitudinal variation of functional traits driven by genetic adaptation in European oak and beech?" Tree Physiology **31**: 1164-1174.

Bulmer, M. G. G. (1972). "Multiple niche polymorphism." The American Naturalist **106**: 254-257.

Butelli, E., C. Licciardello, et al. (2012). "Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges." The Plant Cell **24**: 1242-1255.

Byars, S. G., W. Papst, et al. (2007). "Local adaptation and cogradient selection in the alpine plant, *Poa hiemata*, along a narrow altitudinal gradient." Evolution **61**: 2925-2941.

Casacuberta, E. and J. González (2013). "The impact of transposable elements in environmental adaptation." Molecular Ecology **22**: 1503-1517.

Cavrak, V. V., N. Lettner, et al. (2014). "How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation." PLoS Genetics **10**.

Coop, G., D. Witonsky, et al. (2010). "Using environmental correlations to identify loci underlying local adaptation." Genetics **185**: 1411-1423.

Da Fonseca, R. R., B. D. Smith, et al. (2015). "The origin and evolution of maize in the Southwestern United States." Nature Plants **1**(1): 14003.

De Jesús Sánchez González, J., J. A. R. Corral, et al. (2018). "Ecogeography of teosinte." PLoS ONE.

De Mita, S., A. C. Thuillet, et al. (2013). "Detecting selection along environmental gradients: Analysis of eight methods and their effectiveness for outbreeding and selfing populations." Molecular Ecology **22**: 1383-1399.

Devos, K. M., J. K. M. Brown, et al. (2010). "Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis." Genome Reseach: 1075-1079.

Diez, C. M., B. S. Gaut, et al. (2013). "Genome size variation in wild and cultivated maize along altitudinal gradients." New Phytologist **199**: 264-276.

Dobzhansky, T. (1964). "Biology, molecular and organismic." American Zoologist: 443-452.

Doebley, J. F. and H. H. Iltis (1980). "Taxonomy of Zea (Gramineae). I. A Subgeneric Classification with Key to Taxa." American Journal of Botany **67**: 982.

Dooner, H. K., Q. Wang, et al. (2019). "Spontaneous mutations in maize pollen are frequent in some lines and arise mainly from retrotranspositions and deletions." Proceedings of the National Academy of Sciences of the United States of America **166**: 10734-10743.

Eckert, A. J., P. E. Maloney, et al. (2015). "Local adaptation at fine spatial scales: an example from sugar pine (Pinus lambertiana, Pinaceae)." Tree Genetics and Genomes **11**.

Edelaar, P., P. Burraco, et al. (2011). "Comparisons between Q(ST) and F(ST) --how wrong have we been?" Mol Ecol **20**: 4830-4839.

Estill, J. C. and J. L. Bennetzen (2009). "The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes transposable elements in plant genomes." Plant Methods **5**: 1-11.

Evans, M. M. S. and J. L. Kermicle (2001). "Teosinte crossing barrier1, a locus governing hybridization of teosinte with maize." Theoretical and Applied Genetics **103**: 259-265.

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

Excoffier, L., T. Hofer, et al. (2009). "Detecting loci under selection in a hierarchically structured population." Heredity **103**: 285-298.

Fang, Z., T. Pyhäjärvi, et al. (2012). "Megabase-scale inversion polymorphism in the wild ancestor of maize." Genetics **191**: 883-894.

Faria, R., K. Johannesson, et al. (2019). "Evolving Inversions." Trends in Ecology and Evolution **34**: 239-248.

Feschotte, C. (2008). "The contribution of transposable elements ot the evolution of regulatory networks." Nature Reviews Genetics **9**: 397-405.

Feschotte, C., N. Jiang, et al. (2002). "Plant Transposable Elements: Where Genetics Meets Genomics." Nature Reviews Genetics **3**: 329-341.

Fischer, M. C., C. Rellstab, et al. (2013). "Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis halleri from the Alps." Molecular Ecology **22**: 5594-5607.

Foll, M. and O. Gaggiotti (2008). "A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective." Genetics **180**: 977-993.

Foll, M., O. E. Gaggiotti, et al. (2014). "Widespread signals of convergent adaptation to high altitude in Asia and America." The American Journal of Human Genetics **95**(4): 394-407.

Fournier-Level, A., A. Korte, et al. (2011). "A map of local adaptation in Arabidopsis thaliana." Science **334**: 86-89.

Fukunaga, K., J. Hill, et al. (2005). "Genetic diversity and population structure of teosinte." Genetics **169**: 2241-2254.

Fustier, M. A., J. T. Brandenburg, et al. (2017). "Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples." Molecular Ecology **26**: 2738-2756.

Gilbert, C. and C. Feschotte (2018). "Horizontal acquisition of transposable elements and viral sequences: patterns and consequences." Current Opinion in Genetics and Development **49**: 15-24.

Gilbert, K. J. and M. C. Whitlock (2015). "QST-FST comparisons with unbalanced half-sib designs." Molecular Ecology Resources **15**: 262-267.

Goerner-Potvin, P. and G. Bourque (2018). "Computational tools to unmask transposable elements." Nature Reviews Genetics **19**: 688-704.

Gonzalo-Turpin, H. and L. Hazard (2009). "Local adaptation occurs along altitudinal gradient despite the existence of gene flow in the alpine plant species Festuca eskia." Journal of Ecology **97**(4): 742-751.

Günther, T. and G. Coop (2016). "A Short Manual for Bayenv2.0."

Halbritter, A. H., S. Fior, et al. (2018). "Trait differentiation and adaptation of plants along elevation gradients." Journal of Evolutionary Biology **31**(6): 784-800.

Haldane, J. B. S. (1930). "Theoretical genetics of autopolyploids."

Hallatschek, O., P. Hersen, et al. (2007). "Genetic drift at expanding frontiers promotes gene segregation." Proceedings of the National Academy of Sciences **104**(50): 19926-19930.

Hämälä, T., T. M. Mattila, et al. (2018). "Local adaptation and ecological differentiation under selection, migration, and drift in Arabidopsis lyrata*." Evolution **72**: 1373-1386.

Hermisson, J. and P. S. Pennings (2017). "Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation." Methods in Ecology and Evolution **8**: 700-716.

Hoban, S., J. L. Kelley, et al. (2016). "Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions." The American Naturalist **188**: 379-397.

Hollister, J. D. and B. S. Gaut (2009). "Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression." Genome Research **19**: 1419-1428.

Holsinger, K. E. and B. S. Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting F(ST)." Nature reviews. Genetics **10**: 639-650.

Hori, Y., R. Fujimoto, et al. (2007). "A novel wx mutation caused by insertion of a retrotransposon-like sequence in a glutinous cultivar of rice (Oryza sativa)." Theoretical and Applied Genetics **115**(2): 217-224.

Hufford, M. B., P. Bilinski, et al. (2012). "Teosinte as a model system for population and ecological genomics." Trends in Genetics **28**: 606-615.

Hufford, M. B., P. Lubinksy, et al. (2013). "The Genomic Signature of Crop-Wild Introgression in Maize." PLoS Genetics **9**.

Iltis, H. H. and J. F. Doebley (1980). "Taxonomy of Zea (Gramineae). II. Subspecific Categories in the Zea Mays Complex and a Generic Synopsis." American Journal of Botany **67**: 994.

Jeong, C. and A. Di Rienzo (2014). Adaptations to local environments in modern human populations. Current Opinion in Genetics and Development. **29:** 1-8.

Jiang, N., Z. Bao, et al. (2003). "An active DNA transposon family in rice." Nature **421**: 163-167.

Jiang, N. and S. R. Wessler (2001). "Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements." **13**: 2553-2564.

Jiao, Y., P. Peluso, et al. (2017). "Improved maize reference genome with." Nature.

Johanson, U., J. West, et al. (2000). "Molecular Analysis of *FRIGIDA*, a Major Determinant of Natural Variation in *Arabidopsis* Flowering Time." Science **290**(5490): 344-347.

Josephs, E. B., J. J. Berg, et al. (2019). "Detecting adaptive differentiation in structured populations with genomic data and common gardens." Genetics **211**: 989-1004.

Kapitonov, V. V. and J. Jurka (2001). "Rolling-circle transposons in eukaryotes." Proceedings of the National Academy of Sciences **98**: 8714-8719.

Karhunen, M., J. Merilä, et al. (2013). "driftsel: An R package for detecting signals of natural selection in quantitative traits." Molecular Ecology Resources **13**: 746-754.

Kawakami, T., T. J. Morgan, et al. (2011). "Natural selection drives clinal life history patterns in the perennial sunflower species, *Helianthus maximiliani*." Molecular Ecology **20**: 2318-2328.

Kawase, M., K. Fukunaga, et al. (2005). "Diverse origins of waxy foxtail millet crops in East and Southeast Asia mediated by multiple transposable element insertions." Molecular Genetics and Genomics **274**(2): 131-140.

Kawecki, T. J. (1997). "Sympatric Speciation via Habitat Specialization Driven by Deleterious Mutations." Evolution **51**: 1751.

Kawecki, T. J. and D. Ebert (2004). "Conceptual issues in local adaptation." Ecology Letters **7**: 1225-1241.

Kejnovsky, E., J. Hawkins, et al. (2012). "Plant genome diversity." Springer Vienna: Vienna doi **10**: 978-973.

Kim, N. H., G. Leea, et al. (2016). "Real-time transposable element activity in individual live cells." Proceedings of the National Academy of Sciences of the United States of America **113**: 7278-7283.

Kirkpatrick, M. (2010). "How and why chromosome inversions evolve." PLoS Biology **8**.

Kirkpatrick, M. and N. Barton (2006). "Chromosome inversions, local adaptation and speciation." Genetics.

Kistler, L., S. Y. Maezumi, et al. (2018). "Multiproxy evidence highlights a complex evolutionary legacy of maize in South America." Science **362**: 1309-1313.

Kobayashi, S., N. Goto-Yamamoto, et al. (2004). "Retrotransposon-Induced Mutations in Grape Skin Color." Science **304**: 982.

Körner, C. (2007). "The use of 'altitude' in ecological research." Trends in Ecology and Evolution **22**: 569-574.

Lauter, N., C. Gustus, et al. (2004). "The inheritance and evolution of leaf pigmentation and pubescence in teosinte." Genetics Society of America **167**: 1949-1959.

Le Rouzic, A., T. S. Boutin, et al. (2007). "Long-term evolution of transposable elements." Proceedings of the National Academy of Sciences **104**: 19375-19380.

Legrand, D., N. Larranaga, et al. (2016). "Evolution of a butterfly dispersal syndrome."

Leimu, R. and M. Fischer (2008). "A meta-analysis of local adaptation in plants." PLoS ONE **3**.

Leinonen, T., R. J. S. McCairns, et al. (2013). "Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity." Nature reviews. Genetics **14**: 179-190.

Lenormand, T. (2002). "Gene flow and the limits to natural selection." Trends in Ecology and Evolution **17**: 183-189.

Lewontin, R. C. and J. Krakauer (1973). "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms." Genetics **74**: 175-195.

Li, Y. F., J. C. Costello, et al. (2008). ""Reverse ecology" and the power of population genomics." Evolution **62**: 2984-2994.

Lisch, D. and R. K. Slotkin (2011). Strategies for Silencing and Escape. The Ancient Struggle Between Transposable Elements and Their Hosts. International Review of Cell and Molecular Biology, Elsevier Inc. **292:** 119-152.

Lockton, S. and B. S. Gaut (2010). "The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata." BMC evolutionary biology **10**: 10.

Loreto, E. L. S., C. M. A. Carareto, et al. (2008). "Revisiting horizontal transfer of transposable elements in Drosophila." Heredity **100**: 545-554.

Lovell, J. T., T. E. Juenger, et al. (2013). "Pleiotropy of *FRIGIDA* enhances the potential for multivariate adaptation." Proceedings of the Royal Society B: Biological Sciences **280**.

Lowry, D. B. (2012). "Local adaptation in The model plant." New Phytologist **194**: 888-890.

Lu, Y., S. A. Hokin, et al. (2019). "A pistil-expressed pectin methylesterase confers cross-incompatibility between strains of Zea mays." Nature Communications **10**: 2304.

Makarevitch, I., A. J. Waters, et al. (2015). "Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress." PLoS Genetics **11**: 1-15.

Maron, L. G., C. T. Guimarães, et al. (2013). "Aluminum tolerance in maize is associated with higher *MATE1* gene copy number." Proceedings of the National Academy of Sciences **110**(13): 5241-5246.

Matsuoka, Y., Y. Vigouroux, et al. (2002). "A single domestication for maize shown by multilocus microsatellite genotyping." Proceedings of the National Academy of Sciences of the United States of America **99**: 6080-6084.

Maumus, F. and H. Quesneville (2016). "Impact and insights from ancient repetitive elements in plant genomes." Current Opinion in Plant Biology **30**: 41-46.

Maynard Smith, J. and J. Haigh (2008). "The hitch-hiking effect of a favourable gene." Genetics Research **89**: 391-403.

McClintock, B. (1950). "The origin and behavior of mutable loci in maize."

McClintock, B. (1956). "Controlling elements and the gene." Cold Spring Harbor symposia on quantitative biology **21**: 197-216.

McVean, G. (2007). "The structure of linkage disequilibrium around a selective sweep." Genetics **175**: 1395-1406.

Mei, W., M. G. Stetter, et al. (2018). "Adaptation in plant genomes: Bigger is different." American Journal of Botany **105**(1): 16-19.

Mendez-Vigo, B., F. X. Pico, et al. (2011). "Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in Arabidopsis." Plant Physiology **157**: 1942-1955.

Merchant, S. S., S. E. Prochnik, et al. (2007). "The Chlamydomonas genome reveals the evolution of key animal and plant functions." Science **318**(5848): 245-250.

Messer, P. W. (2013). "SLiM: Simulating evolution with selection and linkage." Genetics **194**: 1037-1039.

Mitchell-Olds, T., J. H. Willis, et al. (2007). "Which evolutionary processes influence natural genetic variation for phenotypic traits?" Nature Reviews Genetics **8**: 845-856.

Moeller, D. A. and P. Tiffin (2008). "Geographic variation in adaptation at the molecular level: A case study of plant immunity genes." Evolution **62**: 3069-3081.

Mojica, J. P., Y. W. Lee, et al. (2012). "Spatially and temporally varying selection on intrapopulation quantitative trait loci for a life history trade-off in Mimulus guttatus." Molecular Ecology **21**: 3718-3728.

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

Moyers, B. T. and L. H. Rieseberg (2016). "Remarkable life history polymorphism may be evolving under divergent selection in the silverleaf sunflower." Molecular Ecology **25**: 3817-3830.

Nielsen, R. (2005). "Molecular signatures of natural selection." Annual Review of Genetics **39**: 197-218.

Oleksyn, J., J. Modrzynski, et al. (1998). "Growth and physiology of Picea abies populations from elevational transects: common garden evidence for altitudinal ecotypes and cold adaptation." 573-590.

Oliver, K. R., J. A. McComb, et al. (2013). "Transposable elements: Powerful contributors to angiosperm evolution and diversity." Genome Biology and Evolution **5**: 1886-1901.

Ovaskainen, O., M. Karhunen, et al. (2011). "A new method to uncover signatures of divergent and stabilizing selection in quantitative traits." Genetics **189**: 621-632.

Panaud, O. (2016). "Horizontal transfers of transposable elements in eukaryotes: The flying genes." Comptes Rendus - Biologies **339**: 296-299.

Paterson, A. H., J. E. Bowers, et al. (2009). "The Sorghum bicolor genome and the diversification of grasses." Nature **457**: 551-556.

Pereira, V. (2004). "Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome." Genome Biology **5**(10): R79.

Petousi, N. and P. A. Robbins (2013). "Human adaptation to the hypoxia of high altitude: the Tibetan paradigm from the pregenomic to the postgenomic era." Journal of applied physiology **116**(7): 875-884.

Petrov, D. A., Y. T. Aminetzach, et al. (2003). "Size matters: Non-LTR retrotransposable elements and ectopic recombination in drosophila." Molecular Biology and Evolution **20**: 880-892.

Piperno, D. R. and K. V. Flannery (2001). "The earliest archaeological maize (zea mays l.) from highland mexico: new accelerator mass spectrometry dates and their implications. Proceedings of the National Academy of Sciences of the United States of America **98**: 2101-2013.

Piperno, D. R., A. J. Ranere, et al. (2009). "Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico." Proceedings of the National Academy of Sciences of the United States of America **106**: 5019-5024.

Pool, J. E. and R. Nielsen (2007). "Population size changes reshape genomic patterns of diversity." Evolution **61**: 3001-3006.

Pool, J. J. E., I. Hellmann, et al. (2010). "Population genetic inference from genomic sequence variation." Genome research **20**: 291-300.

Price, N., B. T. Moyers, et al. (2018). "Combining population genomics and fitness QTLs to identify the genetics of local adaptation in Arabidopsis thaliana." Proceedings of the National Academy of Sciences **115**: 5028-5033.

Pujol, B., S. Blanchet, et al. (2018). "The Missing Response to Selection in the Wild." Trends in Ecology and Evolution **33**: 337-346.

Pyhäjärvi, T., M. B. Hufford, et al. (2013). "Complex patterns of local adaptation in teosinte." Genome Biology and Evolution **5**: 1594-1609.

Quadrana, L., A. B. Silveira, et al. (2016). "The Arabidopsis thaliana mobilome and its impact at the species level." eLife **5**: 1-25.

Rellstab, C., F. Gugerli, et al. (2015). A practical guide to environmental association analysis in landscape genomics. Molecular Ecology. **24:** 4348-4370.

Rivera-Rodríguez, D. M., J. de Jesús Sánchez González, et al. (2019). " Morphological and Climatic Variability of Teosinte (Zea spp.) and Relationships Among Taxa " Systematic Botany **44**: 41-51.

Ronce, O. and J. Clobert (2012). "Dispersal syndromes." Dispersal ecology and evolution **155**: 119-138.

Ross-ibarra, J., P. L. Morrell, et al. (2007). "Ross-Ibarra, et al. PNAS May 15, 2007 vol. 104 suppl. 1 8641–8648.pdf." **104**: 8641-8648.

Ross-Ibarra, J., M. Tenaillon, et al. (2009). "Historical divergence and gene flow in the genus Zea." Genetics **181**: 1399-1413.

Sanchez, J. d. J., T. A. Kato Yamakake, et al. (1998). "Distribucion y caracterización del teocintle." 165.

Sanmiguel, P., B. S. Gaut, et al. (1998). The paleontology of intergene retrotransposons of maize.

Savolainen, O., M. Lascoux, et al. (2013). "Ecological genomics of local adaptation." Nature reviews. Genetics **14**: 807-820.

Schmutz, J., S. B. Cannon, et al. (2010). "Genome sequence of the palaeopolyploid soybean." Nature **463**(7278): 178.

Schnable, J. C., N. M. Springer, et al. (2011). "Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss." Proceedings of the National Academy of Sciences of the United States of America **108**: 4069-4074.

Schnable, P. S., D. Ware, et al. (2009). "The B73 Maize Genome: Complexity, Diversity, and Dynamics."

Schnee, F. B. and J. N. Thompson and Jr (1984). "Conditional Neutrality of Polygene Effects." **38**: 42-46.

Shi, J. and J. Lai (2015). "Patterns of genomic changes with crop domestication and breeding." Current Opinion in Plant Biology **24**: 47-53.

Sigman, M. J. and R. K. Slotkin (2016). "The First Rule of Plant Transposable Element Silencing: Location, Location, Location." The Plant Cell **28**: 304-313.

Simons, Y. B., K. Bullaughey, et al. (2018). "A population genetic interpretation of GWAS findings for human quantitative traits." PLoS Biology **16**: 1-20.

Sohail, M., R. M. Maier, et al. (2019). "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies." eLife **8**.

Spitze, K. (1993). "Population structure in Daphina obtusa: quantitative genetic and allozymic variation." Genetics Society of America **135**: 367-374.

Stitzer, M. C., S. N. Anderson, et al. (2019). "The Genomic Ecosystem of Transposable Elements in Maize." bioRxiv: 559922.

Stritt, C., S. P. Gordon, et al. (2017). "Recent Activity in Expanding Populations and Purifying Selection Have Shaped Transposable Element Landscapes." Genome Biology and EvolutionB **10**: 304-318.

Stuart, T., S. R. Eichten, et al. (2016). "Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation." eLife **5**.

Studer, A., Q. Zhao, et al. (2011). "Identification of a functional transposon insertion in the maize domestication gene tb1." Nature Genetics **43**: 1160-1163.

Swigoňová, Z., J. Lai, et al. (2004). "Close split of sorghum and maize genome progenitors." Genome Research **14**: 1916-1923.

Takuno, S., P. Ralph, et al. (2015). "Independent molecular basis of convergent highland adaptation in maize." Genetics **200**: 1297-1312.

Tenaillon, M. I., J. D. Hollister, et al. (2010). "A triptych of the evolution of plant transposable elements." Trends in Plant Science **15**: 471-478.

Tenaillon, M. I. and D. Manicacci (2011). Maize origins: an old question under the spotlights. Advances in Maize (Essential Reviews in Experimental Biology), The Society for Experimental Biology**:** 89-110.

Tenaillon, M. I., J. U'Ren, et al. (2004). "Selection versus demography: A multilocus investigation of the domestication process in maize." Molecular Biology and Evolution **21**: 1214-1225.

Tigano, A. and V. L. Friesen (2016). "Genomics of local adaptation with gene flow." Molecular Ecology **25**: 2144-2164.

Tishkoff, S. A., F. A. Reed, et al. (2007). "Convergent adaptation of human lactase persistance in Africa and Europe." Nat Genet **39**: 31-40.

Turchin, M. C., C. W. Chiang, et al. (2012). "Evidence of widespread selection on standing variation in Europe at height-associated SNPs." Nature Genetics **44**: 1015-1019.

Twyford, A. D. and J. Friedman (2015). "Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion." Evolution **69**: 1476-1486.

Vallebueno-Estrada, M., I. Rodríguez-Arévalo, et al. (2016). "The earliest maize from san marcos tehuacán is a partial domesticate with genomic evidence of inbreeding." Proceedings of the National Academy of Sciences of the United States of America **113**: 14151-14156.

van Heerwaarden, J., J. Doebley, et al. (2011). "Genetic signals of origin, spread, and introgression in a large sample of maize landraces." Proceedings of the National Academy of Sciences of the United States of America **108**: 1088-1092.

Varagona, M. J., M. Purugganan, et al. (1992). "Alternative splicing induced by insertion of retrotransposons into the maize waxy gene." The Plant Cell **4**(7): 811-820.

Vicient, C. M. (2010). "Transcriptional activity of transposable elements in maize." BMC genomics **11**: 601.

Vigouroux, Y., J. C. Glaubitz, et al. (2008). "Population structure and genetic diversity of New World maize races assessed by DNA microsatellites." American Journal of Botany **95**: 1240-1253.

Vitte, C., M. A. Fustier, et al. (2014). "The bright side of transposons in crop evolution." Briefings in Functional Genomics and Proteomics **13**: 276-295.

Vitte, C. and O. Panaud (2003). "Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L." Molecular Biology and Evolution **20**: 528-540.

Wang, L., A. Stec, et al. (1999). "The limits of selection during maize domestication." Nature **398**: 236-239.

Wang, M., X. Huang, et al. (2014). "Detecting Recent Positive Selection with High Accuracy and Reliability by Conditional Coalescent Tree." Molecular Biology and Evolution.

Weng, M. L., C. Becker, et al. (2019). "Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*." Genetics **211**: 703-714.

Wicker, T., F. Sabot, et al. (2007). "A unified classification system for eukaryotic transposable elements." Nature reviews. Genetics **8**: 973-982.

Wilkes, H. G. (1977). "Hybridization of maize and teosinte, in Mexico and Guatemala and the improvement of Maize 1." 254-293.

Wright, S. (1951). "The genetical structure of populations." Annals of Eugenics **15**: 323-354.

Wright, S. I., I. V. Bi, et al. (2005). "The effects of artificial selection on the maize genome." Science (New York, N.Y.) **308**: 1310-1314.

Xu, Y. and J. Du (2014). "Young but not relatively old retrotransposons are preferentially located in gene-rich euchromatic regions in tomato (*Solanum lycopersicum*) plants." The Plant Journal **80**(4): 582-591.

Yang, L. and J. L. Bennetzen (2009). "Distribution, diversity, evolution, and survival of Helitrons in the maize genome." Proceedings of the National Academy of Sciences.

Yang, Q., Z. Li, et al. (2013). "CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize." Proceedings of the National Academy of Sciences **110**: 16969-16974.

Yeaman, S. and M. C. Whitlock (2011). "The genetic architecture of adaptation under migration-selection balance." Evolution **65**: 1897-1911.

Yi, X., Y. Liang, et al. (2010). "Sequencing of fifty human exomes reveals adaptations to high altitude." Science **329**: 75-78.

Zemach, A., M. Y. Kim, et al. (2013). "The arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin." Cell **153**: 193-205.

Zhang, L., J. Hu, et al. (2019). "A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour." Nature Communications **10**: 1-13.

II. CHAPTER 1: COMMON GARDENS IN TEOSINTES REVEAL THE ESTABLISHMENT OF A SYNDROME OF ADAPTATION TO ALTITUDE

The present chapter aims to study teosinte local adaptation along altitudinal gradients with emphasis on unveiling its genetic determinants through a reverse ecology approach. First, population genomics methods on a small set of populations enabled to highlight a set of candidate SNPs. On a much wider population sample, we analyzed their frequencies in correlation with environmental descriptors, and determined their link to phenotypic variation through an association mapping method using common garden experiments.

This ambitious project was initiated during the PhD of Margaux-Alison Fustier under the supervision of Maud Tenaillon and Domenica Manicacci. During her PhD, M-A Fustier ran the common garden experiments with collaborators in Mexico. This involved planting, growing and measuring the plants. She also analyzed High-Throughput Sequencing data on six teosinte populations and determined regions with signals of natural selection (Fustier, Brandenburg et al. 2017). Among these, she identified a list of candidate SNPs, which were then genotyped on a larger set of populations as well as on the genetic association panel. Neutral SSR markers were also genotyped on the genetic association panel. She ran a first series of analyses on these data sets including the description of neutral genetic structure, of phenotypic variations and pairwise correlations among traits. She also proposed a preliminary association mapping model.

During my PhD I made adjustments to M-A. Fustier's analyses and developed new analyses. I improved the genetic association model including modifications in input matrices, and testing different models. To correct for neutral structure I reran STRUCTURE analyses on SSR data following different sub-sampling schemes and compared their performance. I also built kinship matrices from all SSRs as well as excluding one chromosome at a time and compared their results on the association mapping outputs. Since teosintes are an outcrossing taxa, I devised a way to calculate their kinship matrix diagonal to benefit from information included in their heterozygocity when applying software inspired for homozygous lines. I then generated association mapping results of candidate SNPs for five group and 11 population neutral structure corrections. Furthermore, I tested several more complex models including interactions between genotypes and population of origin.

I also re-estimated phenotypic values correcting for the experimental design, reran the phenotypic PCA and re-analyzed the correlations between traits. I evaluated the altitudinal effect on phenotypes by running a model that includes altitude as a covariate. I evaluated linkage disequilibrium among candidate SNPs, whilst correcting for neutral structure through Bayesian clustering of individuals as well as a kinship matrix calculated from SSR data. I undertook all the

Q_{ST}-F_{ST} analyses with the Q_{ST}-F_{ST}*comp* package using phenotypic data, the half-sib design pedigree and SSR data. I also performed all environmental correlation analyses of candidate SNPs along both gradients, which entailed recalculating environmental PCA coordinates on a subset of 28 populations.

Finally, I analyzed and discussed all these results and greatly contributed to the writing of the corresponding paper which I present here as Chapter 1. This paper has currently been accepted for publication at PLOS Genetics. It is co-first authored by M-A. Fustier and myself.

Common gardens in teosintes reveal the establishment of a syndrome of adaptation to altitude

M-A. Fustier[1,†], N.E. Martínez-Ainsworth[1,†], J.A. Aguirre-Liguori[2], A. Venon[1], H. Corti[1], A. Rousselet[1], F. Dumas[1], H. Dittberner[3], M.G. Camarena[4], D. Grimanelli[5], O. Ovaskainen[6,7], M. Falque[1], L. Moreau[1], J. de Meaux[3], S. Montes[4], L.E. Eguiarte[2], Y. Vigouroux[5], D. Manicacci[1,*], M.I. Tenaillon[1,*]

[1]: Génétique Quantitative et Evolution – Le Moulon, Institut National de la Recherche Agronomique, Université Paris-Sud, Centre National de la Recherche Scientifique, AgroParisTech, Université Paris-Saclay, France

[2]: Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

[3]: Institute of Botany, University of Cologne Biocenter, 47b Cologne, Germany

[4]: Campo Experimental Bajío, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Celaya, Mexico

[5]: Université de Montpellier, Institut de Recherche pour le développement, UMR Diversité, Adaptation et Développement des plantes, Montpellier, France

[6]: Organismal and Evolutionary Biology Research Programme, PO Box, 65, FI-00014 University of Helsinki, Finland

[7]: Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

[†]: The authors equally contributed to this work* co-corresponding authors

E-mail: domenica.manicacci@inra.fr, maud.tenaillon@inra.fr

Abstract

In plants, local adaptation across species range is frequent. Yet, much has to be discovered on its environmental drivers, the underlying functional traits and their molecular determinants. Genome scans are popular to uncover outlier loci potentially involved in the genetic architecture of local adaptation, however links between outliers and phenotypic variation are rarely addressed. Here we focused on adaptation of teosinte populations along two elevation gradients in Mexico that display continuous environmental changes at a short geographical scale. We used two common gardens, and phenotyped 18 traits in 1664 plants from 11 populations of annual teosintes. In parallel, we genotyped these plants for 38 microsatellite markers as well as for 171 outlier single nucleotide polymorphisms (SNPs) that displayed excess of allele differentiation between pairs of lowland and highland populations and/or correlation with environmental variables. Our results revealed that phenotypic differentiation at 10 out of the 18 traits was driven by local selection. Trait covariation along the elevation gradient indicated that adaptation to altitude results from the assembly of multiple co-adapted traits into a complex syndrome: as elevation increases, plants flower earlier, produce less tillers, display lower stomata density and carry larger, longer and heavier grains. The proportion of outlier SNPs associating with phenotypic variation, however, largely depended on whether we considered a neutral structure with 5 genetic groups (73.7%) or 11 populations (13.5%), indicating that population stratification greatly affected our results. Finally, chromosomal inversions were enriched for both SNPs whose allele frequencies shifted along elevation as well as phenotypically-associated SNPs. Altogether, our results are consistent with the establishment of an altitudinal syndrome promoted by local selective forces in teosinte populations in spite of detectable gene flow. Because elevation mimics climate change through space, SNPs that we found underlying phenotypic variation at adaptive traits may be relevant for future maize breeding.

Keywords: spatially-varying selection; F_{ST}-scan; association mapping; altitudinal syndrome; pleiotropy; chromosomal inversions.

Author summary

Across their native range species encounter a diversity of habitats promoting local adaptation of geographically distributed populations. While local adaptation is widespread, much has yet to be discovered about the conditions of its emergence, the targeted traits, their molecular determinants and the underlying ecological drivers. Here we employed a reverse ecology approach, combining phenotypes and genotypes, to mine the determinants of local adaptation of teosinte populations distributed along two steep altitudinal gradients in Mexico. Evaluation of 11 populations in two common gardens located at mid-elevation pointed to adaptation via an altitudinal multivariate syndrome, in spite of gene flow. We scanned genomes to identify loci with allele frequencies shifts along elevation, a subset of which associated to trait variation. Because elevation mimics climate change through space, these polymorphisms may be relevant for future maize breeding.

II.1 INTRODUCTION

Local adaptation is key for the preservation of ecologically useful genetic variation (Whitlock, 2015). The conditions for its emergence and maintenance have been the focus of a long-standing debate nourished by ample theoretical work (Bulmer 1972; Lande 1976; Bradshaw 1984; Endler 1986; Lenormand 2002; Whitlock and Gomulkiewicz 2005; Gay, Crochet et al. 2008; Yeaman and Otto 2011). On the one hand, spatially-varying selection promotes the evolution of local adaptation, provided that there is genetic diversity underlying the variance of fitness-related traits (Rundle and Nosil 2005). On the other hand, opposing forces such as neutral genetic drift, temporal fluctuations of natural selection, recurrent introduction of maladaptive alleles via migration and homogenizing gene flow may hamper local adaptation (reviewed in (Kawecki and Ebert 2004)). Meta-analyzes indicate that local adaptation is pervasive in plants, with evidence of native-site fitness advantage in reciprocal transplants detected in 45% to 71% of populations (Leimu and Fischer 2008; Hereford 2009).

While local adaptation is widespread, much has yet to be discovered about the traits affected by spatially-varying selection, their molecular determinants and the underlying ecological drivers (Tiffin and Ross-Ibarra 2014). Local adaptation is predicted to increase with phenotypic, genotypic and environmental divergence among populations (Lande 1976; Slatkin 1985; Garcia-Ramos and Kirkpatrick 1997). Comparisons of the quantitative genetic divergence of a trait (Q_{ST}) with the neutral genetic differentiation (F_{ST}) can provide hints on whether trait divergence is driven by spatially-divergent selection (Wright 1951; Lande 1992; Spitze 1993; Whitlock 1999). Striking examples of divergent selection include developmental rate in the common toad (Luquet, Léna et al. 2015), drought and frost tolerance in alpine populations of the European silver fir (Roschanski, Csilléry et al. 2016), and traits related to plant phenology, size and floral display among populations of *Helianthus* species (Kawakami, Morgan et al. 2011; Moyers and Rieseberg 2016). These studies have reported covariation of physiological, morphological and/or life-history traits across environmental gradients which collectively define adaptive syndromes. Such syndromes may result from several non-exclusive mechanisms: plastic responses, pleiotropy, non-adaptive genetic correlations among traits (constraints), and joint selection of traits encoded by different sets of genes resulting in adaptive correlations. In some cases, the latter mechanism may involve selection and rapid spread of chromosomal inversions that happen to capture multiple locally favored alleles (Kirkpatrick and Barton 2006) as exemplified in *Mimulus guttatus* (Lowry and Willis 2010). While

distinction between these mechanisms is key to decipher the evolvability of traits, empirical data on the genetic bases of correlated traits are currently lacking (Legrand, Larranaga et al. 2016).

The genes mediating local adaptation are usually revealed by genomic regions harboring population-specific signatures of selection. These signatures include alleles displaying greater-than-expected differentiation among populations (Bierne, Welch et al. 2011) and can be identified through F_{ST}-scans (Lewontin and Krakauer 1973; Beaumont and Nichols 1996; Vitalis, Dawson et al. 2001; Foll and Gaggiotti 2008; Excoffier 2009; Bonhomme, Chevalet et al. 2010; Günther and Coop 2013). However, F_{ST}-scans and its derivative methods (Bierne, Welch et al. 2011) suffer from a number of limitations, among them a high number of false positives (reviewed in (Lotterhos and Whitlock 2014; Haasl and Payseur 2016)) and the lack of power to detect true positives (Le Corre and Kremer 2012). Despite these caveats, F_{ST}-outlier approaches have helped in the discovery of emblematic adaptive alleles such as those segregating at the *EPAS1 locus* in Tibetan human populations adapted to high altitude (Yi, Liang et al. 2010). An alternative to detect locally adaptive loci is to test for genotype-environment correlations (Joost, Bonin et al. 2007; Coop, Witonsky et al. 2010; Poncet, Herrmann et al. 2010; Guillot, Renaud et al. 2012; Frichot, Schoville et al. 2013; Günther and Coop 2013; Gautier 2015). Correlation-based methods can be more powerful than differentiation-based methods (De Mita, Thuillet et al. 2013), but spatial autocorrelation of population structure and environmental variables can lead to spurious signatures of selection (Hoban, Kelley et al. 2016).

Ultimately, to identify the outlier loci that have truly contributed to improve local fitness, a link between outliers and phenotypic variation needs to be established. The most common approach is to undertake association mapping. However, recent literature in humans has questioned our ability to control for sample stratification in such approach (Barton, Hermisson et al. 2019). Detecting polymorphisms responsible for trait variation is particularly challenging when trait variation and demographic history follow parallel environmental (geographic) clines. Plants however benefit from the possibility of conducting replicated phenotypic measurements in common gardens, where environmental variation is controlled. Hence association mapping has been successfully employed in the model plant species *Arabidopsis thaliana*, where broadly distributed ecotypes evaluated in replicated common gardens have shown that fitness-associated alleles display geographic and climatic patterns indicative of selection (Fournier-Level, Korte et al. 2011). Furthermore, the relative fitness of *A. thaliana* ecotypes in a given environment could be predicted from climate-associated SNPs (Hancock, Brachi et al. 2011). While climatic selection over broad latitudinal scales produces genomic and phenotypic patterns of local adaptation in the selfer plant

A. thaliana, whether similar patterns exist at shorter spatial scale in outcrossing species remains to be elucidated.

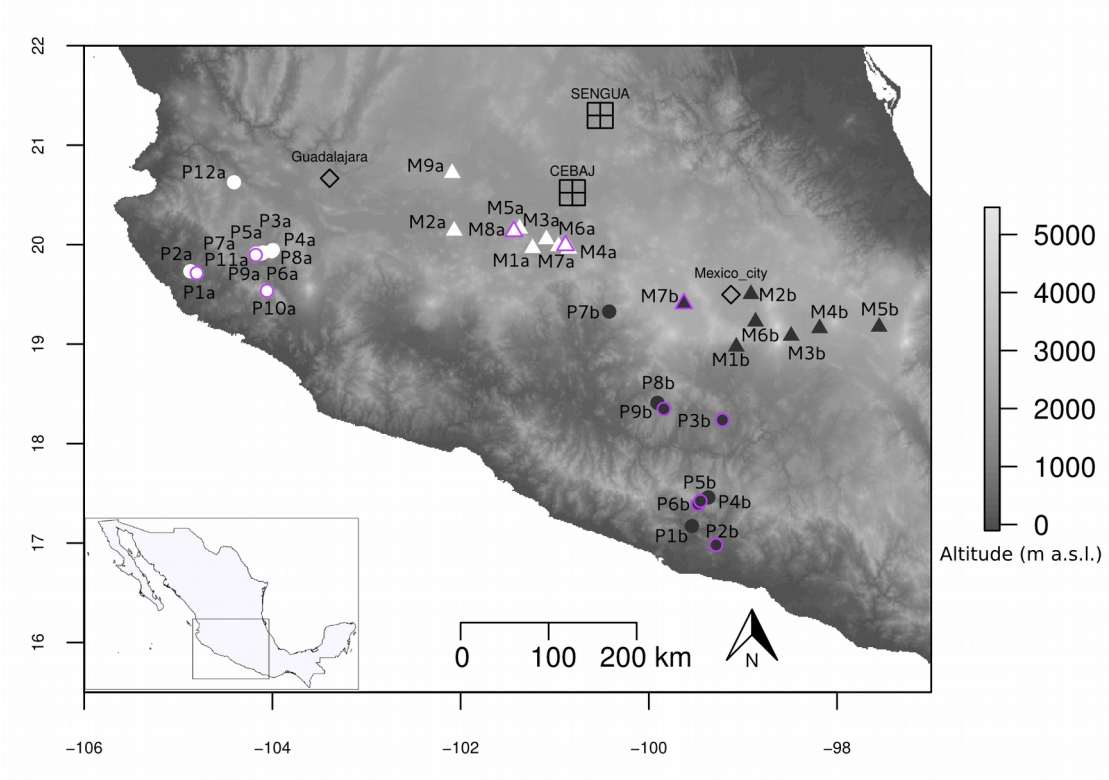
We focused here on a well-established outcrossing plant system, the teosintes, to investigate the relationship of molecular, environmental, and phenotypic variation in populations sampled across two elevation gradients in Mexico. The gradients covered a relatively short yet climatically diverse, spatial scale. They encompassed populations of two teosinte subspecies that are the closest wild relatives of maize, *Zea mays* ssp. *parviglumis* (hereafter *parviglumis*) and *Z. mays* ssp. *mexicana* (hereafter *mexicana*). The two subspecies display large effective population sizes (Ross-Ibarra, Tenaillon et al. 2009), and span a diversity of climatic conditions, from warm and mesic conditions below 1800 m for *parviglumis*, to drier and cooler conditions up to 3000 m for *mexicana* (Hufford, Martínez-Meyer et al. 2012). Previous studies have discovered potential determinants of local adaptation in these systems. At a genome-wide scale, decrease in genome size correlates with increasing altitude, which likely results from the action of natural selection on life cycle duration (Diez, Gaut et al. 2013; Bilinski, Albert et al. 2018). More modest structural changes include megabase-scale inversions that harbor clusters of SNPs whose frequencies are associated with environmental variation (Fang, Pyhäjärvi et al. 2012; Pyhäjärvi, Hufford et al. 2013). Also, differentiation- and correlation-based genome scans in teosinte populations succeeded in finding outlier SNPs potentially involved in local adaptation (Aguirre-Liguori, Tenaillon et al. 2017; Fustier, Brandenburg et al. 2017). But a link with phenotypic variation has yet to be established.

In this paper, we genotyped a subset of these outlier SNPs on a broad sample of 28 teosinte populations, for which a set of neutral SNPs was also available; as well as on an association panel encompassing 11 populations. We set up common gardens in two locations to evaluate the association panel for 18 phenotypic traits over two consecutive years. Individuals from this association panel were also genotyped at 38 microsatellite markers to enable associating genotypic to phenotypic variation while controlling for sample structure and kinship among individuals. We addressed three main questions: What is the extent of phenotypic variation within and among populations? Can we define a set of locally-selected traits that constitute a syndrome of adaptation to altitude? What are the genetic bases of such syndrome? We further discuss the challenges of detecting phenotypically-associated SNPs when trait and genetic differentiation parallel environmental clines.

II.2 RESULTS

II.2.1 Trait-by-trait analysis of phenotypic variation within and among populations.

In order to investigate phenotypic variation, we set up two common garden experiments located in Mexico to evaluate individuals from 11 teosinte populations (Fig 1). The two experimental fields were chosen because they were located at intermediate altitudes (S1 Fig). Although natural teosinte populations are not typically encountered around these locations (Hufford, Martínez-Meyer et al. 2012), we verified that environmental conditions were compatible with both subspecies (S2 Fig). The 11 populations were sampled among 37 populations (S1 Table) distributed along two altitudinal gradients that range from 504 to 2176 m in altitude over ~460 kms for gradient *a*, and from 342 to 2581m in altitude over ~350 kms for gradient *b* (S1 Fig). Lowland populations of the subspecies *parviglumis* (n=8) and highland populations of the subspecies *mexicana* (n=3) were climatically contrasted as can be appreciated in the Principal Component Analysis (PCA) computed on 19 environmental variables (S2 Fig). The corresponding set of individuals grown from seeds sampled from the 11 populations formed the association panel.



(Caption on following page)

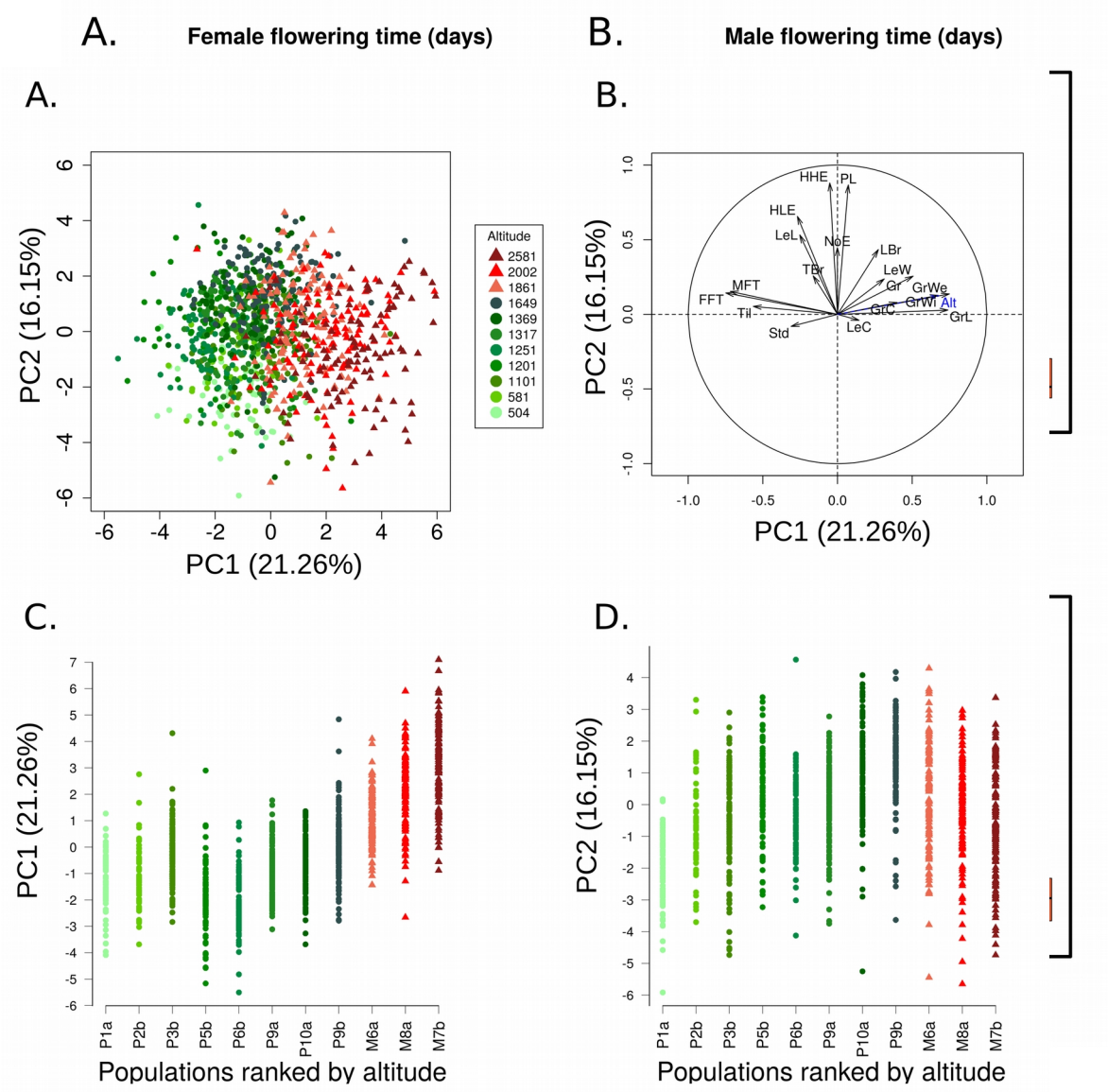
Figure 1. Geographical location of sampled populations and experimental fields. The entire set of 37 Mexican teosinte populations is shown with *parviglumis* (circles) and *mexicana* (triangles) populations sampled along gradient *a* (white) and gradient *b* (black). The 11 populations indicated with a purple outline constituted the association panel. This panel was evaluated in a four-block design over two years in two experimental fields located at mid-elevation, SENGUA and CEBAJ. Two major cities (Mexico City and Guadalajara) are also indicated. Topographic surfaces have been obtained from International Centre for Tropical Agriculture (Jarvis A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT), available from http://srtm.csi.cgiar.org).

We gathered phenotypic data during two consecutive years (2013 and 2014). We targeted 18 phenotypic traits that included six traits related to plant architecture, three traits related to leaves, three traits related to reproduction, five traits related to grains, and one trait related to stomata (S2 Table). Each of the four experimental assays (year-field combinations) encompassed four blocks. In each block, we evaluated one offspring (half-sibs) of ~15 mother plants from each of the 11 teosinte populations using a semi-randomized design. After filtering for missing data, the association panel included 1664 teosinte individuals. We found significant effects of Field, Year and/or their interaction for most traits, and a highly significant Population effect for all of them (model M1, S3 Table).

We investigated the influence of altitude on each trait independently. All traits, except for the number of nodes with ears (NoE), exhibited a significant effect of altitude (S3 Table, M3 model). Note that after accounting for elevation, the population effect remained significant for all traits, suggesting that factors other than altitude contributed to shape phenotypic variation among populations. Traits related to flowering time and tillering displayed a continuous decrease with elevation, and traits related to grain size increased with elevation (Fig 2 & S3 Fig). Stomata density also diminished with altitude (Fig 2). In contrast, plant height, height of the highest ear, number of nodes with ear in the main tiller displayed maximum values at intermediate altitudes (highland *parviglumis* and lowland *mexicana*) (S3 Fig).

We estimated narrow-sense heritabilites (additive genotypic effect) per population for all traits using a mixed animal model. Average per-trait heritability ranged from 0.150 for tassel branching to 0.664 for female flowering time, albeit with large standard errors (S2 Table). We obtained higher heritability for grain related traits when mother plant measurements were included in the model with 0.631 (sd = 0.246), 0.511 (sd = 0.043) and 0.274 (sd = 0.160) for grain length, weight and width, respectively, suggesting that heritability was under-estimated for other traits where mother plant values were not available.

Figure 2: Population-level box-plots of adjusted means for four traits. Traits are female flowering time (A), male flowering time (B), grain length (C) and stomata density (D). Populations are ranked by altitude. *Parviglumis* populations are shown in green and *mexicana* in red, lighter colors are used for gradient 'a' and darker colors for gradient 'b'. In the case of male and female flowering time, we report data for 9 out of 11 populations because most



individuals from the two lowland populations (P1a and P2b) did not flower in our common gardens. Covariation with elevation was significant for the four traits. Corrections for experimental setting are detailed in the material and methods section (Model M'1).

II.2.2 Multivariate analysis of phenotypic variation and correlation between traits.

Principal component analysis including all phenotypic measurements highlighted that 21.26% of the phenotypic variation scaled along PC1 (Fig 3A), a PC axis that is strongly collinear with altitude (Fig 3B). Although populations partly overlapped along PC1, we observed a consistent tendency for population phenotypic differentiation along altitude irrespective of the gradient (Fig

3C). Traits that correlated the most to PC1 were related to grain characteristics, tillering, flowering and to a lesser extent to stomata density (Fig 3B). PC2 correlated with traits exhibiting a trend toward increase-with-elevation within *parviglumis*, but decrease-with-elevation within *mexicana* (Fig 3D). Those traits were mainly related to vegetative growth (Fig 3B). Together, both axes explained 37.41% of the phenotypic variation.

Figure 3: Principal Component Analysis on phenotypic values corrected for the experimental setting. Individuals factor map (A) and corresponding correlation circle (B) on the first two principal components with altitude (Alt) added as a supplementary variable (in blue). Individual phenotypic values on PC1 (C) and PC2 (D) are plotted against population ranked by altitude and color-coded following A. For populations from the two subspecies, *parviglumis* (circles) and *mexicana* (triangles), color intensity indicates ascending elevation in green for *parviglumis* and red for *mexicana*. Corrections for experimental setting are detailed in the material and methods (Model M2).

We assessed more formally pairwise-correlations between traits after correcting for experimental design and population structure (K=5). We found 82 (54%) significant correlations among 153 tested pairs of traits. The following pairs of traits had the strongest positive correlations: male and female flowering time, plant height and height of the highest ear, height of the highest and lowest ear, grain length with grain weight and width (S4 Fig). The correlation between flowering time (female or male) with grain weight and length were among the strongest negative correlations (S4 Fig).

II.2.3 Neutral structuring of the association panel

We characterized the genetic structure of the association panel using SSRs. The highest likelihood from Bayesian classification was obtained at K=2 and K=5 clusters (S5 Fig). At K=2, the clustering separated the lowland of gradient *a* from the rest of the populations. From K=3 to K=5, a clear separation between the eight *parviglumis* and the three *mexicana* populations emerged. Increasing K values finally split the association panel into the 11 populations it encompassed (S6 Fig). The K=5 structure associated to both altitude (lowland *parviglumis* versus highland *mexicana*) and gradients *a* and *b* (Fig 4A & B). TreeMix analysis for a subset of 10 of these populations further confirmed those results with an early split separating the lowlands from gradient *a* (cf. K=2, S6 Fig) followed by the separation of the three *mexicana* from the remaining populations (Fig 4C). TreeMix further supported three migration edges, a model that explained 98.75% of the variance and represented a significant improvement over a model without admixture (95.7%, Figure S7).

This admixture model was consistent with gene flow between distant lowland *parviglumis* populations from gradient *a* and *b*, as well as between *parviglumis* and *mexicana* populations (Fig 4C). Likewise, structure analysis also suggested admixture among some of the lowland populations, and to a lesser extent between the two subspecies (Fig 4B).

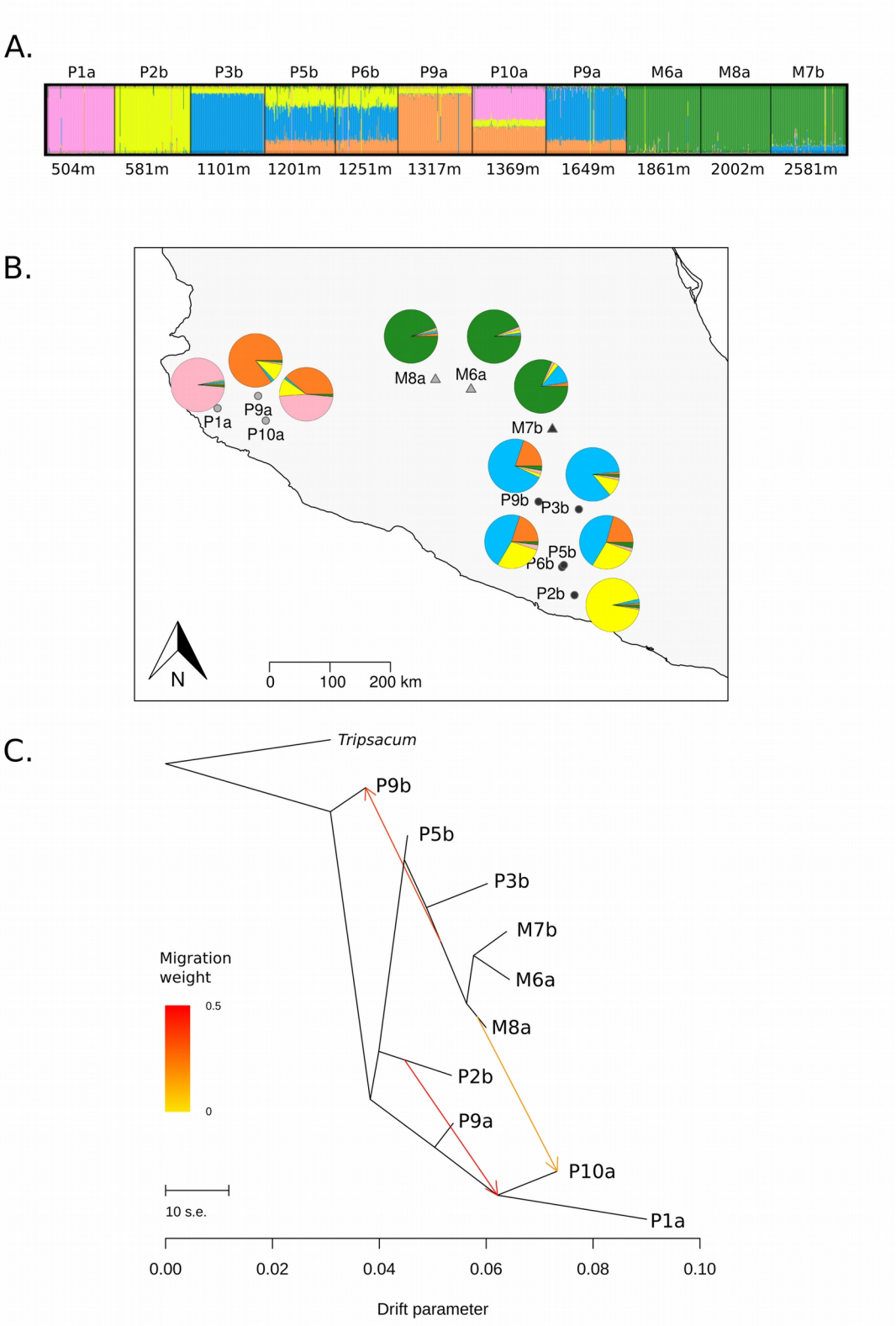


Figure 4: Genetic clustering, historical splits and admixture among populations of the association panel. Genetic clustering visualization based on 38 SSRs is shown for *K=5* (A). Colors represent the *K* clusters. Individuals (vertical lines) are partitioned into colored segments whose length represents the membership proportions to the *K* clusters. Populations (named after the subspecies M: *mexicana*, P: *parviglumis* and gradient 'a' or 'b') are ranked by altitude indicated in meters above sea level. The corresponding geographic distribution of populations along with their average membership probabilities are plotted (B). Historical splits and admixtures between populations were inferred from SNP data for a subset of 10 populations of the association panel (C). Admixtures are colored according to their weight.

II.2.4 Identification of traits evolving under spatially-varying selection

We estimated the posterior mean (and 95% credibility interval) of genetic differentiation (F_{ST}) among the 11 populations of the association panel using DRIFTSEL. Considering 1125 plants for which we had both individual phenotypes and individual genotypes for 38 SSRs (S4 Table), we estimated the mean F_{ST} to 0.22 (0.21-0.23). Note that we found a similar estimate on a subset of 10 of these populations with 1000 neutral SNPs (F_{ST} (CI)=0.26 (0.25-0.27)). To identify traits whose variation among populations was driven primarily by local selection, we employed the Bayesian method implemented in DRIFTSEL, that infers additive genetic values of traits from a model of population divergence under drift (Ovaskainen, Karhunen et al. 2011). Selection was inferred when observed phenotypic differentiation exceeded neutral expectations for phenotypic differentiation under random genetic drift. Single-trait analyses revealed evidence for spatially-varying selection at 12 traits, with high consistency between SSRs and neutral SNPs (Table 1). Another method that contrasted genetic and phenotypic differentiation (Q_{ST}- F_{ST}) uncovered a large overlap with nine out of the 12 traits significantly deviating from the neutral model (Table 1) and one of the remaining ones displaying borderline significance (Plant height=PL, S8 Fig). Together, these two methods indicated that phenotypic divergence among populations was driven by local selective forces.

Table 1. Signals of selection (posterior probability S) for each trait considering SSR markers (11 populations) or SNPs (10 populations).

Traits[a]	SSR[b]	SNP[b]
Plant height	0.995	0.972
Height of the lowest ear*	0.950	0.959
Height of the highest ear	0.982	0.966
Number of tillers*	1.000	1.000
Number of lateral branches*	1.000	0.990
Number of nodes with ears	0.682	0.699
Leaf length	0.888	0.875
Leaf width	0.999	0.996
Leaf color	0.633	0.583
Female flowering time*	1.000	1.000
Male flowering time*	1.000	1.000
Tassel branching*	0.925	0.908
Number of grains per ear	0.832	0.622
Grain length*	1.000	1.000
Grain width*	0.995	0.984
Grain weight*	1.000	0.999
Grain color	0.717	0.689
Stomata density*	0.999	0.999

[a]: Traits displaying signal of selection (spatially-varying traits, S > 0.95) are indicated in bold, and marked by an asterisk when significant in Q_{ST}-F_{ST}*Comp* analysis. We considered the underlined traits as spatially varying. For a detailed description of traits see S2 Table.

[b]: Values reported correspond to S from DRIFTSEL. S is the posterior probability that divergence among populations was not driven by drift only. Following (McKinney, Varian et al. 2014), we used here a conservative credibility value of S > 0.95 to declare divergent selection.

Altogether, evidence of spatially varying selection at 10 traits (Table 1) as well as continuous variation of a subset of traits across populations in both elevation gradients (Fig 2, S3 Fig) was consistent with a syndrome where populations produced less tillers, flowered earlier, displayed lower stomata density and carried larger, longer and heavier grains with increasing elevation.

II.2.5 Outlier detection and correlation with environmental variables

We successfully genotyped 218 (~81%) out of 270 outlier SNPs on a broad set of 28 populations, of which 141 were previously detected in candidate regions for local adaptation (Fustier, Brandenburg et al. 2017). Candidate regions were originally identified from re-sequencing data of only six teosinte populations (S1 Table) following an approach that included high differentiation between highlands and lowlands, environmental correlation, and in some cases their intersection with genomic regions involved in quantitative trait variation in maize. The remaining outlier SNPs (77) were discovered in the present study by performing F_{ST}-scans on the same re-sequencing data (S5 Table). We selected outlier SNPs that were both highly differentiated between highland and lowland populations within gradient (high/low in gradient *a* or *b* or both), and between highland and lowland populations within subspecies in gradient *b* (high/low within *parviglumis*, *mexicana* or both). F_{ST}-scans pinpointed three genomic regions of particularly high differentiation (S9 Fig) that corresponded to previously described inversions (Fang, Pyhäjärvi et al. 2012; Pyhäjärvi, Hufford et al. 2013): one inversion on chromosome 1 (*Inv1n*), one on chromosome 4 (*Inv4m*) and one on the far end of chromosome 9 (*Inv9e*).

A substantial proportion of outlier SNPs was chosen based on their significant correlation among six populations between variation of allele frequency and their coordinate on the first environmental principal component (Fustier, Brandenburg et al. 2017). We extended environmental analyses to all

171 outlier SNPs on a broader sample of 28 populations (S1 Table) and used the two first components (PCenv1 and PCenv2) to summarize environmental information. When considering all 37 populations, the first component, that explained 56% of the variation, correlated with altitude but displayed no correlation to either latitude or longitude. PCenv1 was defined both by temperature- and precipitation- related variables (S2 B Fig) including Minimum Temperature of Coldest Month (T6), Mean Temperature of Driest and Coldest Quarter (T9 and T11) and Precipitation of Driest Month and Quarter (P14 and P17). The second PC explained 20.5% of the variation and was mainly defined (S2 B Fig) by Isothermality (T3), Temperature Seasonality (T4) and Temperature Annual Range (T7).

We first employed multiple regression to test whether the pairwise F_{ST} matrix across 28 populations correlated to the environmental (distance along PCenv1) and/or the geographical distance. As expected, we found a significantly greater proportion of environmentally-correlated SNPs among outliers compared with neutral SNPs (χ^2 =264.07, P-value=2.2 10^{-16}), a pattern not seen with geographically-correlated SNPs. That outlier SNPs displayed a greater isolation-by-environment than isolation-by-distance, indicated that patterns of allele frequency differentiation among populations were primarily driven by adaptive processes. We further tested correlations between allele frequencies and environmental variation. Roughly 60.82% (104) of the 171 outlier SNPs associated with at least one of the two first PCenvs, with 87 and 33 associated with PCenv1 and PCenv2, respectively, and little overlap (S5 Table). As expected, the principal component driven by altitude (PCenv1) correlated to allele frequency for a greater fraction of SNPs than the second orthogonal component. Interestingly, we found enrichment of environmentally-associated SNPs within inversions both for PCenv1 (χ^2 = 14.63, P-value=1.30 10^{-4}) and PCenv2 (χ^2 = 33.77, P-value=6.22 10^{-9}).

II.2.6 Associating genotypic variation to phenotypic variation

We tested the association between phenotypes and 171 of the outlier SNPs (MAF>5%) using the association panel. For each SNP-trait combination, the sample size ranged from 264 to 1068, with a median of 1004 individuals (S6 Table). We used SSRs to correct for both structure (at K=5) and kinship among individual genotypes. This model (M5) resulted in a uniform distribution of P-values when testing the association between genotypic variation at SSRs and phenotypic trait variation (S10 Fig). Under this model, we found that 126 outlier SNPs (73.7%) associated to at least one trait (Fig 5 and S11 Fig) at an FDR of 10%. The number of associated SNPs per trait varied

from 0 for leaf and grain coloration, to 55 SNPs for grain length, with an average of 22.6 SNPs per trait (S5 Table). Ninety-three (73.8%) out of the 126 associated SNPs were common to at least two traits, and the remaining 33 SNPs were associated to a single trait (S5 Table). The ten traits displaying evidence of spatially varying selection in the Q_{ST}-F_{ST} analyses displayed more associated SNPs per trait (30.5 on average), than the non-spatially varying traits (12.75 on average).

A growing body of literature stresses that incomplete control of population stratification may lead to spurious associations (Sohail, Maier et al. 2019). Hence, highly differentiated traits along environmental gradients are expected to co-vary with any variant whose allele frequency is differentiated along the same gradients, without underlying causal link. We therefore expected false positives in our setting where both phenotypic traits and outlier SNPs varied with altitude. We found a slightly significant correlation (r=0.5, P-value=0.03) between the strength of the population effect for each trait – a measure of trait differentiation (S3 Table) – and its number of associated SNPs (S5 Table).

To verify that additional layers of structuring among populations did not cause an excess of associations, we repeated the association analyzes considering a structuring with 11 populations (instead of *K*=5) as covariate (M5'), a proxy of the structuring revealed at *K*=11 (S6 Fig). With this level of structuring, we retrieved much less associated SNPs (S5 Table). Among the 126 SNPs associating with at least one trait at *K*=5, only 22 were recovered considering 11 populations. An additional SNP was detected with structuring at 11 populations that was absent at *K*=5. Eight traits displayed no association, and the remaining traits varied from a single associated SNP (Leaf length – LeL and the number of tillers – Til) to 8 associated SNPs for grain weight (S5 Table). For instance, traits such as female or male flowering time that displayed 45 and 43 associated SNPs at *K*=5, now displayed only 4 and 3 associated SNPs, respectively (Fig 5). Note that one trait (Leaf color) associated with 4 SNPs considering 11 populations while displaying no association at *K*=5. Significant genetic associations were therefore highly contingent on the population structure. Noteworthy, traits under spatially varying selection still associated with more SNPs (2.00 on average) than those with no spatially varying selection (1.25 SNPs on average).

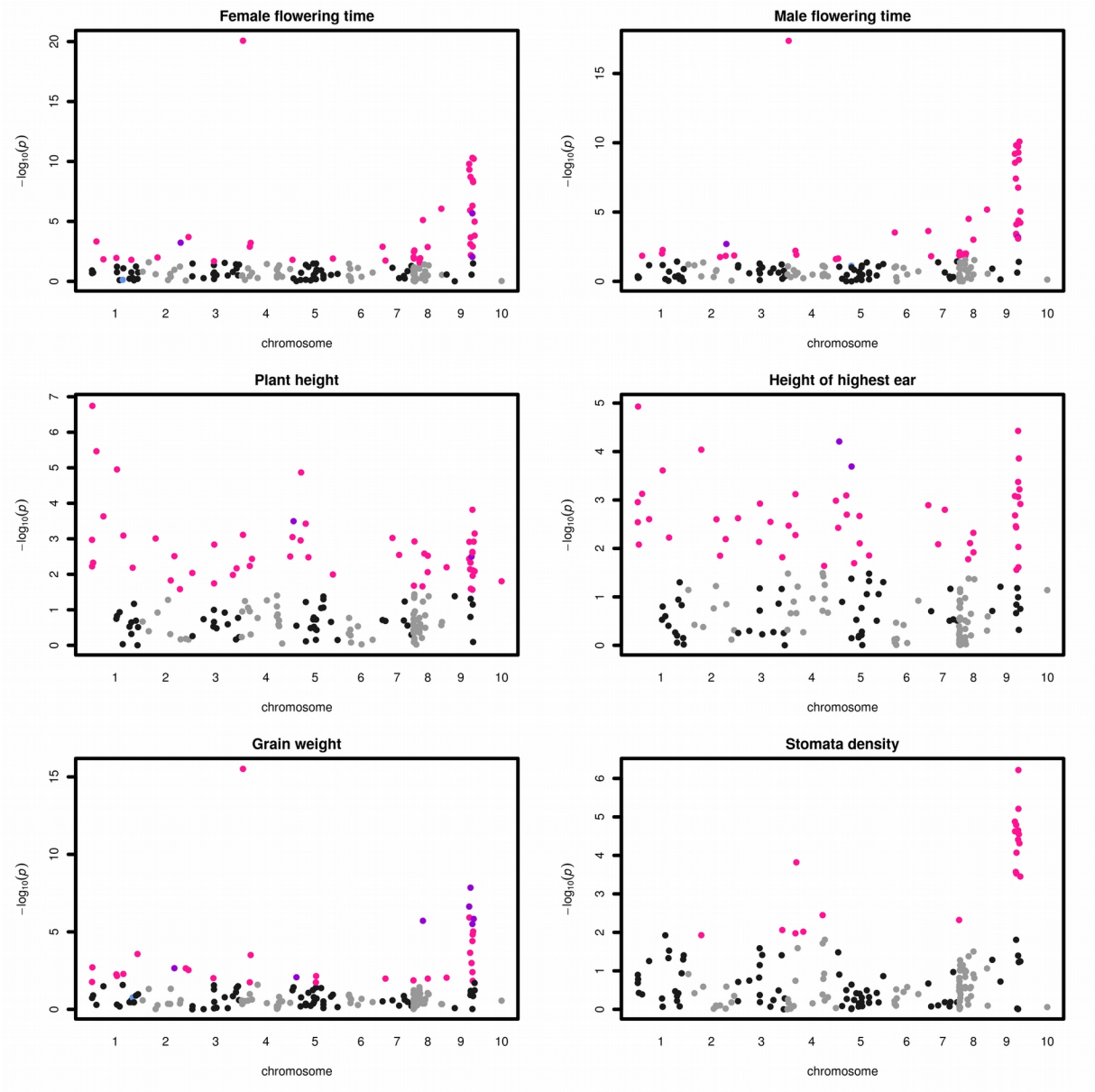


Figure 5: Manhattan plots of associations between 171 outlier SNPs and 6 phenotypic traits. X-axis indicates the positions of outlier SNPs on chromosomes 1 to 10, black and gray colors alternating per chromosome. We plotted on the Y-axis the negative Log_{10}-transformed *P* values obtained for the *K*=5 model. Significant associations (10% FDR) are indicated considering either a structure matrix at *K*=5 (pink dots), 11 populations (blue dots) or both *K*=5 and 11 populations (purple dots) models.

Altogether the 23 SNPs recovered considering a neutral genetic structure with 11 populations corresponded to 30 associations, 7 of the SNPs being associated to more than one trait (S5 Table). For all these 30 associations except in two cases (FFT with SNP_7, and MFT with SNP_28), the SNP effect did not vary among populations (non-significant SNP-by-population interaction in model M5' when we included the SNP interactions with year*field and population).

For a subset of two SNPs, we illustrated the regression between the trait value and the shift of allele frequencies with altitude (Fig 6 A&B). We estimated corresponding additive and dominance effects (S7 Table). In some cases, the intra-population effect corroborated the inter-population variation with relatively large additive effects of the same sign (Fig 6 C&D). Note that in both examples shown in Fig 6, one or the other allele was dominant. In other cases, the results were more difficult to interpret with negligible additive effect but extremely strong dominance (S7 Table, SNP_210 for instance).

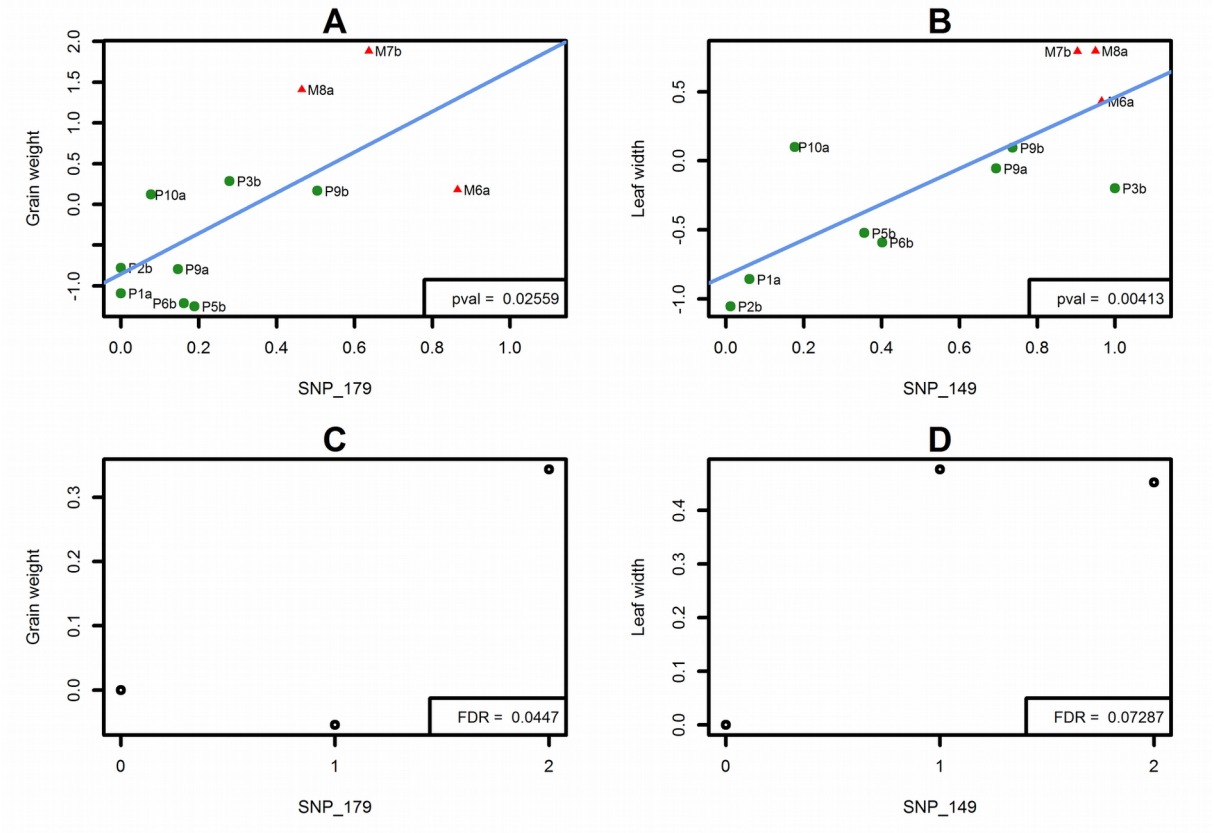


Figure 6: Regression of phenotypic average value on SNP allele frequency across populations, and within-population average phenotypic value for each SNP genotype. Per-population phenotypic average values of traits are regressed on alleles frequencies at SNP_179 (A) and SNP_149 (B) with corresponding within-population average phenotypic value per genotype (C & D). In A and B, the 11 populations of the association panel are shown with *parviglumis* (green circles) and *mexicana* (red triangles) populations sampled along gradient *a* and gradient *b*. Phenotypic average values were corrected for the experimental design (calculated as the residues of model M2). Pval refers to the P-value of the linear regression represented in blue. In C and D, genotypic effects from model M5' are expressed as the average phenotypic value of heterozygotes (1) and homozygotes for the reference (0) and the alternative allele (2). FDR values are obtained from the association analysis on 171 SNPs with correction for genetic structure using 11 population.

II.2.7 Independence of SNPs associated to phenotypes

We computed the pairwise linkage disequilibrium (LD) as measured by r^2 between the 171 outlier SNPs using the R package LDcorSV (Desrousseaux, Sandron et al. 2017). Because we were specifically interested by LD pattern between phenotypically-associated SNPs, as for the association analyses we accounted for structure and kinship computed from SSRs while estimating LD (Mangin, Siberchicot et al. 2012). The 171 outlier SNPs were distributed along the 10 chromosomes of maize, and exhibited low level of linkage disequilibrium (LD), except for SNPs located on chromosomes eight, nine, and a cluster of SNPs located on chromosome 4 (S12 Fig).

Among the 171, the subset of 23 phenotypically-associated SNPs (detected when considering the 11-population structure) displayed an excess of elevated LD values – out of 47 pairs of SNPs phenotypically-associated to a same trait, 16 pairs were contained in the 5% higher values of the LD distribution of all outlier SNP pairs. Twelve out of the 16 pairs related to grain weight, the remaining four to leaf coloration, and one pair of SNPs was associated to both traits. Noteworthy was that inversions on chromosomes 1, 4, and 9, taken together, were enriched for phenotypically-associated SNPs ($\chi^2 = 8.95$, P-value=0.0028). We recovered a borderline significant enrichment with the correction K=5 ($\chi^2 = 3.82$, P-value=0.051).

Finally, we asked whether multiple SNPs contributed independently to the phenotypic variation of a single trait. We tested a multiple SNP model where SNPs were added incrementally when significantly associated (FDR < 0.10). We found 2, 3 and 2 SNPs for female, male flowering time and height of the highest ear, respectively (S5 Table). Except for the latter trait, the SNPs were located on different chromosomes.

II.3 DISCUSSION

Plants are excellent systems to study local adaptation. First, owing to their sessile nature, local adaptation of plant populations is pervasive (Leimu and Fischer 2008). Second, environmental effects can be efficiently controlled in common garden experiments, facilitating the identification of the physiological, morphological and phenological traits influenced by spatially-variable selection (Savolainen, Lascoux et al. 2013). Identification of the determinants of complex trait variation and their covariation in natural populations is however challenging (Anderson, Willis et al. 2011).

While population genomics has brought a flurry of tools to detect footprints of local adaptation, their reliability remains questioned (Sohail, Maier et al. 2019). In addition, local adaptation and demographic history frequently follow the same geographic route, making the disentangling of trait, molecular, and environmental variation, particularly arduous. Here we investigated those links on a well-established outcrossing system, the closest wild relatives of maize, along altitudinal gradients that display considerable environmental shifts over short geographical scales.

II.3.1 The syndrome of altitudinal adaptation results from selection at multiple co-adapted traits

Common garden studies along elevation gradients have been conducted in European and North American plants species (Halbritter, Fior et al. 2018). Together with other studies, they have revealed that adaptive responses to altitude are multifarious (Körner 2007). They include physiological responses such as high photosynthetic rates (Friend, Woodward et al. 1989), tolerance to frost (Neuner 2014), biosynthesis of UV-induced phenolic components (Frohnmeyer and Staiger 2014); morphological responses with reduced stature (Byars, Papst et al. 2007; Luo, Widmer et al. 2015), modification of leaf surface (Guerin, Wen et al. 2012), increase in leaf non-glandular trichomes (Kofidis, Bosabalidis et al. 2003), modification of stomata density; and phenological responses with variation in flowering time (Mendez-Vigo, Pico et al. 2011), and reduced growth period (Oleksyn, Modrzynski et al. 1998).

Our multivariate analysis of teosinte phenotypic variation revealed a marked differentiation between teosinte subspecies along an axis of variation (21.26% of the total variation) that also discriminated populations by altitude (Fig 2A & B). The combined effects of assortative mating and environmental elevation variation may generate, in certain conditions, trait differentiation along gradients without underlying divergent selection (Soularue and Kremer 2014). While we did not measure flowering time differences among populations *in situ*, we did find evidence for long distance gene flow between gradients and subspecies (Fig 4 A & C). In addition, several lines of arguments suggest that the observed clinal patterns result from selection at independent traits and is not solely driven by differences in flowering time among populations. First, two distinct methods accounting for shared population history concur with signals of spatially-varying selection at ten out of the 18 traits (Table 1). Nine of them exhibited a clinal trend of increase/decrease of population phenotypic values with elevation (S3 Fig) within at least one of the two subspecies. This number is actually conservative, because these approaches disregard the impact of selective constraints which

in fact tend to decrease inter-population differences in phenotypes. Second, while male and female flowering times were positively correlated, they displayed only subtle correlations ($|r|<0.16$) with other spatially-varying traits except for grain weight and length ($|r| <0.33$). Third, we observed convergence at multiple phenotypes between the lowland populations from the two gradients that occurred despite their geographical and genetical distance (Fig 4) again arguing that local adaptation drives the underlying patterns.

Spatially-varying traits that displayed altitudinal trends, collectively defined a teosinte altitudinal syndrome of adaptation characterized by early-flowering, production of few tillers albeit numerous lateral branches, production of heavy, long and large grains, and decrease in stomata density. We also observed increased leaf pigmentation with elevation, although with a less significant signal (S3 Table), consistent with the pronounced difference in sheath color reported between *parviglumis* and *mexicana* (Doebley 1984; Lauter, Gustus et al. 2004). Because seeds were collected from wild populations, a potential limitation of our experimental setting is the confusion between genetic and environmental maternal effects. Environmental maternal effects could bias upward our heritability estimates. However, our results corroborate previous findings of reduced number of tillers and increased grain weight in *mexicana* compared with *parviglumis* (Smith, Goodman et al. 1981). Thus although maternal effects could not be fully discarded, we believe they were likely to be weak.

The trend towards depleted stomata density at high altitudes (S3 Fig) could arguably represent a physiological adaptation as stomata influence components of plant fitness through their control of transpiration and photosynthetic rate (Raven 2002). Indeed, in natural accessions of *A. thaliana*, stomatal traits showed signatures of local adaptation and were associated with both climatic conditions and water-use efficiency (Dittberner, Korte et al. 2018). Furthermore, previous work has shown that in arid and hot highland environments, densely-packed stomata may promote increased leaf cooling in response to desiccation (Carlson, Adams et al. 2016) and may also counteract limited photosynthetic rate with decreasing pCO_2 (Körner and Mayr 1981). Accordingly, increased stomata density at high elevation sites has been reported in alpine species such as the European beech (Bresson, Vitasse et al. 2011) as well as in populations of *Mimulus guttatus* subjected to higher precipitations in the Sierra Nevada (Kooyers, Greenlee et al. 2015). In our case, higher elevations display both *arid* environment and *cooler* temperatures during the growing season, features perhaps more comparable to other tropical mountains for which a diversity of patterns in stomatal density variation with altitude has been reported (Körner, Neumayer et al. 1989). Further work will be needed to decipher the mechanisms driving the pattern of declining

stomata density with altitude in teosintes. Altogether, the altitudinal syndrome was consistent with natural selection for rapid life-cycle shift, with early-flowering in the shorter growing season of the highlands and production of larger propagules than in the lowlands. This altitudinal syndrome evolved in spite of detectable gene flow.

Although we did not formally measure biomass production, the lower number of tillers and higher amount and size of grains in the highlands when compared with the lowlands may reflect trade-offs between allocation to grain production and vegetative growth (Jakobsson and Eriksson 2000). Because grains fell at maturity and a single teosinte individual produces hundreds of ears, we were unable to provide a proxy for total grain production. The existence of fitness-related trade-offs therefore still needs to be formally addressed.

Beyond trade-offs, our results more generally question the extent of correlations between traits. In maize, for instance, we know that female and male flowering time are positively correlated and that their genetic control is in part determined by a common set of genes (Buckler, Holland et al. 2009). They themselves further increase with yield-related traits (Moreau, Charcosset et al. 2004). Response to selection for late-flowering also led to a correlated increase in leaf number in cultivated maize (Durand, Bouchet et al. 2012), and common genetic loci have been shown to determine these traits as well (Li, Wang et al. 2015). Here we found strong positive correlations between traits: male and female flowering time, grain length and width, plant height and height of the lowest or highest ear. Strong negative correlations were observed instead between grain weight and both male and female flowering time. Trait correlations were therefore partly consistent with previous observations in maize, suggesting that they were inherited from wild ancestors.

II.3.2 Footprints of past adaptation are relevant to detect variants involved in present phenotypic variation

The overall level of differentiation in our outcrossing system fell within the range of previous estimates (23% (Aguirre-Liguori, Gaut et al. 2019)and 33% (Pyhäjärvi, Hufford et al. 2013) for samples encompassing both teosinte subspecies). It is relatively low (F_{ST} ≈22%) compared to other systems such as the selfer *Arabidopsis thaliana*, where association panels typically display maximum values of F_{ST} around 60% within 10kb-windows genome-wide (Consortium 2016). Nevertheless, correction for sample structure is key for statistical associations between genotypes and phenotypes along environmental gradients. This is because outliers that

display lowland/highland differentiation co-vary with environmental factors, which themselves may affect traits (Novembre and Barton 2018). Consistently, we found that 73.7% SNPs associated with phenotypic variation at *K*=5, but only 13.5% of them did so when considering a genetic structure with 11 populations. Except for one, the latter set of SNPs represented a subset of the former. Because teosinte subspecies differentiation was fully accounted for at *K*=5 (as shown by the clear distinction between *mexicana* populations and the rest of the samples, Fig 4A), the inflation of significant associations at *K*=5 is not due to subspecies differentiation, but rather to residual stratification among populations within genetic groups. Likewise, recent studies in humans, where global differentiation is comparatively low (Guo, Wu et al. 2018) have shown that incomplete control for population structure within European samples strongly impacts association results (Berg, Harpak et al. 2019; Sohail, Maier et al. 2019). Controlling for such structure may be even more critical in domesticated plants, where genetic structure is inferred *a posteriori* from genetic data (rather than a priori from population information) and pedigrees are often not well described. Below, we show that considering more than one correction using minor peaks delivered by the Evanno statistic (S5 Fig) can be informative.

Considering a structure with 5 genetic groups, the number of SNPs associated per trait varied from 1 to 55, with no association for leaf and grain coloration (S5 Table). False positives likely represent a greater proportion of associations at *K*=5 as illustrated by a slight excess of small P-values when compared with a correction with 11 populations for most traits (S11 Fig). Nevertheless, our analysis recovered credible candidate adaptive loci that were no longer associated when a finer-grained population structure was included in the model. For instance at *K*=5, we detected *Sugary1* (*Su1*), a gene encoding a starch debranching enzyme that was selected during maize domestication and subsequent breeding (Whitt, Wilson et al. 2002; Jaenicke-Despres, Buckler et al. 2003). We found that *Su1* was associated with variation at six traits (male and female flowering time, tassel branching, height of the highest ear, grain weight and stomata density) pointing to high pleiotropy. A previous study reported association of this gene to oil content in teosintes (Weber, Briggs et al. 2008). In maize, this gene has a demonstrated role in kernel phenotypic differences between maize genetic groups (Bouchet, Servin et al. 2013). *Su1* is therefore most probably a true-positive. That this gene was no longer recovered with the 11-population structure correction indicated that divergent selection acted among populations. Indeed, allelic frequency was highly contrasted among populations, with most populations fixed for one or the other allele, and a single population with intermediate allelic frequency. With the 11-population correction, very low power is thus left to detect the effect of *Su1* on phenotypes.

Although the confounding population structure likely influenced the genetic associations, experimental evidence indicates that an appreciable proportion of the variants recovered with both *K*=5 and 11 populations are true-positives (S5 Table). One SNP associated with female and male flowering time, as well as with plant height and grain length (at *K*=5 only for the two latter traits) maps within the *phytochrome B2* (SNP_210; *phyB2*) gene. Phytochromes are involved in perceiving light signals and are essential for growth and development in plants. The maize gene *phyB2* regulates the photoperiod-dependent floral transition, with mutants producing early flowering phenotypes and reduced plant height (Sheehan, Kennedy et al. 2007). Genes from the phosphatidylethanolamine-binding proteins (PEBPs) family – *Zea mays CENTRORADIALIS* (*ZCN*) family in maize – are also well-known to act as promotor and repressor of the floral transition in plants (Danilevskaya, Meng et al. 2007). Z*CN8* is the main floral activator of maize (Meng, Muszynski et al. 2011), and both *ZCN8* and *ZCN5* strongly associate with flowering time variation (Bouchet, Servin et al. 2013; Li, Li et al. 2016). Consistently, we found associations of male and female flowering time with *PEBP18* (SNP_15). It is interesting to note that SNPs at two flowering time genes, *phyB2* and *PEBP18*, influenced independently as well as in combination both female and male flowering time variation (S5 Table).

The proportion of genic SNPs associated to phenotypic variation was not significantly higher than that of non-genic SNPs (i.e, SNPs >1kb from a gene) ($\chi^2_{(df=1)}$ = 0.043, P-value = 0.84 at *K*=5 and $\chi^2_{(df=1)}$ =1.623 , P-value =0.020 with 11 populations) stressing the importance of considering both types of variants (Yu, Li et al. 2012). For instance, we discovered a non-genic SNP (SNP_149) that displayed a strong association with leaf width variation as well as a pattern of allele frequency shift with altitude among populations (Fig 6B).

II.3.3 Physically-linked and independent SNPs both contribute to the establishment of adaptive genetic correlations

We found limited LD among our outlier SNPs (S12 Fig) corroborating previous reports (LD decay within <100bp, (Fustier, Brandenburg et al. 2017; Aguirre-Liguori, Gaut et al. 2019)). However, the subset of phenotypically-associated SNPs displayed greater LD, a pattern likely exacerbated by three Mb-scale inversions located on chromosomes 1 (*Inv1n*), 4 (*Inv4m*) and 9 (*Inv9e*) that, taken together, were enriched for SNPs associated with environmental variables related to altitude and/or SNPs associated with phenotypic variation. Previous work (Fang, Pyhäjärvi et al. 2012; Pyhäjärvi, Hufford et al. 2013) has shown that *Inv1n* and *Inv4m* segregate within both

parviglumis and *mexicana*, while two inversions on chromosome 9, *Inv9d* and *Inv9e*, are present only in some of the highest *mexicana* populations; such that all four inversions also follow an altitudinal pattern. Our findings confirmed that three of these inversions possessed an excess of SNPs with high F_{ST} between subspecies and between low- and high-*mexicana* populations for *Inv9e* (Aguirre-Liguori, Tenaillon et al. 2017). Noteworthy *Inv9d* contains a large ear leaf width quantitative trait locus in maize (Yu, Li et al. 2012). Corroborating these results, we found consistent association between the only SNP located within this inversion and leaf width variation in teosinte populations (S5 Table). Overall, our results further strengthen the role of chromosomal inversions in teosinte altitudinal adaptation.

Because inversions suppress recombination between inverted and non-inverted genotypes, their spread has likely contributed to the emergence and maintenance of locally adaptive allelic combinations in the face of gene flow, as reported in a growing number of other models (reviewed in (Wellenreuther and Bernatchez 2018)) including insects (Ayala, Ullastres et al. 2014), fish (Barth, Berg et al. 2017), birds (Lundberg, Liedvogel et al. 2017) and plants (Lowry and Willis 2010; Twyford and Friedman 2015). But we also found three cases of multi-SNP determinism of traits (male and female flowering time and height of the highest ear, Table S5) supporting selection of genetically independent loci. Consistently with Weber et al. (Weber, Briggs et al. 2008), we found that individual SNPs account for small proportions of the phenotypic variance (S7 Table). Altogether, these observations are consistent with joint selection of complex traits determined by several alleles of small effects, some of which being maintained in linkage through selection of chromosomal rearrangements.

II.4 CONCLUSION

Elevation gradients provide an exceptional opportunity for investigating variation of functional traits in response to continuous environmental factors at short geographical scales. Here we documented patterns indicating that local adaptation, likely facilitated by the existence of chromosomal inversions, allows teosintes to cope with specific environmental conditions in spite of gene flow. We detected an altitudinal syndrome in teosintes composed of sets of independent traits evolving under spatially-varying selection. Because traits co-varied with environmental differences along gradients, however, statistical associations between genotypes and phenotypes largely

depended on control of population stratification. Yet, several of the variants we uncovered seem to underlie adaptive trait variation in teosintes. Adaptive teosinte trait variation is likely relevant for maize evolution and breeding. Whether the underlying SNPs detected in teosintes bear similar effects in maize or whether their effects differ in domesticated backgrounds will have to be further investigated.

II.5 Material and Methods

II.5.1 Description of teosinte populations and sampling

We used 37 teosinte populations of *mexicana* (16) and *parviglumis* (21) subspecies from two previous collections (Díez, Gaut et al. 2013; Aguirre-Liguori, Tenaillon et al. 2017; Fustier, Brandenburg et al. 2017) to design our sampling. These populations (S1 Table) are distributed along two altitudinal gradients (Fig 1). We plotted their altitudinal profiles using R 'raster' package (Hijmans, van Etten et al. 2018) (S1 Fig). We further obtained 19 environmental variable layers from http://idrisi.uaemex.mx/distribucion/superficies-climaticas-para-mexico. These high-resolution layers comprised monthly values from 1910 to 2009 estimated via interpolation methods (Cuervo-Robayo, Téllez-Valdés et al. 2014). We extracted values of the 19 climatic variables for each population (S1 Table). Note that high throughput sequencing (HTS) data were obtained in a previous study for six populations out of the 37 (M6a, P1a, M7b, P2b, M1b and P8b; Fig 1, S1 Table) to detect candidate genomic regions for local adaptation (Fustier, Brandenburg et al. 2017). The four highest and lowest of these populations were included in the association panel described below.

We defined an association panel of 11 populations on which to perform a genotype-phenotype association study (S1 Table). Our choice was guided by grain availability as well as the coverage of the whole climatic and altitudinal ranges. Hence, we computed Principal Component Analyses (PCA) for each gradient from environmental variables using the FactoMineR package in R (Husson, Josse et al. 2016) and added altitude to the PCA graphs as a supplementary variable. Our association panel comprised five populations from a first gradient (*a*) – two *mexicana* and three *parviglumis,* and six populations from a second gradient (*b*) – one *mexicana* and five *parviglumis* (Fig 1).

Finally, we extracted available SNP genotypes generated with the MaizeSNP50 Genotyping BeadChip for 28 populations out of our 37 populations (Aguirre-Liguori, Tenaillon et al. 2017) (S1 Table). From this available SNP dataset, we randomly sampled 1000 SNPs found to display no selection footprint (Aguirre-Liguori, Tenaillon et al. 2017), hereafter neutral SNPs. We used this panel of 28 populations to investigate correlation with environmental variation. Note that 10 out of

the 28 populations were common to our association panel, and genotypes were available for 24 to 34 individuals per population, albeit different from the ones of our association mapping panel.

II.5.2 Common garden experiments

We used two common gardens for phenotypic evaluation of the association panel (11 populations). Common gardens were located at INIFAP (Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias) experimental field stations in the state of Guanajuato in Mexico, one in Celaya municipality at the Campo Experimental Bajío (CEBAJ) (20°31’20’’ N, 100°48’44’’W) at 1750 meters of elevation, and one in San Luis de la Paz municipality at the Sitio Experimental Norte de Guanajuato (SENGUA) (21°17’55’’N, 100°30’59’’W) at 2017 meters of elevation. These locations were selected because they present intermediate altitudes (S1 Fig). The two common gardens were replicated in 2013 and 2014.

The original sampling contained 15 to 22 mother plants per population. Eight to 12 grains per mother plant were sown each year in individual pots. After one month, seedlings were transplanted in the field. Each of the four fields (2 locations, 2 years) was separated into four blocks encompassing 10 rows and 20 columns. We evaluated one offspring of ~15 mother plants from each of the 11 teosinte populations in each block, using a semi-randomized design, i.e. each row containing one or two individuals from each population, and individuals being randomized within row, leading to a total of 2,640 individual teosinte plants evaluated.

II.5.3 SSR genotyping and genetic structuring analyses on the association panel

In order to quantify the population structure and individual kinship in our association panel, we genotyped 46 SSRs (S4 Table). Primers sequences are available from the maize database project (Andorf, Cannon et al. 2016) and genotyping protocol were previously published (Camus-Kulandaivelu, Veyrieras et al. 2006). Genotyping was done at the GENTYANE platform (UMR INRA 1095, Clermont-Ferrand, France). Allele calling was performed on electropherograms with the GeneMapper® Software Applied Biosystems®. Allele binning was carried out using Autobin software (Guichoux, Lagache et al. 2011), and further checked manually.

We employed STRUCTURE Bayesian classification software to compute a genetic structure matrix on individual genotypes. Individuals with over 40% missing data were excluded from

analysis. For each number of clusters (*K* from 2 to 13), we performed 10 independent runs of 500,000 iterations after a burn-in period of 50,000 iterations, and combined these 10 replicates using the LargeKGreedy algorithm from the CLUMPP program (Jakobsson and Rosenberg 2007). We plotted the resulting clusters using DISTRUCT software. We then used the Evanno method (Evanno, Regnaut et al. 2005) to choose the optimal *K* value.

We inferred a kinship matrix **K** from the same SSRs using SPAGeDI (Hardy and Vekemans 2002). Kinship coefficients were calculated for each pair of individuals as correlation between allelic states (Loiselle, Sork et al. 1995). Since teosintes are outcrossers and expected to exhibit an elevated level of heterozygosity, we estimated intra-individual kinship to fill in the diagonal. We calculated ten kinship matrices, each excluding the SSRs from one out of the 10 chromosomes. Microsatellite data are available at: 10.6084/m9.figshare.9901472.

In order to gain insights into population history of divergence and admixture, we used 1000 neutral SNPs (i.e. SNPs genotyped by Aguirre-Liguori and collaborators (Aguirre-Liguori, Tenaillon et al. 2017) and that displayed patterns consistent with neutrality among 49 teosinte populations) genotyped on 10 out of the 11 populations of the association panel to run a TreeMix analysis (TreeMix version 1.13 (Pickrell and Pritchard 2012). TreeMix models genetic drift to infer populations splits from an outgroup as well as migration edges along a bifurcating tree. We oriented the SNPs using the previously published MaizeSNP50 Genotyping BeadChip data from the outgroup species *Tripsacum dactyloides* (Pyhäjärvi, Hufford et al. 2013). We tested from 0 to 10 migration edges. We fitted both a simple exponential and a non-linear least square model (threshold of 1%) to select the optimal number of migration edges as implemented in the OptM R package (Fitak 2019). We further verified that the proportion of variance did not substantially increase beyond the optimal selected value.

II.5.4 Phenotypic trait measurements

We evaluated a total of 18 phenotypic traits on the association panel (S2 Table). We measured six traits related to plant architecture (PL: Plant Height, HLE: Height of the Lowest Ear, HHE: Height of the Highest Ear, Til: number of Tillers, LBr: number of Lateral Branches, NoE: number of Nodes with Ears), three traits related to leave morphologies (LeL: Leaf Length, LeW: Leaf Width, LeC: Leaf Color), three traits related to reproduction (MFT: Male Flowering Time, FFT: Female Flowering Time, TBr : Tassel Branching), five traits related to grains (Gr: number of

Grains per ear, GrL: Grain Length, GrWi: Grain Width, GrWe: Grain Weight, GrC: Grain Color), and one trait related to Stomata (StD: Stomata Density). These traits were chosen because we suspected they could contribute to differences among teosinte populations based on a previous report of morphological characterization on 112 teosinte collections grown in five localities (Sanchez, Kato Yamakake et al. 1998).

We measured the traits related to plant architecture and leaves after silk emergence. Grain traits were measured at maturity. Leaf and grain coloration were evaluated on a qualitative scale. For stomata density, we sampled three leaves per plant and conserved them in humid paper in plastic bags. Analyses were undertaken at the Institute for Evolution and Biodiversity (University of Münster) as follows: 5mm blade discs were cut out from the mid length of one of the leaves and microscopic images were taken after excitation with a 488nm laser. Nine locations ($0.15mm^2$) per disc were captured with 10 images per location along the z-axis (vertically along the tissue). We automatically filtered images based on quality and estimated leaf stomata density using custom image analysis algorithms implemented in Matlab. For each sample, we calculated the median stomata density over the (up to) nine locations. To verify detection accuracy, manual counts were undertaken for 54 random samples. Automatic and manual counts were highly correlated (R^2=0.82), indicating reliable detection (see S1 Annex StomataDetection, Dittberner and de Meaux, for a detailed description). The filtered data set of phenotypic measurements is available at: 10.6084/m9.figshare.9901472.

II.5.5 Statistical analyses of phenotypic variation

In order to test for genetic effects on teosinte phenotypic variation, we decomposed phenotypic values of each trait considering a fixed population effect plus a random mother-plant effect (model M1):

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \theta_{ij} + \gamma_{k/ij} + \delta_l + \chi_{il} + \psi_{jl} + P_{m/l} + \varepsilon_{ijklm} \quad \text{(M1)}$$

where the response variable Y is the observed phenotypic value, μ is the total mean, α_i is the fixed year effect (i = 2013, 2014), β_j the fixed field effect (j = field station, *SENGUA*, *CEBAJ)*, θ_{ij} is the year by field interaction, $\gamma_{k/ij}$ is the fixed block effect (k = 1, 2, 3, 4) nested within the year-by-field combination, δ_l is the fixed effect of the population of origin (l = 1 to 11*)*, χ_{il} is the year by population interaction, ψ_{jl} is the field by population interaction, $P_{m/l}$ is the random effect of mother plant (m = 1 to 15) nested within population, and ε_{ijklm} is the individual residue. Identical notations

were used in all following models. For the distribution of the effects, the same variance was estimated within all populations. Mixed models were run using ASReml v.3.0 (Butler, Cullis et al. 2007) and MM4LMM v2.0.1 [https://rdrr.io/cran/MM4LMM/man/MM4LMM-package.html, update by F. Laporte] R packages, which both gave very similar results, and fixed effects were tested through Wald tests.

For each trait, we represented variation among populations using box-plots on mean values per mother plant adjusted for the experimental design following model M'1:

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \theta_{ij} + \gamma_{k/ij} + p_{m/l} + \varepsilon_{ijklm} \quad \text{(M1')}$$

where mother plant within population is considered as fixed. We used the function *predict* to obtain least-square means (ls-means) of each mother plant, and looked at the tendencies between population's values. All fixed models were computed using *lm* package in R, and we visually checked the assumptions of residues independence and normal distribution.

We performed a principal component analysis (PCA) on phenotypic values corrected for the experimental design, using FactoMineR package in R (Husson, Josse et al. 2016) from the residues of model M2 computed using the *lm* package in R:

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \theta_{ij} + \gamma_{k/ij} + \varepsilon_{ijklm} \quad \text{(M2)}$$

Finally, we tested for altitudinal effects on traits by considering the altitude of the sampled population (*l*) as a covariate (ALT) and its interaction with year and field in model M3:

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \theta_{ij} + \gamma_{k/ij} + c\,.\,ALT_l + a_i\,.\,ALT_l + b_j\,.\,ALT_l + P_{m/l} + \varepsilon_{ijklm} \quad \text{(M3)}$$

where all terms are equal to those in model M1 except that the fixed effect of the population of origin was replaced by a regression on the population altitude (ALT_l).

II.5.6 Detection of selection acting on phenotypic traits

We aimed at detecting traits evolving under spatially varying selection by comparing phenotypic to neutral genotypic differentiation. Q_{st} is a statistic analogous to F_{ST} but for quantitative traits, which can be described as the proportion of phenotypic variation explained by differences

among populations (Spitze 1993; Gilbert and Whitlock 2015). Significant differences between Q_{ST} and F_{ST} can be interpreted as evidence for spatially-varying ($Q_{ST}>F_{ST}$) selection (Holsinger and Weir 2009). We used the R package *QstFstComp* (Gilbert and Whitlock 2015) that is adequate for experimental designs with randomized half-sibs in outcrossing species. We used individuals that were both genotyped and phenotyped on the association panel to establish the distribution of the difference between statistics (Q_{ST}-F_{ST}) under the neutral hypothesis of evolution by drift - using the half-sib dam breeding design and 1000 resamples. We next compared it to the observed difference with 95% threshold cutoff value in order to detect traits under spatially-varying selection.

In addition to Q_{ST}-F_{ST} analyses, we employed the DRIFTSEL R package (Karhunen, Merilä et al. 2013) to test for signal of selection of traits while accounting for drift-driven population divergence and genetic relatedness among individuals (half-sib design). DRIFTSEL is a Bayesian method that compares the probability distribution of predicted and observed mean additive genetic values. It provides the S statistic as output, which measures the posterior probability that the observed population divergence arose under divergent selection ($S\sim1$), stabilizing selection ($S\sim0$) or genetic drift (intermediate S values) (Ovaskainen, Karhunen et al. 2011). It is particularly powerful for small datasets, and can distinguish between drift and selection even when Q_{ST}-F_{ST} are equal (Ovaskainen, Karhunen et al. 2011). We first applied RAFM to estimate the F_{ST} value across populations, and the population-by-population coancestry coefficient matrix. We next fitted both the RAFM and DRIFTSEL models with 15,000 MCMC iterations, discarded the first 5,000 iterations as a transient, and thinned the remaining by 10 to provide 1000 samples from the posterior distribution. Note that DRIFTSEL was slightly modified because we had information only about the dams, but not the sires, of the phenotyped individuals. We thus modified DRIFTSEL with the conservative assumption of all sires being unrelated. Because DRIFTSEL does not require that the same individuals were both genotyped and phenotyped, we used SSRs and phenotype data of the association panel as well as the set of neutral SNPs and phenotype data on 10 out of the 11 populations. For the SNP analyses, we selected out of the 1000 neutral SNPs the 465 most informative SNPs based on the following criteria: frequency of the less common variant at least 10%, and proportion of missing data at most 1%. Finally, we estimated from DRIFTSEL the posterior probability of the ancestral population mean for each trait as well as deviations of each population from these values.

Both Q_{ST}-F_{ST} and DRIFTSEL rely on the assumption that the observed phenotypic variation was determined by additive genotypic variation. We thus estimated narrow-sense heritability for each trait in each population to estimate the proportion of additive variance in performance. We

calculated per population narrow-sense heritabilites as the ratio of the estimated additive genetic variance over the total phenotypic variance on our common garden measurements using the MCMCglmm R package (Hadfield 2010) where half sib family is the single random factor, and the design (block nested within year and field) is corrected as fixed factor. For three grain-related traits, we also ran the same model but including mother plants phenotypic values calculated from the remaining grains not sown. We ran 100,000 iterations with 10,000 burn-in, inverse gamma (0.001; 0.001) as priors. We then calculated the mean and standard deviation of the 11 per population h^2 estimates.

II.5.7 Pairwise correlations between traits

We evaluated pairwise-correlations between traits by correlating the residues obtained from model M4, that corrects the experiment design (year, field and blocks) as well as the underlying genetic structure estimated from SSRs:

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \theta_j + \gamma_{k/ij} + \sum_{n=1}^{4} b_n . C^n_{ijklm} + \varepsilon_{ijklm} \quad \text{(M4)}$$

where b_n is the slope of the regression of Y on the n^{th} structure covariate C^n. Structure covariate values (C^n covariates, from STRUCTURE output) were calculated at the individual level, i.e. for each offspring of mother plant *m* from population *l,* grown in the year *i* field *j* and block *k*. C^n are thus declared with *ijklm* indices, although they are purely genetic covariates.

II.5.8 Genotyping of outlier SNPs on 28 populations

We extracted total DNA from each individual plant of the association panel as well as 20 individuals from each of the 18 remaining populations that were not included in the association panel (Table 1). Extractions were performed from 30 mg of lyophilized adult leaf material following recommendations of DNeasy 96 Plant Kit manufacturer (QIAGEN, Valencia, CA, USA). We genotyped outlier SNPs using Kompetitive Allele Specific PCR technology (KASPar, LGC Group) (Semagn, Babu et al. 2014). Data for outlier SNPs are available at: 10.6084/m9.figshare.9901472.

Among SNPs identified as potentially involved in local adaptation, 270 were designed for KASPar assays, among which 218 delivered accurate quality data. Of the 218 SNPs, 141 were detected as outliers in two previous studies using a combination of statistical methods – including F_{ST}-scans (Weir and Hill 2002), Bayescan (Foll and Gaggiotti 2008) and Bayenv2 (Günther and Coop 2013; Günther and Coop 2016), Bayescenv (De Villemereuil and Gaggiotti 2015) – applied to either six of our teosinte populations (Fustier, Brandenburg et al. 2017) or to a broader set of 49 populations genotyped by the Illumina® MaizeSNP50 BeadChip (Aguirre-Liguori, Tenaillon et al. 2017). The remaining outlier SNPs (77) were detected by F_{ST}-scans from six populations (S7 Fig, S5 Table), following a simplified version of the rationale in (Fustier, Brandenburg et al. 2017) by considering only differentiation statistics: SNPs were selected if they displayed both a high differentiation (5% highest F_{ST} values) between highland and lowland populations in at least one of the two gradients, and a high differentiation (5% highest F_{ST} values) between highland and lowland populations either within *parviglumis* (P2b and P8b) or within *mexicana* (M7b and M1b) or both in gradient *b* (S1 Fig). We thereby avoided SNPs fixed between the two subspecies.

II.5.9 Association mapping

We tested the association of phenotypic measurements with outlier SNPs on a subset of individuals for which (1) phenotypic measurements were available, (2) at least 60% of outlier SNPs were adequately genotyped, and (3) kinship and cluster membership values were available from SSR genotyping. For association, we removed SNPs with minor allele frequency lower than 5%.

In order to detect statistical associations between outlier SNPs and phenotypic variation, we used the following mixed model derived from (Yu and Pressoir et al., 2005):

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \theta_{ij} + \gamma_{k/ij} + \sum_{n=1}^{4} b_n . C_{ijklm}^{n} + \zeta_o + u_{ijklm} + \varepsilon_{ijklm} \qquad \text{(M5)}$$

where ζ is the fixed bi-allelic SNP factor with one level for each of the three genotypes (o=0, 1, 2; with o=1 for heterozygous individuals), and u_{ijklm} is the random genetic effect of the individual. We assumed that the vector of u_{ijklm} effects followed a $N(0,\mathbf{K}\,\sigma^2_u)$ distribution, where **K** is the kinship matrix computed as described above.

A variant of model M5 was employed to test for SNP association to traits, while correcting for structure as the effect of population membership (δ_l), δ being a factor with 11 levels (populations):

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \theta_{ij} + \gamma_{k/ij} + \delta_l + \zeta_o + u_{ijklm} + \varepsilon_{ijklm} \qquad \text{(M5')}$$

In order to avoid overcorrection of neutral genetic structure and improve power, we ran the two models independently for each chromosome using a kinship matrix **K** estimated from all SSRs except those contained in the chromosome of the tested SNP (Rincent, Moreau et al. 2014). We tested SNP effects through the Wald statistics, and applied a 10% False Discovery Rate (FDR) threshold for each phenotype separately. In order to validate the correction for genetic structure, the 38 multiallelic SSR genotypes were transformed into biallelic genotypes, filtered for MAF > 5%, and used to run associations with the complete M5 and M5' models, as well as the M5 models excluding either kinship or both structure and kinship. For each trait, we generated QQplots of P-values for each of these models.

Multiple SNP models were built by successively adding at each step the most significant SNP, as long as its FDR was lower than 0.10. We controlled for population structure at *Pop*=11 and used the kinship matrix that excluded the SSR on the same chromosome as the last tested SNP.

II.5.10 Environmental correlation of outlier SNPs

We tested associations between allelic frequency at 171 outlier SNPs and environmental variables across 28 populations, using Bayenv 2.0 (Coop, Witonsky et al. 2010; Twyford and Friedman 2015). Because environmental variables are highly correlated, we used the first two principal component axes from the environmental PCA analysis (PCenv1 and PCenv2) to run Bayenv 2.0. This software requires a neutral covariance matrix, that we computed from the available dataset of 1000 neutral SNPs (S1 Table). We performed 100,000 iterations, saving the matrix every 500 iterations. We then tested the correlation of these to the last matrix obtained, as well as to an F_{ST} matrix calculated with BEDASSLE (Bradburd, Ralph et al. 2013), as described in (Aguirre-Liguori, Tenaillon et al. 2017).

For each outlier SNP, we compared the posterior probability of a model that included an environmental factor (PCenv1 or PCenv2) to a null model. We determined a 5% threshold for significance of environmental association by running 100,000 iterations on neutral SNPs. We carried out five independent runs for each outlier SNP and evaluated their consistency from the coefficient of variation of the Bayes factors calculated among runs.

In order to test whether environmental distance was a better predictor of allele frequencies at candidate SNPs than geography, we used multiple regression on distance matrices (MRM, (Lichstein 2007)) implemented in the ecodist R package (Goslee and Urban 2007) for each outlier SNP. We used pairwise F_{ST} values as the response distance matrix and the geographic and environmental distance matrices as explanatory matrices. We evaluated the significance of regression coefficients by 1000 permutations and iterations of the MRM. We determined the total number of environmentally and geographically associated SNPs (P-value<0.05) among outliers. We employed the same methodology for our set of 1000 neutral SNPs.

ETHICS STATEMENT

All the field work has been done in Mexico in collaboration with Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Celaya in Celaya.

ACKNOWLEDGMENTS

We thank Jessica Melique for her help in gathering stomata data. We are very grateful to Angelica Cibrian at Langebio (Cinvestav, Irapuato, Mexico) who let us use her lab to lyophilize our samples, and store them. Valeria Souza greatly helped us with logistic support in Mexico. Insights from Delphine Legrand, Pierre de Villemereuil and Elodie Marchadier were important for analyzing correlations between traits and to estimate heritabilities. We would like to thank warmly five anonymous reviewers for their insightful comments on our work.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: MIT, DM. Contributed plant material: LEE, JAA-L, SM, MIT. DNA extractions: HC, AR, M-AF, DM. Conducted field experiments and performed phenotypic measurements: M-AF, MIT, MGC, SM. Outlier SNP detection: M-AF. Microsatellite genotyping: AV, HC, FD, MF. Neutral SNP genotyping: LEE, JAA-L. Stomata density images and pipeline: JdD, HD. Statistical methodology: DM, M-AF, NEM-A, OO, LM, MIT, JAA-L. Implemented methods, wrote scripts: M-AF, NEM-A, OO, JAA-L. Analyzed the data: M-AF, NEM-A, DM. Prepared Tables and Figures: M-AF, NEM-A. Discussed interpretations: M-AF,

NEM-A, DG, OO, LM, JAA-L, JdM, YV, DM, MIT. Wrote the paper: NEM-A, M-AF, MIT. Corrected and commented the manuscript: all authors.

FUNDING

This work was supported by two grants overseen by the French National Research Agency (ANR) (Project ANR 12-ADAP-0002-01) to MIT and YV, and the ECOSNord/ANUIES/CONACYT/SEP project M12A01, CONACYT-ANUIES 207571 to MIT and LEE. LEE research and JAA-L and M-AF postdoctoral salaries were supported by CONACYT-Mexico Investigación Científica Básica CB2011/167826 awarded to LEE. M-AF was funded by the Project ANR 12-ADAP-0002-01 and NEM-A was funded by a CONACYT PhD fellowship grant no. 57916/310738. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. OO was funded by the Academy of Finland (grant 109581), the Jane and Aatos Erkko Foundation, and the Research Council of Norway (SFF-III grant 221257).

II.6 REFERENCES

Aguirre-Liguori, J. A., M. I. Tenaillon, et al. (2017). "Connecting genomic patterns of local adaptation and niche suitability in teosintes." Molecular Ecology **26**: 4226-4240.

Aguirre-Liguori, J. A., B. S. Gaut, et al. (2019). " Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (Zea mays parviglumis and Zea mays mexicana) " Molecular Ecology: 2814-2830.

Anderson, J. T., J. H. Willis, et al. (2011). "Evolutionary genetics of plant adaptation." Trends in Genetics **27**: 258-266.

Andorf, C. M., E. K. Cannon, et al. (2016). "MaizeGDB update: New tools, data and interface for the maize model organism database." Nucleic Acids Research **44**: 1195-1201.

Ayala, D., A. Ullastres, et al. (2014). "Adaptation through chromosomal inversions in Anopheles." Frontiers in Genetics **5**: 1-10.

Barth, J. M. I., P. R. Berg, et al. (2017). "Genome architecture enables local adaptation of Atlantic cod despite high connectivity." Molecular Ecology **26**: 4452-4466.

Barton, N., J. Hermisson, et al. (2019). "Why structure matters." eLife **8**.

Beaumont, M. A. and R. A. Nichols (1996). "Evaluating loci for use in the genetic analysis of population structure." Proceedings of the Royal Society B: Biological Sciences **263**: 1619-1626.

Berg, J. J., A. Harpak, et al. (2019). "Reduced signal for polygenic adaptation of height in UK Biobank." eLife **8**: 1-47.

Bierne, N., J. Welch, et al. (2011). "The coupling hypothesis: Why genome scans may fail to map local adaptation genes." Molecular Ecology **20**: 2044-2072.

Bilinski, P., P. S. Albert, et al. (2018). "Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays." PLoS Genetics **14**.

Bonhomme, M., C. Chevalet, et al. (2010). "Detecting selection in population trees: The Lewontin and Krakauer test extended." Genetics **186**: 241-262.

Bouchet, S., B. Servin, et al. (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**.

Bradburd, G. S., P. L. Ralph, et al. (2013). "Disentangling the effects of geographic and ecological isolation on genetic differentiation." Evolution **67**: 3258-3273.

Bradshaw, A. D. (1984). "Ecological significance of genetic variation between populations." Perspectives on plant population ecology: 213-228.

Bresson, C. C., Y. Vitasse, et al. (2011). "To what extent is altitudinal variation of functional traits driven by genetic adaptation in European oak and beech?" Tree Physiology **31**: 1164-1174.

Buckler, E. S., J. B. Holland, et al. (2009). "The genetic architecture of maize flowering time." Science **325**: 714-718.

Bulmer, M. G. G. (1972). "Multiple niche polymorphism." The American Naturalist **106**: 254-257.

Butler, D., B. R. Cullis, et al. (2007). "ASReml-R reference manual." Technical Report.

Byars, S. G., W. Papst, et al. (2007). "Local adaptation and cogradient selection in the alpine plant, Poa hiemata, along a narrow altitudinal gradient." Evolution **61**: 2925-2941.

Camus-Kulandaivelu, L., J. B. Veyrieras, et al. (2006). "Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the Dwarf8 gene." Genetics **172**: 2449-2463.

Carlson, J. E., C. A. Adams, et al. (2016). "Intraspecific variation in stomatal traits, leaf traits and physiology reflects adaptation along aridity gradients in a South African shrub." Annals of Botany **117**: 195-207.

Consortium, G. (2016). "1,135 Genomes reveal the global pattern of polymorphism in Arabidopsis thaliana." Cell **166**: 481-491.

Coop, G., D. Witonsky, et al. (2010). "Using environmental correlations to identify loci underlying local adaptation." Genetics **185**: 1411-1423.

Cuervo-Robayo, A. P., O. Téllez-Valdés, et al. (2014). "An update of high-resolution monthly climate surfaces for Mexico." International Journal of Climatology **34**: 2427-2437.

Danilevskaya, O. N., X. Meng, et al. (2007). "A genomic and expression compendium of the expanded PEBP gene family from maize." Plant Physiology **146**: 250-264.

De Mita, S., A. C. Thuillet, et al. (2013). "Detecting selection along environmental gradients: Analysis of eight methods and their effectiveness for outbreeding and selfing populations." Molecular Ecology **22**: 1383-1399.

De Villemereuil, P. and O. E. Gaggiotti (2015). "A new FST-based method to uncover local adaptation using environmental variables." Methods in Ecology and Evolution.

Desrousseaux, A. D., F. Sandron, et al. (2017). "Package ‘ LDcorSV ’."

Diez, C. M., B. S. Gaut, et al. (2013). "Genome size variation in wild and cultivated maize along altitudinal gradients." New Phytologist **199**: 264-276.

Díez, C. M., B. S. Gaut, et al. (2013). "Genome size variation in wild and cultivated maize along altitudinal gradients." New Phytologist **199**(1): 264-276.

Dittberner, H., A. Korte, et al. (2018). "Natural variation in stomata size contributes to the local adaptation of water-use efficiency in Arabidopsis thaliana." Molecular Ecology: 4052-4065.

Doebley, J. F. (1984). "Maize introgression into teosinte -- a reappraisal." Annals of the Missouri Botanical Garden **71**: 1100-1113.

Durand, E., S. Bouchet, et al. (2012). "Flowering time in maize: Linkage and epistasis at a major effect locus." Genetics **190**: 1547-1562.

Endler, J. A. (1986). "Natural selection in the wild." 354.

Evanno, G., S. Regnaut, et al. (2005). "Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study." Molecular Ecology **14**: 2611-2620.

Excoffier, L. a. H., T and Foll, Matthieu (2009). "Detecting loci under selection in a hierarchically structured population." Heredity **103**: 285-298.

Fang, Z., T. Pyhäjärvi, et al. (2012). "Megabase-scale inversion polymorphism in the wild ancestor of maize." Genetics **191**: 883-894.

Fitak, R. R. s. (2019). "optM: an R package to optimize the number of migration edges using threshold models." Journal of Heredity.

Foll, M. and O. Gaggiotti (2008). "A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective." Genetics **180**: 977-993.

Fournier-Level, A., A. Korte, et al. (2011). "A map of local adaptation in Arabidopsis thaliana." Science **334**: 86-89.

Frichot, E., S. D. Schoville, et al. (2013). "Testing for associations between loci and environmental gradients using latent factor mixed models." Molecular Biology and Evolution **30**: 1687-1699.

Friend, A. D., F. I. Woodward, et al. (1989). "Field measurements of photosynthesis, stomatal conductance, leaf nitrogen and δ 13 C along altitudinal gradients in Scotland." Functional Ecology **3**: 117.

Frohnmeyer, H. and D. Staiger (2014). "Update on ultraviolet-B light responses ultraviolet-B radiation-mediated responses in plants. Balancing damage and protection." **133**: 1420-1428.

Fustier, M. A., J. T. Brandenburg, et al. (2017). "Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples." Molecular Ecology **26**: 2738-2756.

Garcia-Ramos, G. and M. Kirkpatrick (1997). "Genetic models of adaptation and gene flow in peripherial populations." Evolution **51**: 21-28.

Gautier, M. (2015). "Genome-wide scan for adaptive divergence and association with population-specific covariates." Genetics **201**: 1555-1579.

Gay, L., P. A. Crochet, et al. (2008). "Comparing clines on molecular and phenotypic traits in hybrid zones: A window on tension zone models." Evolution **62**: 2789-2806.

Gilbert, K. J. and M. C. Whitlock (2015). "QST-FST comparisons with unbalanced half-sib designs." Molecular Ecology Resources **15**: 262-267.

Goslee, S. C. and D. L. Urban (2007). "Journal of Statistical Software The ecodist Package for Dissimilarity-based Analysis of Ecological Data."

Guerin, G. R., H. Wen, et al. (2012). "Leaf morphology shift linked to climate change." Population Ecology: 882-886.

Guichoux, E., S. Lagache, et al. (2011). "Current trends in microsatellite genotyping." Molecular Ecology Resources **11**: 591-611.

Guillot, G., S. Renaud, et al. (2012). "A unifying model for the analysis of phenotypic, genetic, and geographic data." **61**: 897-911.

Günther, T. and G. Coop (2013). "Robust identification of local adaptation from allele frequencies." Genetics Society of America **195**: 205-220.

Günther, T. and G. Coop (2016). "A Short Manual for Bayenv2.0."

Guo, J., Y. Wu, et al. (2018). "Global genetic differentiation of complex traits shaped by natural selection in humans." Nature Communications **9**(1): 1865.

Haasl, R. J. and B. A. Payseur (2016). "Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication." Molecular Ecology **25**: 5-23.

Hadfield, J. D. (2010). "MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package." Journal of Statistical Software **33**.

Halbritter, A. H., S. Fior, et al. (2018). "Trait differentiation and adaptation of plants along elevation gradients." Journal of Evolutionary Biology **31**(6): 784-800.

Hancock, A. M., B. Brachi, et al. (2011). "Adaptation to climate across the Arabidopsis thaliana genome." Science **334**: 83-86.

Hardy, O. J. and X. Vekemans (2002). "spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels." Molecular Ecology Notes **2**: 618-620.

Hereford, J. (2009). "A quantitative survey of local adaptation and fitness trade-offs." The American Naturalist **173**: 579-588.

Hijmans, R. J., J. van Etten, et al. (2018). "Package 'raster': geographic data analysis and modeling."

Hoban, S., J. L. Kelley, et al. (2016). "Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions." The American Naturalist **188**: 379-397.

Holsinger, K. E. and B. S. Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting F(ST)." Nature reviews. Genetics **10**: 639-650.

Hufford, M. B., E. Martínez-Meyer, et al. (2012). "Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight." PLoS ONE **7**.

Husson, F., J. Josse, et al. (2016). "Package ' FactoMineR '. An R package." 96.

Jaenicke-Despres, V., E. S. Buckler, et al. (2003). "Early allelic selection in maize as revealed by ancient DNA." **302**: 1206-1209.

Jakobsson, A. and O. Eriksson (2000). "A comparative study of seed number, seed size, seedling size and recruitment in grassland plants." Oikos **88**: 494-502.

Jakobsson, M. and N. A. Rosenberg (2007). "CLUster Matching and Permutation Program Version 1.1.2."

Joost, S., A. Bonin, et al. (2007). "A spatial analysis method (SAM) to detect candidate loci for selection: Towards a landscape genomics approach to adaptation." Molecular Ecology **16**: 3955-3969.

Karhunen, M., J. Merilä, et al. (2013). "driftsel: An R package for detecting signals of natural selection in quantitative traits." Molecular Ecology Resources **13**: 746-754.

Kawakami, T., T. J. Morgan, et al. (2011). "Natural selection drives clinal life history patterns in the perennial sunflower species, Helianthus maximiliani." Molecular Ecology **20**: 2318-2328.

Kawecki, T. J. and D. Ebert (2004). "Conceptual issues in local adaptation." Ecology Letters **7**: 1225-1241.

Kirkpatrick, M. and N. Barton (2006). "Chromosome inversions, local adaptation and speciation." Genetics.

Kofidis, G., A. M. Bosabalidis, et al. (2003). "Contemporary seasonal and altitudinal variations of leaf structural features in oregano (Origanum vulgare L.)." Annals of Botany **92**(5): 635-645.

Kooyers, N. J., A. B. Greenlee, et al. (2015). "Replicate altitudinal clines reveal that evolutionary flexibility underlies adaptation to drought stress in annual Mimulus guttatus." New Phytologist **206**: 152-165.

Körner, C. (2007). "The use of 'altitude' in ecological research." Trends in Ecology and Evolution **22**: 569-574.

Körner, C. and R. Mayr (1981). Stomatal behaviour in alpine plant communities between 600 and 2600 metres above sea level. Plants and their Atmospheric Environment. J. Grace, E. D. Ford and P. G. Jarvis, Blackwell, Oxford,**:** pp 205-218.

Körner, C., M. Neumayer, et al. (1989). "Functional morphology of mountain plants." Flora **182**: 353-383.

Lande, R. (1976). "Natural selection and random genetic drift in phenotypic evolution." Evolution **30**: 314-334.

Lande, R. (1992). "Neutral theory of quantitative genetic variance in an island model with local extinction and colonization." Evolution **46**: 381-389.

Lauter, N., C. Gustus, et al. (2004). "The inheritance and evolution of leaf pigmentation and pubescence in teosinte." Genetics Society of America **167**: 1949-1959.

Le Corre, V. and A. Kremer (2012). "The genetic differentiation at quantitative trait loci under local adaptation." Molecular Ecology **21**: 1548-1566.

Legrand, D., N. Larranaga, et al. (2016). "Evolution of a butterfly dispersal syndrome."

Leimu, R. and M. Fischer (2008). "A meta-analysis of local adaptation in plants." PLoS ONE **3**.

Lenormand, T. (2002). "Gene flow and the limits to natural selection." Trends in Ecology and Evolution **17**: 183-189.

Lewontin, R. C. and J. Krakauer (1973). "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms." Genetics **74**: 175-195.

Li, D., X. Wang, et al. (2015). "The genetic architecture of leaf number and its genetic relationship to flowering time in maize." New Phytologist **210**: 256-268.

Li, Y. X., C. Li, et al. (2016). "Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population." The Plant journal : for cell and molecular biology **86**: 391-402.

Lichstein, J. W. (2007). "Multiple regression on distance matrices: A multivariate spatial analysis tool." Plant Ecology **188**: 117-131.

Loiselle, B. A., V. L. Sork, et al. (1995). "Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae)." American Journal of Botany. **82**: 1420-1425.

Lotterhos, K. E. and M. C. Whitlock (2014). "Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests." Molecular Ecology **23**: 2178-2192.

Lowry, D. B. and J. H. Willis (2010). "A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation." PLoS Biology **8**.

Lundberg, M., M. Liedvogel, et al. (2017). "Genetic differences between willow warbler migratory phenotypes are few and cluster in large haplotype blocks." Evolution Letters: 155-168.

Luo, Y., A. Widmer, et al. (2015). "The roles of genetic drift and natural selection in quantitative trait divergence along an altitudinal gradient in Arabidopsis thaliana." Heredity **114**: 220-228.

Luquet, E., J.-P. Léna, et al. (2015). "Phenotypic divergence of the common toad (Bufo bufo) along an altitudinal gradient: evidence for local adaptation." Heredity **114**: 69-79.

Mangin, B., A. Siberchicot, et al. (2012). "Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness." Heredity **108**: 285-291.

McKinney, G. J., A. Varian, et al. (2014). "Genetic and morphological divergence in three strains of brook trout Salvelinus fontinalis commonly stocked in Lake Superior." PLoS ONE **9**(12): e113809.

Mendez-Vigo, B., F. X. Pico, et al. (2011). "Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in Arabidopsis." Plant Physiology **157**: 1942-1955.

Meng, X., M. G. Muszynski, et al. (2011). "The FT-Like ZCN8 Gene Functions as a Floral Activator and Is Involved in Photoperiod Sensitivity in Maize." The Plant Cell **23**: 942-960.

Moreau, L., A. Charcosset, et al. (2004). "Use of trial clustering to study QTL x environment effects for grain yield and related traits in maize." Theoretical and Applied Genetics **110**: 92-105.

Moyers, B. T. and L. H. Rieseberg (2016). "Remarkable life history polymorphism may be evolving under divergent selection in the silverleaf sunflower." Molecular Ecology **25**: 3817-3830.

Neuner, G. (2014). "Frost resistance in alpine woody plants." Frontiers in Plant Science **5**.

Novembre, J. and N. H. Barton (2018). "Tread Lightly Interpreting Polygenic Tests of Selection." Genetics **208**(4): 1351-1355.

Oleksyn, J., J. Modrzynski, et al. (1998). "Growth and physiology of Picea abies populations from elevational transects: common garden evidence for altitudinal ecotypes and cold adaptation." 573-590.

Ovaskainen, O., M. Karhunen, et al. (2011). "A new method to uncover signatures of divergent and stabilizing selection in quantitative traits." Genetics **189**: 621-632.

Pickrell, J. K. and J. K. Pritchard (2012). "Inference of population splits and mixtures from genome-wide allele frequency data." PLOS Genetics **8**(11): e1002967.

Poncet, B. N., D. Herrmann, et al. (2010). "Tracking genes of ecological relevance using a genome scan in two independent regional population samples of Arabis alpina." Molecular Ecology **19**: 2896-2907.

Pyhäjärvi, T., M. B. Hufford, et al. (2013). "Complex patterns of local adaptation in teosinte." Genome Biology and Evolution **5**: 1594-1609.

Raven, J., A (2002). "Selection pressures on stomatal evolution." New Phytologist **153**: 371-386.

Rincent, R., L. Moreau, et al. (2014). "Recovering power in association mapping panels with variable levels of linkage disequilibrium." Genetics **197**: 375-387.

Roschanski, A. M., K. Csilléry, et al. (2016). "Evidence of divergent selection for drought and cold tolerance at landscape and local scales in Abies alba Mill. in the French Mediterranean Alps." Molecular Ecology **25**: 776-794.

Ross-Ibarra, J., M. Tenaillon, et al. (2009). "Historical divergence and gene flow in the genus Zea." Genetics **181**: 1399-1413.

Rundle, H. D. and P. Nosil (2005). "Ecological speciation." Ecology Letters **8**: 336-352.

Sanchez, J. d. J., T. A. Kato Yamakake, et al. (1998). "Distribución y caracterización del teocintle." 165.

Savolainen, O., M. Lascoux, et al. (2013). "Ecological genomics of local adaptation." Nature reviews. Genetics **14**: 807-820.

Semagn, K., R. Babu, et al. (2014). "Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement." Molecular Breeding **33**: 1-14.

Sheehan, M. J., L. M. Kennedy, et al. (2007). "Subfunctionalization of PhyB1 and PhyB2 in the control of seedling and mature plant traits in maize." Plant Journal **49**: 338-353.

Slatkin, M. (1985). "Rare alleles as indicators of gene flow." Evolution **39(1)**: 53-65.

Smith, J. S. C., M. M. Goodman, et al. (1981). "Variation within teosinte. I. Numerical analysis of morphological data." Economic Botany **35**: 187-203.

Sohail, M., R. M. Maier, et al. (2019). "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies." eLife **8**.

Soularue, J. P. and A. Kremer (2014). "Evolutionary responses of tree phenology to the combined effects of assortative mating, gene flow and divergent selection." Heredity **113**: 485-494.

Spitze, K. (1993). "Population structure in Daphina obtusa: quantitative genetic and allozymic variation." Genetics Society of America **135**: 367-374.

Tiffin, P. and J. Ross-Ibarra (2014). "Advances and limits of using population genetics to understand local adaptation." Trends in Ecology and Evolution **29**: 673-680.

Twyford, A. D. and J. Friedman (2015). "Adaptive divergence in the monkey flower Mimulus guttatus is maintained by a chromosomal inversion." Evolution **69**: 1476-1486.

Vitalis, R., K. Dawson, et al. (2001). "Interpretation of variation across marker loci as evidence of selection." Genetics **158**: 1811-1823.

Weber, A. L., W. H. Briggs, et al. (2008). "The genetic architecture of complex traits in teosinte (Zea mays ssp. parviglumis): New evidence from association mapping." Genetics **180**: 1221-1232.

Weir, B. S. and W. G. Hill (2002). "Estimating F-statistics." Annu. Rev. Genet **36**: 721-750.

Wellenreuther, M. and L. Bernatchez (2018). "Eco-evolutionary genomics of chromosomal inversions." Trends in Ecology and Evolution **33**: 427-440.

Whitlock, M. C. (1999). "Neutral additive genetic variance in a metapopulation." Genetics Research **74**: 215-221.

Whitlock, M. C. and R. Gomulkiewicz (2005). "Probability of fixation in a heterogeneous environment." Genetics: 1-40.

Whitlock, M. C. (2015). "Modern approaches to local adaptation." The American Naturalist **186**.

Whitt, S. R., L. M. Wilson, et al. (2002). "Genetic diversity and selection in the maize starch pathway." Proceedings of the National Academy of Sciences of the United States of America **99**: 12959-12962.

Wright, S. (1951). "The genetical structure of populations." Annals of Eugenics **15**: 323-354.

Yeaman, S. and S. P. Otto (2011). "Establishment and maintenance of adaptive genetic divergence under migration, selection and drift." Evolution **67**: 2123-2129.

Yu J, G. Pressoir, et al. (2005) "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." Nature Genetics. **38**(2):203–8.

Yi, X., Y. Liang, et al. (2010). "Sequencing of fifty human exomes reveals adaptations to high altitude." Science **329**: 75-78.

Yu, J., X. Li, et al. (2012). "Genic and non-genic contributions to natural variation of quantitative traits in maize." Genome research: 2436-2444.

II.7 Supplementary Figures

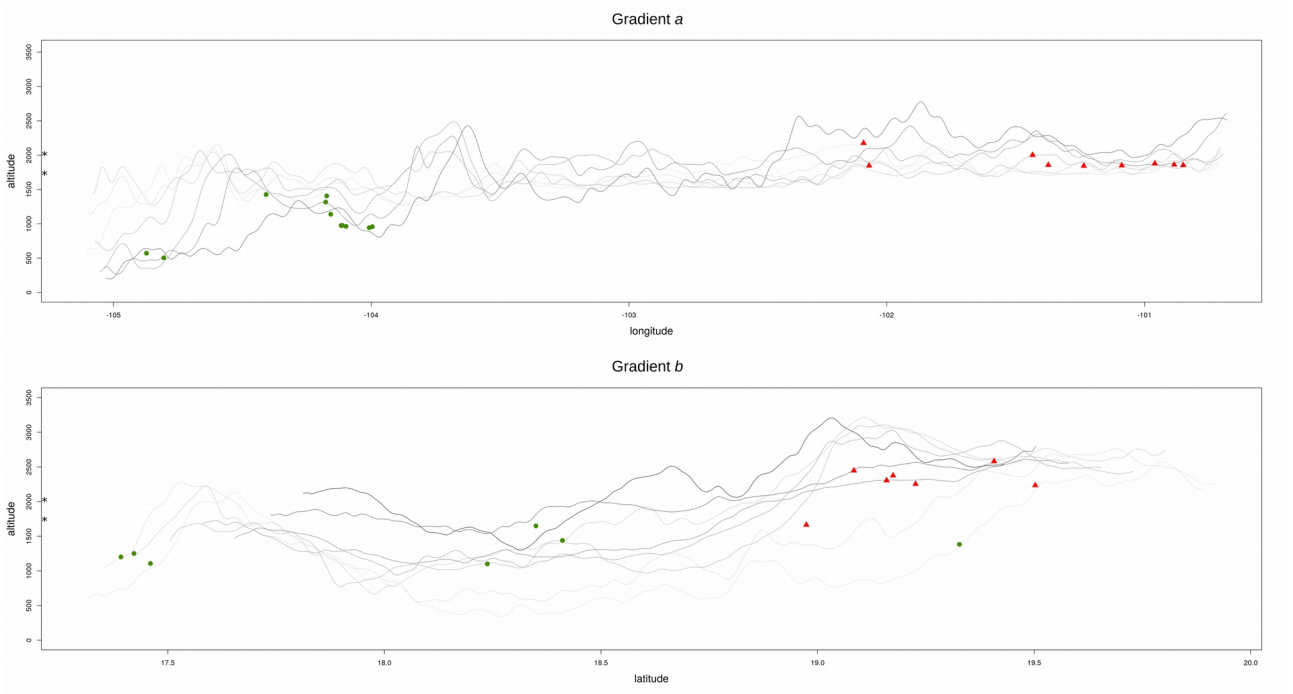


Figure S1: Altitudinal profiles along gradients *a* and *b*. Sampled populations are plotted on parallel altitudinal profiles for gradients *a* and *b*. Darker gray lines indicate lower latitude for gradient *a* and lower longitude for gradient *b*. Sampled populations are plotted by green circles (*parviglumis*) or red triangles (*mexicana*). The altitude of the two experimental fields (CEBAJ: 1750m and SENGUA: 2017m) are marked with stars on y-axes.

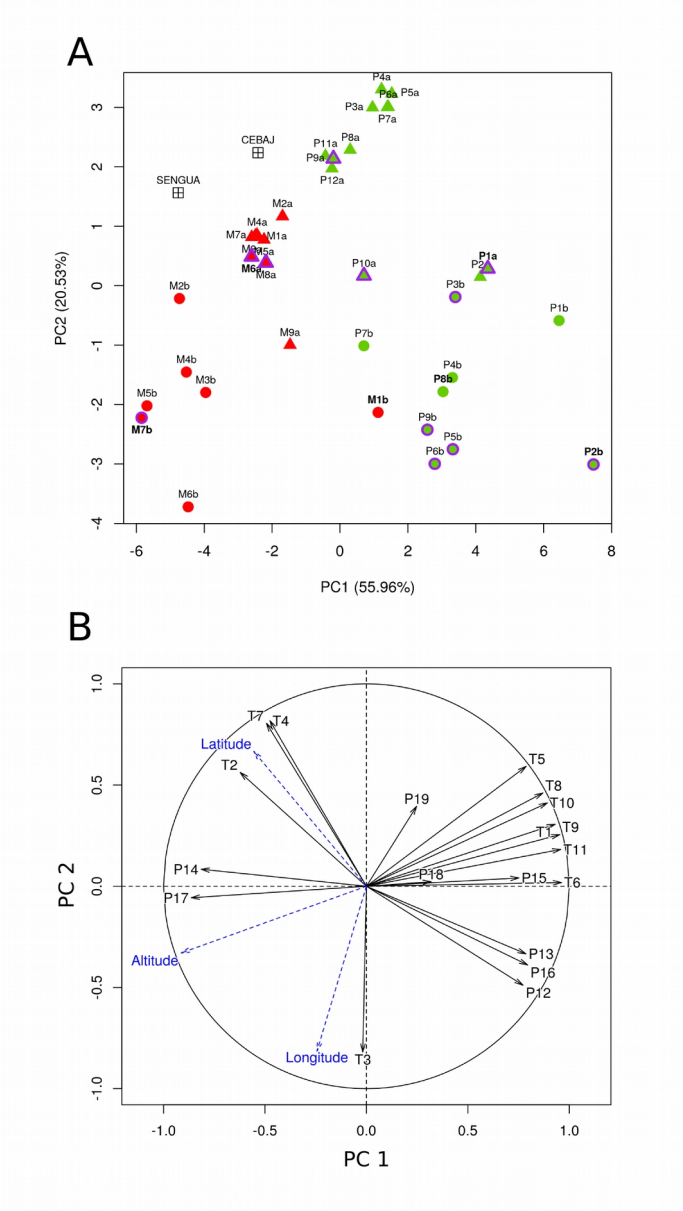


Figure S2: Principal Component Analysis of 19 climate variables for 37 teosinte populations. A: Projection of *parviglumis* (in green) and *mexicana* (in red) populations on the first PCA plane with gradients *a* and *b* indicated by triangles and circles, respectively. The 11 populations evaluated in common gardens are surrounded by a purple outline. Populations that were previously sequenced to detect selection footprints are shown in bold (S1 Table). B: Correlation circle of the 19 climatic variables on the first PCA plane. Climatic variables indicated as Tn (n from 1 to 11) and Pn (n from 12 to 19) are related to temperature and precipitation, respectively. Altitude, Latitude and Longitude (in blue) were added as supplementary variables, and CEBAJ and SENGUA field locations were added as supplementary individuals.

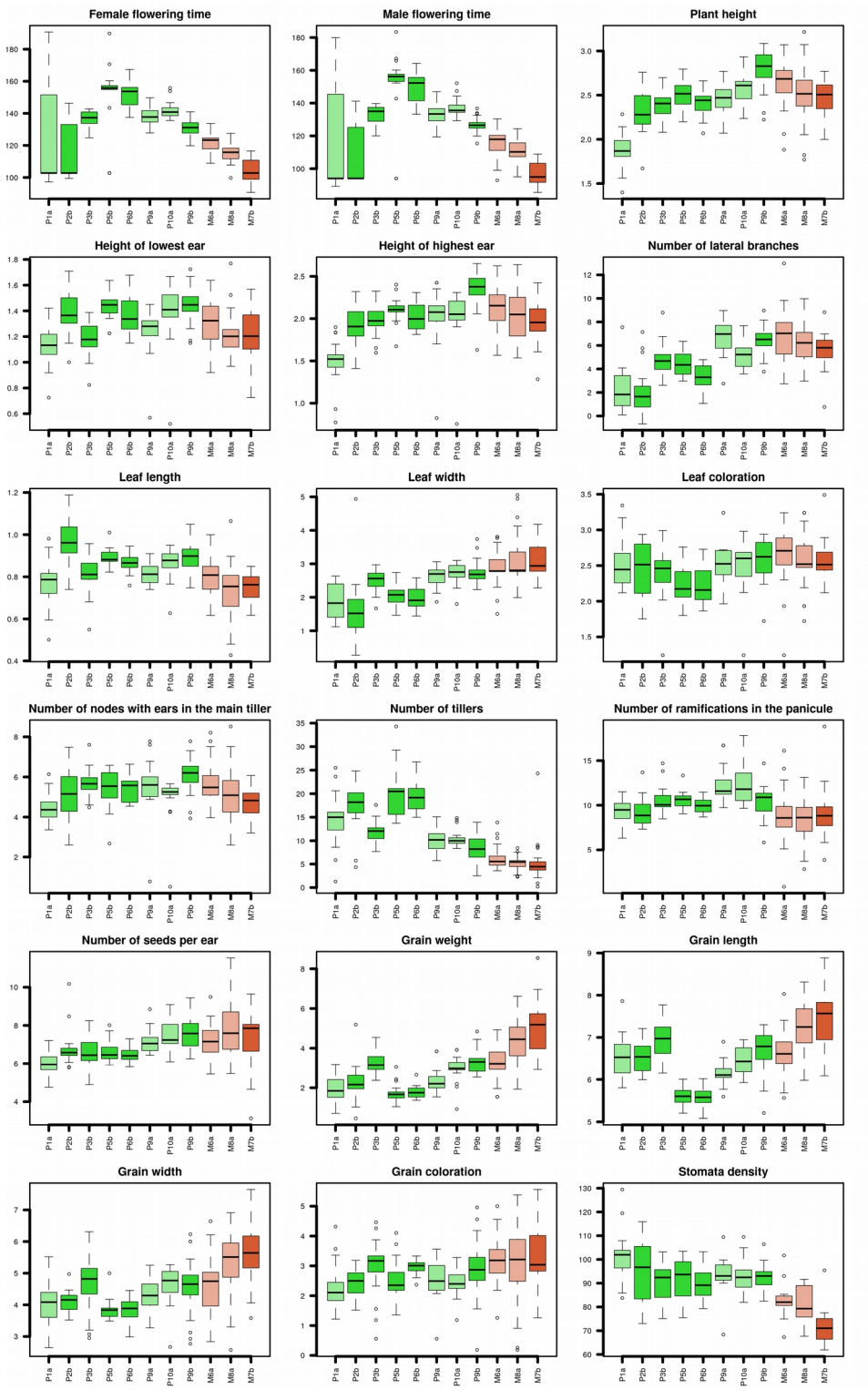


Figure S3: Box-plots of means adjusted by field, year and block, for all traits. Populations are ranked by altitude. *parviglumis* populations are shown in green and *mexicana* in red. Lighter colors are used for gradient 'a' and darker colors for gradient 'b'. Units of measurement correspond to those defined in S2 Table. For male and female flowering time, we report values for all 11 populations although very few individuals from the two most lowland populations (P1a and P2b) flowered. Covariation with altitude was significant for all traits except for the number of nodes with ears on the main tiller (S3 Table).

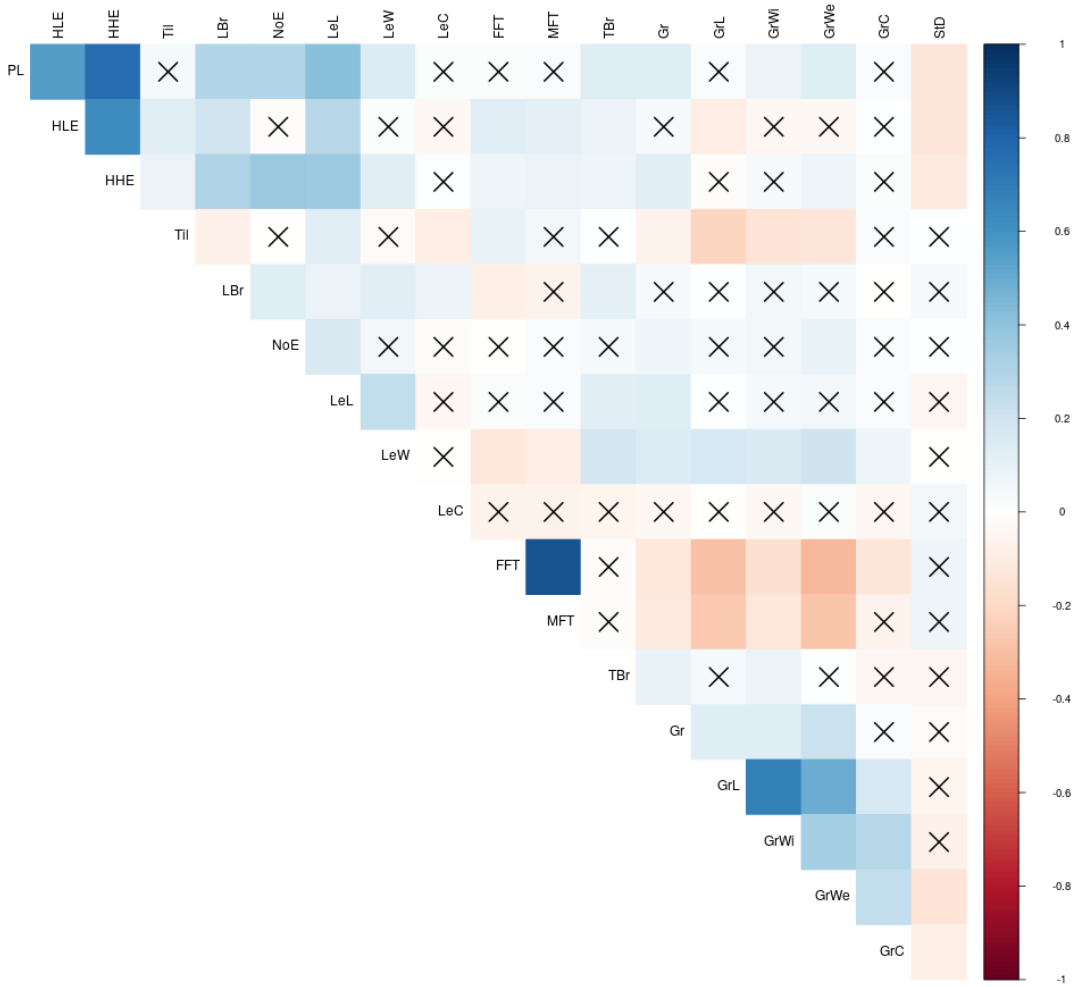


Figure S4: Pairwise correlations between phenotypic traits. Pearson coefficient sign and magnitude for significant correlations between phenotypic traits after correction for experiment design (Model M1'). X: correlations that are not significant.

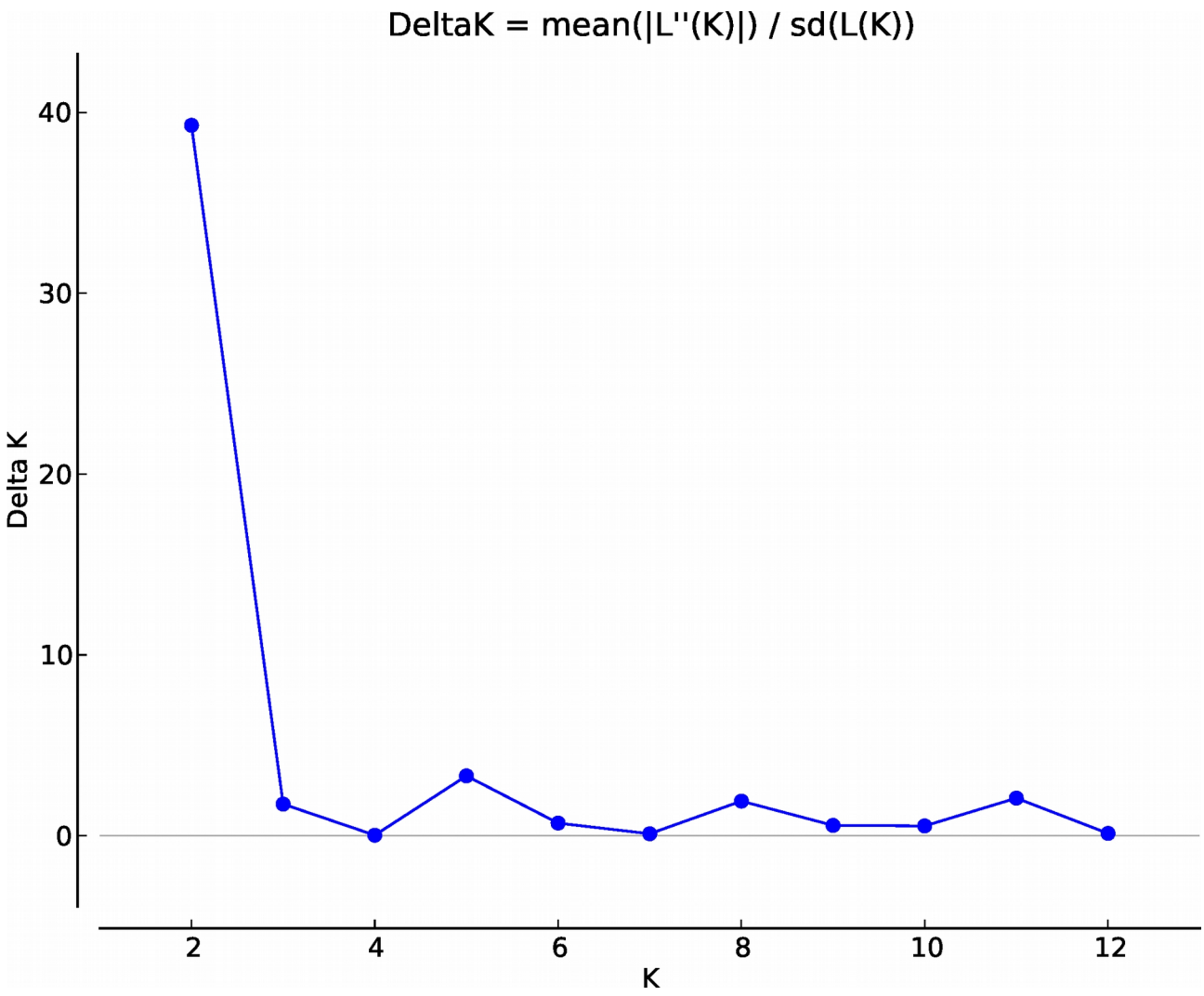


Figure S5. Evanno method calculations for population number ΔK in the association panel genotyped for 38 SSRs.

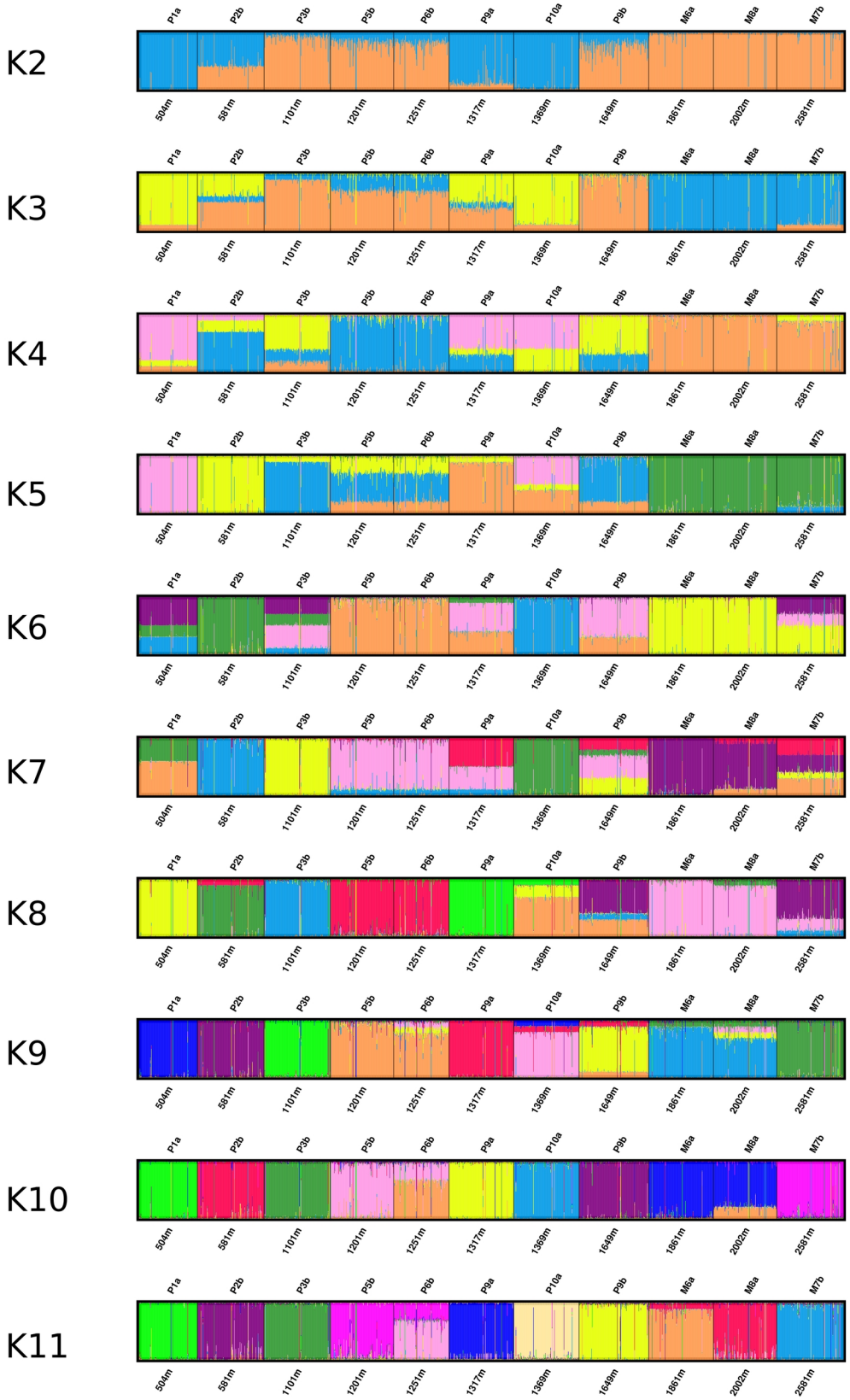


Figure S6. Genetic clustering of ancestry proportions in the association panel genotyped for 38 SSRs. Genetic clustering was computed for *K*=2 to *K*=11. Vertical lines (individuals) are partitioned into coloured segments whose length represents the admixture proportions from the K clusters.

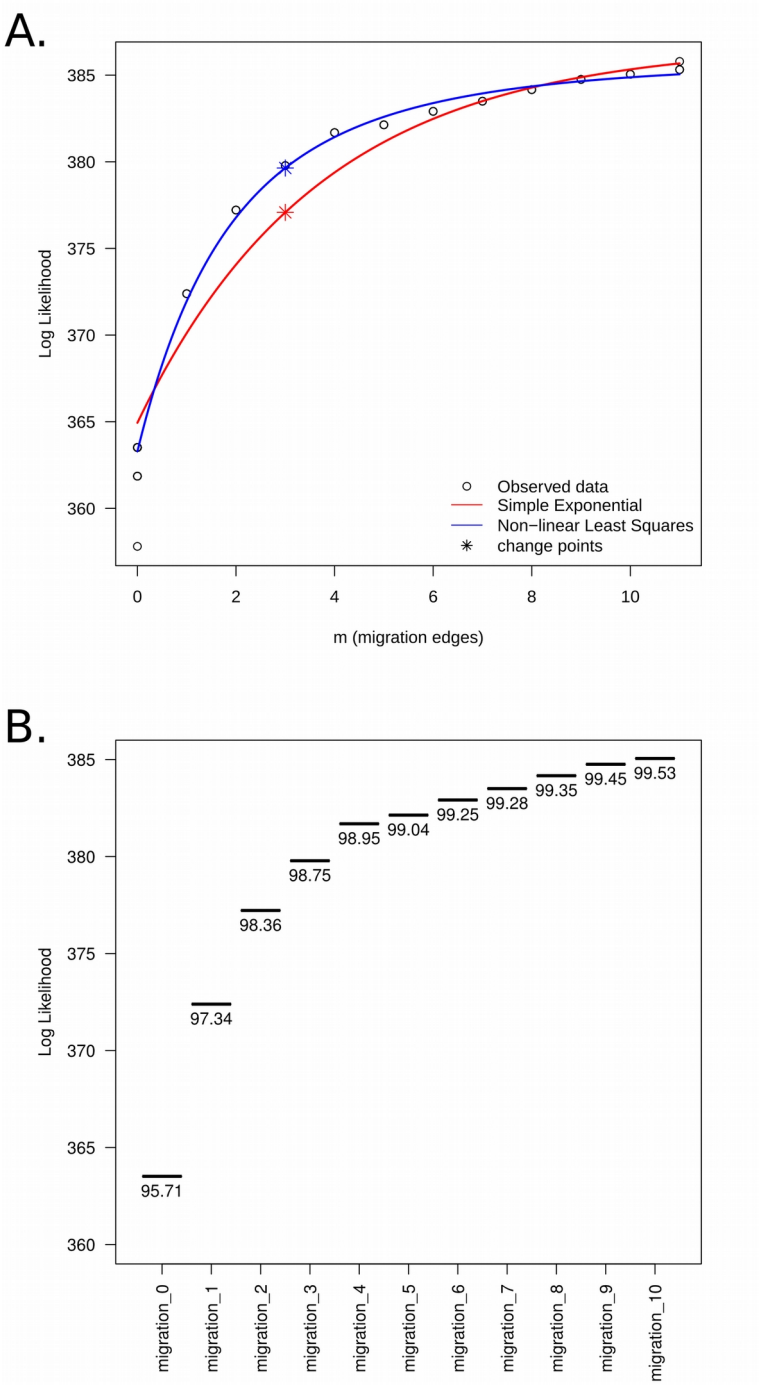


Figure S7. Determination of the migration edge number in the TreeMix model. Observed Log likelihood values are plotted against the number of migration edges tested from 0 to 10, and two models are fitted to the data (A). Both the simple exponential and the non-linear least squares delivered an optimal value of 3 for the number of migration edges (change points). The model with 3 migration edges explained 98.75% of the variance, a substantial increase from the null model with no migration edge which is 95.7% (B).

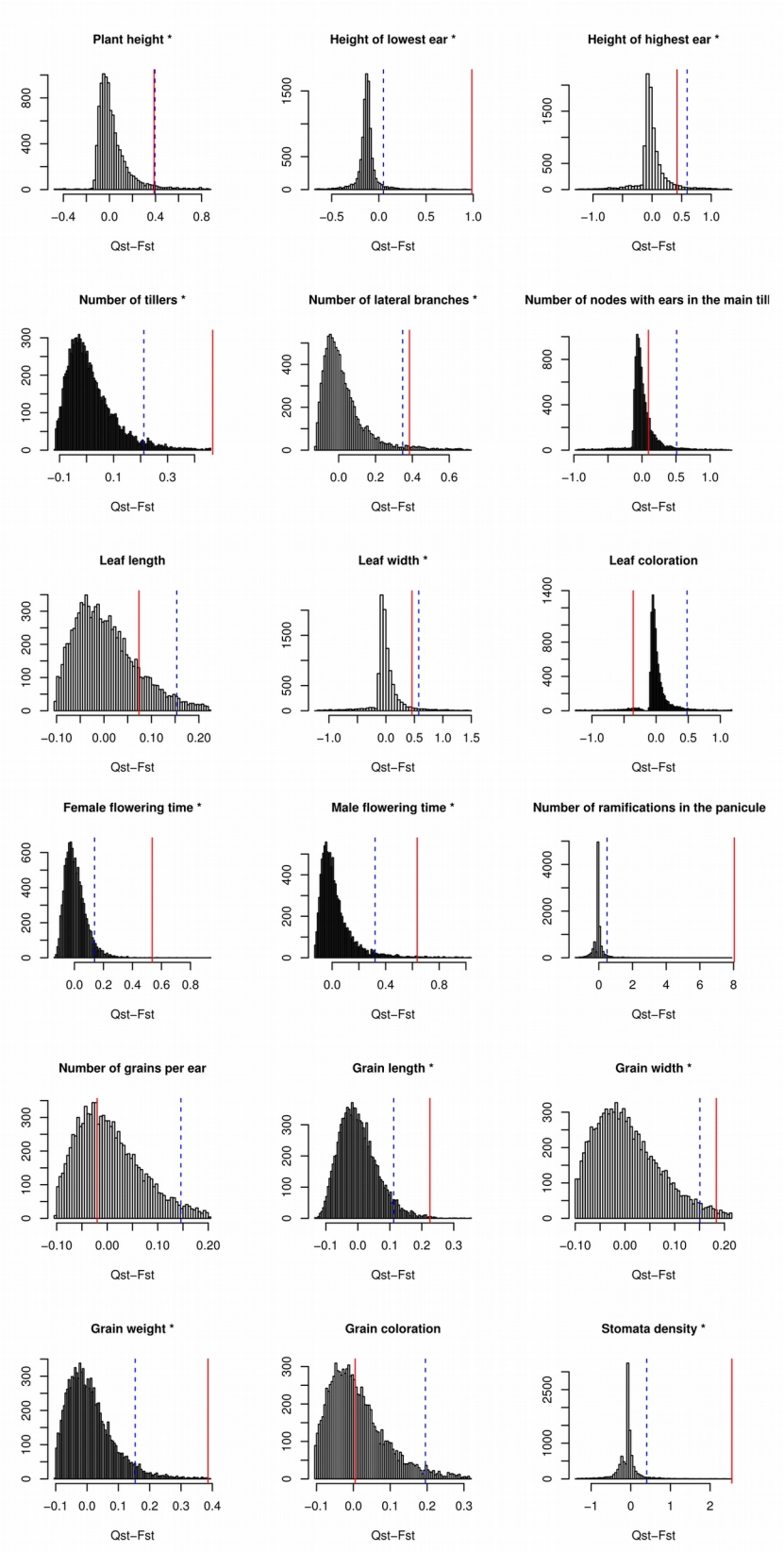


Figure S8: Significance of Q_{ST}-F_{ST} difference for each trait. The dotted blue line indicates the 95% threshold of the simulated distributions and the red line refers to the observed difference. In this analysis, we considered as spatially-varying traits those for which the observed difference fell outside the 95% threshold. Note that Plant height was borderline significant. *: Set of traits detected by DRIFTSEL.

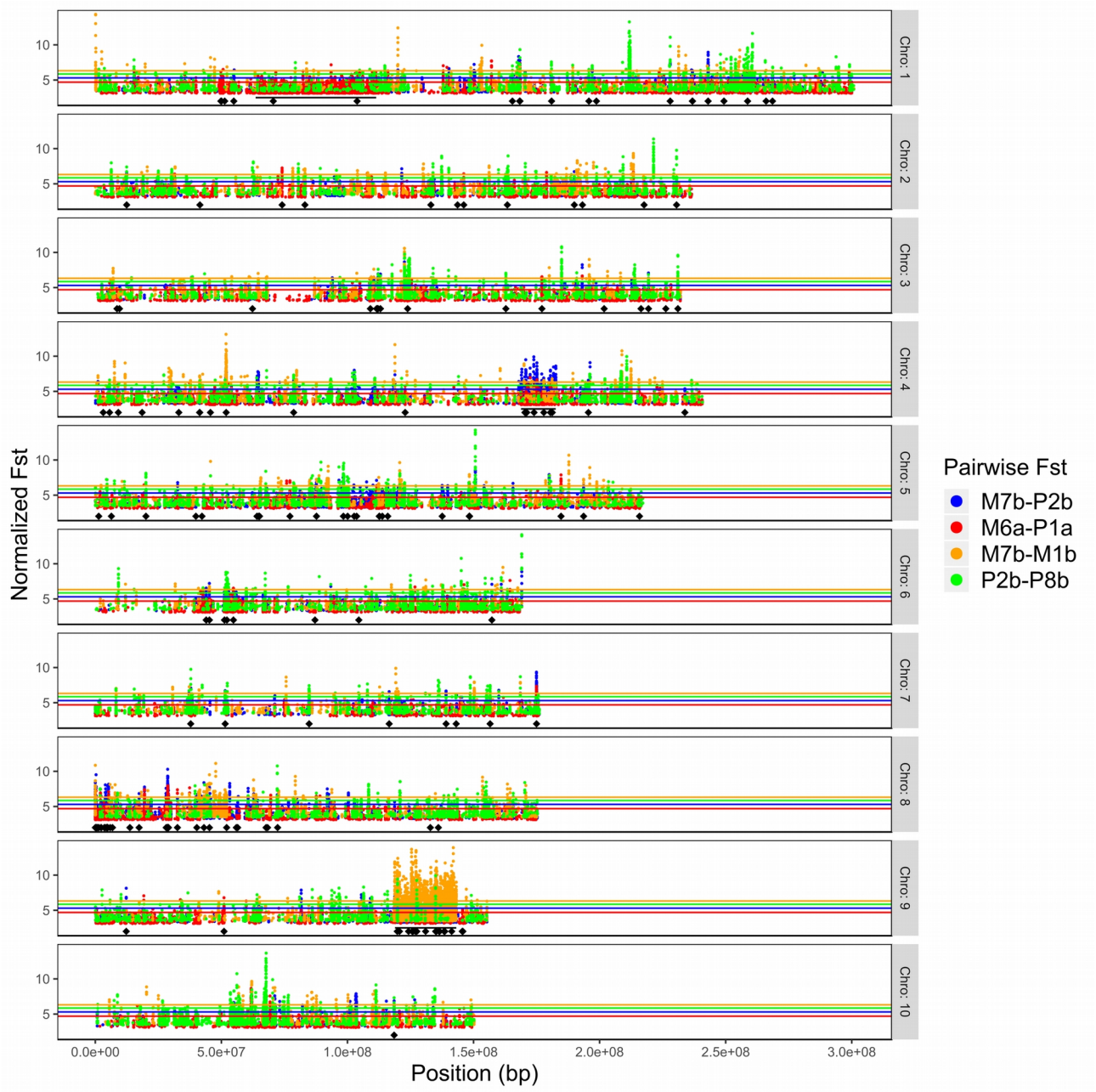


Figure S9: Genomic F_{ST}-scans on 6 teosinte populations. We computed 4 pairwise-F_{ST} values from 6 populations previously sequenced (S1 Table). Those include F_{ST} between lowland and highland populations of each gradient (P1a-M6a, P2b-M7b) as well as within subspecies on gradient *b* (P2b-P8b, M1b-M7b). F_{ST} values are averaged across sliding windows of 20 SNPs with a step of five SNPs (from top to bottom, chromosome 1 to 10) and normalized by subtracting the F_{ST} mean and dividing by the standard deviation across pairwise comparisons. Only the top 1% values are represented. The 1% thresholds for each pairwise comparisons are indicated by colored horizontal lines. Horizontal black bars indicate location of inversions on chromosome 1 (*Inv1n*), chromosome 4 (*Inv4m*) and chromosome 9 (*Inv9d*). The subset of 171 outlier SNPs analyzed in the present study is indicated with black diamond marks along the X axes.

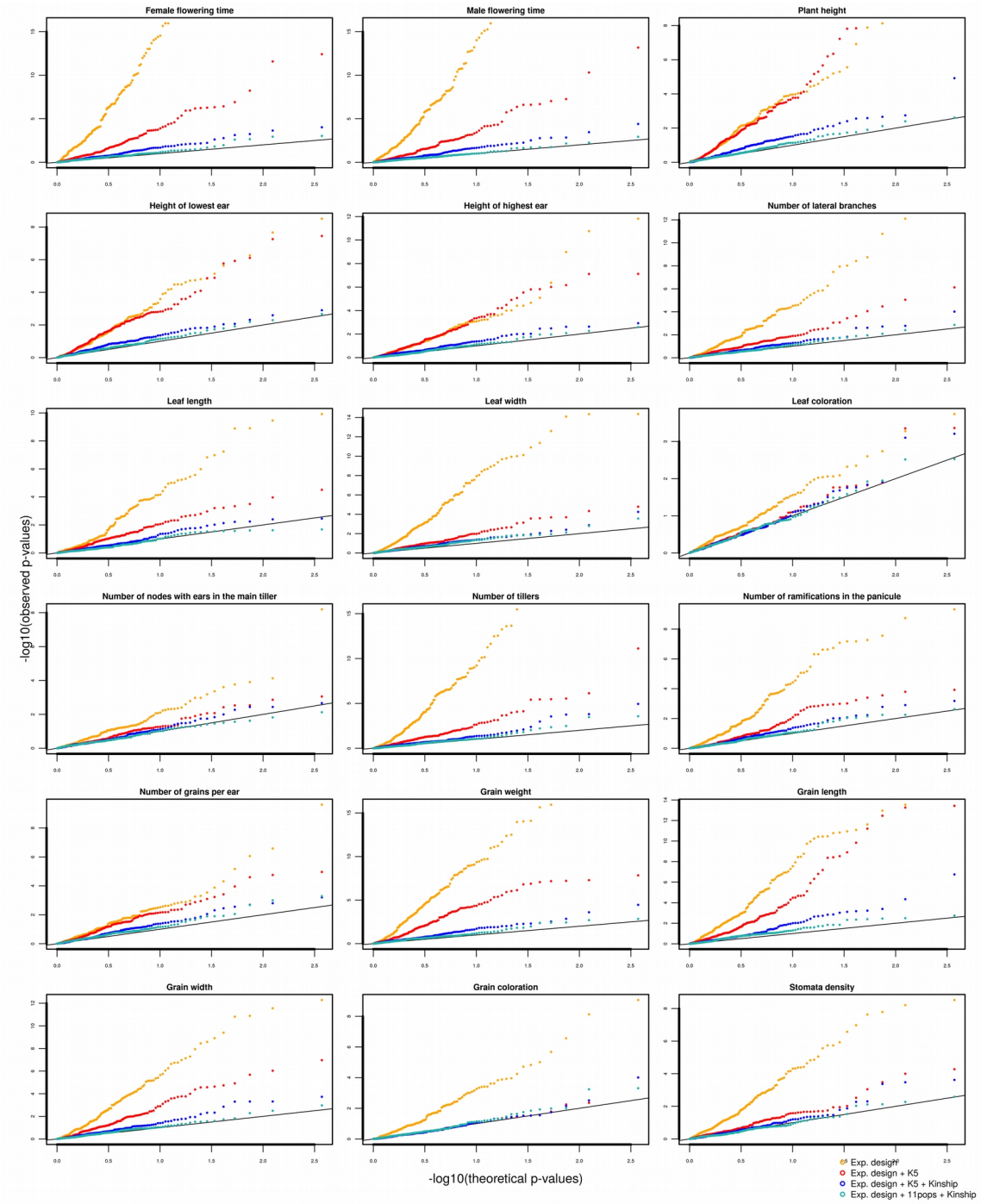


Figure S10: QQ-plots of observed P-values and expected P-values generated from 38 SSRs. We employed three versions of the model M5 with correction for neither structure nor kinship, with correction for genetic structure (at *K*=5), with correction for genetic structure (at *K*=5 and with 11 populations) and kinship.

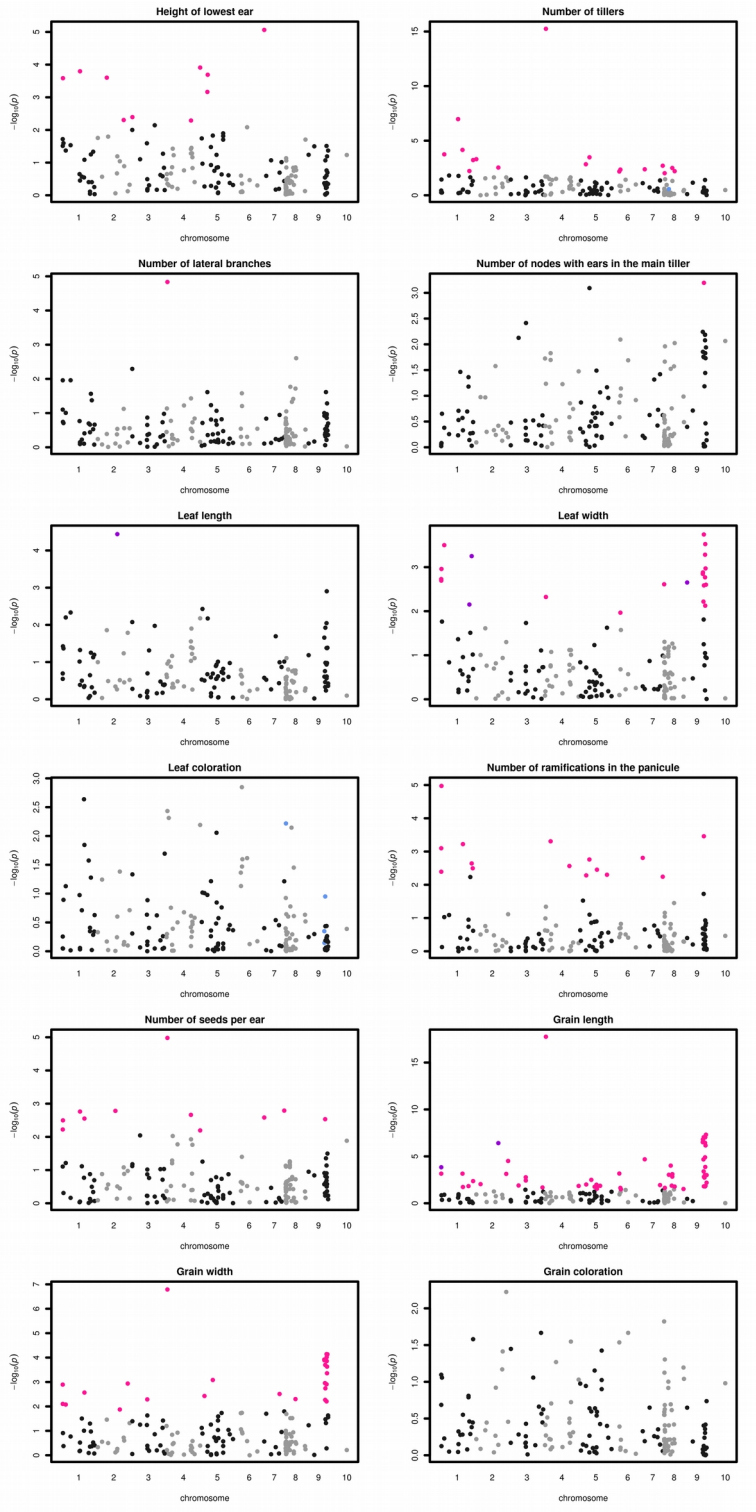


Figure S11: Manhattan plots of associations between 171 outlier SNPs and 12 phenotypic traits. X-axis indicates the positions of outlier SNPs on chromosomes 1 to 10, black and gray colors alternating per chromosome. Plotted on the Y-axis are the negative Log_{10}-transformed *P* values obtained for the *K*=5 model. Significant associations (10% FDR) are indicated considering either a structure matrix at *K*=5 (pink dots), for 11 populations (blue dots), or for both *K*=5 and 11 populations models (purple dots).

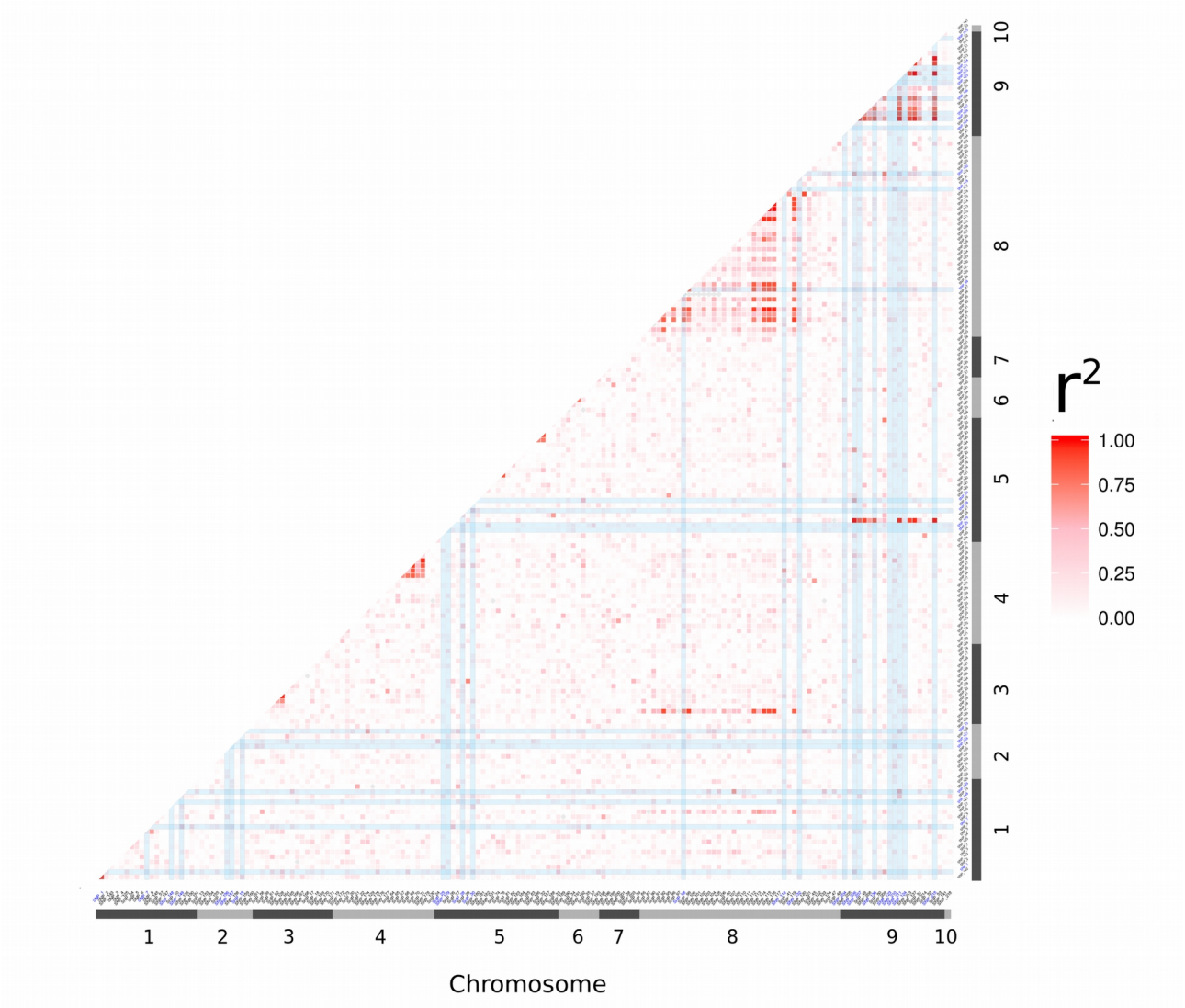


Figure S12: Pairwise Linkage Disequilibrium (LD) between outlier SNPs. Pairwise LD between 171 SNPs was estimated using r^2, and corrected for structure at K=5 and kinship computed from 38 SSRs. Blue shaded bars show the 23 SNPs found to associate with at least one phenotype under the 11 populations structure correction.

II.8 SUPPLEMENTARY TABLES

S1 Table. Description of 37 teosinte populations, and sets of populations used in the present study by data types.

Population code, accession names and corresponding references (indicated in the main text) where they are described are indicated along with the subspecies names, geographical coordinates, locality, state and year of collection.

Colours indicate populations for which different types of data were available (High Throughput Sequencing and Neutral SNP genotyping) or produced in our study.

Values for 19 bioclimatic variables are indicated, along with values obtained for the two first principal components of PCA performed on (37) and (28) populations.

Population code	Population name in [2]	Accession names	References	Subspecies	Latitude	Longitude	Altitude	Locality	State	Year of collection	HTS	Neutral SNP genotyping	SSR genotyping	Phenotyping	Candidate SNP genotyping	T1	T2	T3
Gradient *a*																		
P1a	L2	CM13	[2]	*parviglumis*	19.7154	-104.8035	504	Telpitita	Jalisco	2010						23.247	15.810	65.954
P2a		SMH 577	[2]	*parviglumis*	19.7325	-104.8719	572	El_Llanito	Jalisco	2009						23.050	15.728	66.647
P3a		SMH 570	[1]	*parviglumis*	19.9339	-104.0077	944	La_Labor	Jalisco	2009						23.010	15.939	63.314
P4a		MIT14	[2]	*parviglumis*	19.9458	-103.9958	958	San_Lorenzo	Jalisco	2010						23.417	16.197	63.338
P5a		SMH 569	[2]	*parviglumis*	19.9218	-104.0977	963	Los_Naranjos	Jalisco	2009						23.556	16.102	63.356
P6a		SMH 567	[2]	*parviglumis*	19.9130	-104.1169	976	El_Estanco	Jalisco	2009						23.343	15.883	63.348
P7a		MIT 15	[2]	*parviglumis*	19.9129	-104.1123	976	Ejutla	Jalisco	2010						23.328	15.879	63.347
P8a		SMH 566	[2]	*parviglumis*	19.9027	-104.1572	1140	El_Estanco	Jalisco	2009						21.811	15.548	63.303
P9a		SMH 565	[1]	*parviglumis*	19.8961	-104.1768	1317	Ejutla	Jalisco	2009						20.946	15.813	63.354
P10a		SMH578	[1]	*parviglumis*	19.5354	-104.0583	1369	El Rodeo	Jalisco	2009						19.485	14.482	63.089
P11a		SMH 564	[1]	*parviglumis*	19.9100	-104.1726	1407	Ejutla	Jalisco	2009						20.724	15.977	63.387
P12a		CM12	[1]	*parviglumis*	20.6273	-104.4079	1426	Guachinango	Jalisco	2010						19.249	17.251	63.611
M1a		SMH 580	[2]	*mexicana*	19.9646	-101.2354	1844	Capacho	Michoacán	2010						17.980	15.866	64.116
M2a		CM11	[2]	*mexicana*	20.1394	-102.0684	1846	Churintzio	Michoacán	2010						17.848	16.198	62.964
M3a		SMH575	[1]	*mexicana*	20.0532	-101.0881	1849	San_Rafael	Michoacán	2009						18.010	15.660	64.055
M4a		SMH 572	[2]	*mexicana*	19.9582	-100.8497	1854	Andocutin	Guanajuato	2009						17.950	15.916	63.870
M5a		SMH576	[1]	*mexicana*	20.1607	-101.3731	1856	Tejocote_ de_Clalera	Guanajuato	2009						18.277	15.868	64.828
M6a	H2	SMH571	[1]	*mexicana*	19.9917	-100.8849	1861	San_Jose_ de_las_Pilas	Guanajuato	2009						17.331	15.516	63.419
M7a		SMH 573	[2]	*mexicana*	19.9828	-100.9599	1878	Puerto_ de_Cabras	Guanajuato	2009						17.898	15.731	63.676
M8a		SMH579	[2]	*mexicana*	20.1338	-101.4342	2002	Armadillo	Michoacán	2010						17.999	15.830	65.335
M9a		SMH-MGCH 582	[3]	*mexicana*	20.7195	-102.0898	2176	Jesús_María	Jalisco	2011						16.014	13.839	61.903

Gradient *b*

P1b		CM03	[1]	*parviglumis*	17.1719	-99.5415	343	Tierra Colorada	Guerrero	2010					26.776	13.997	67.688
P2b	L1	CM04	[1]	*parviglumis*	16.9811	-99.2855	581	Tecoanapa	Guerrero	2010					24.805	12.891	68.463
P3b		CM08	[1]	*parviglumis*	18.2377	-99.2180	1101	Paso_Morelos	Guerrero	2010					24.002	15.131	69.466
P4b		CM07	[1]	*parviglumis*	17.4596	-99.3683	1107	Mochitlán	Guerrero	2010					23.135	13.765	68.578
P5b		CM05	[1]	*parviglumis*	17.3918	-99.4776	1201	Chilpancingo	Guerrero	2010					21.882	12.962	68.339
P6b		CM06	[1]	*parviglumis*	17.4219	-99.4507	1251	Mazatlán	Guerrero	2010					21.151	12.788	68.151
P7b		SMH 581	[2]	*parviglumis*	19.3274	-100.4214	1383	Enandio	Michoacán	2011					20.542	14.732	67.570
P8b	IP1	CM02	[1]	*parviglumis*	18.4110	-99.9084	1439	Alcholoa	Guerrero	2010					22.181	13.167	68.141
P9b		CM01	[1]	*parviglumis*	18.3498	-99.8411	1649	Teloloapan	Guerrero	2010					21.330	12.595	67.962
M1b	IM1	CM09	[1]	*mexicana*	18.9741	-99.0703	1669	Huilotepec	Morelos	2010					20.276	14.251	69.252
M2b		Texcoco_ Texcoco	[3]	*mexicana*	19.5025	-98.9146	2234	Texcoco	México	2011					15.711	17.078	67.658
M3b		Cocotitlan_ Cocotitlan	[3]	*mexicana*	19.2262	-98.8661	2252	Cocotitlán	México	2011					15.592	15.463	68.445
M4b		Tenancingo_ Tenancingo	[3]	*mexicana*	19.1595	-98.1854	2306	Tenancingo	Tlaxcala	2011					15.054	17.147	69.916
M5b		SanNicolas_ SnNicolas BuenosAires	[3]	*mexicana*	19.1742	-97.5528	2375	San_Nicolas	Puebla	2011					14.207	17.198	72.221
M6b		Calpan_ Calpan	[3]	*mexicana*	19.0839	-98.4874	2447	Calpan	Puebla	2011					13.853	15.036	71.170
M7b	H1	CM10	[1]	*mexicana*	19.4075	-99.6271	2581	VillaSeca	Mexico	2010					13.410	16.331	69.255
Experimental fields																	
CEBAJ		-	-	-	20.5222	-100.8122	1750	CEBAJ	Guanajuato	-	-	-	-	-	18.739	16.565	63.240
SENGUA		-	-	-	21.2986	-100.5164	2017	SENGUA	Guanajuato	-	-	-	-	-	16.810	15.904	63.307

T1=BIO1, Annual Mean Temperature; T2=BIO2, Mean Diurnal Range; T3=BIO3, Isothermality (BIO2/BIO7); T4=BIO4, Temperature Seasonality; T5=BIO5, Max Temperature of Warmest Month; T6=BIO6, Min Temperature of Coldest Month; T7=BIO7, Temperature Annual Range (BIO5–BIO6); T8=BIO8, Mean Temperature of Wettest Quarter; T9=BIO9, Mean Temperature of Driest Quarter; T10=BIO10, Mean Temperature of Warmest Quarter; P11= BIO11, Mean Temperature of Coldest Quarter

P12=BIO12, Annual Precipitation; P13=BIO13, Precipitation of Wettest Month; P14=BIO14, Precipitation of Driest Month; P15=BIO15, Precipitation Seasonality (coefficient of variation); P16=BIO16, Precipitation of Wettest Quarter; P17=BIO17, Precipitation of Driest Quarter; P18=BIO18, Precipitation of Warmest Quarter; P19=BIO19, Precipitation of Coldest Quarter.

[1] : C. Muñoz-Díez, et al., New Phytol. 199 (2013) 264-276, [2] : M.A. Fustier, et al., Mol Ecol. 26 (2017) 2738–56, [3] J.A. Aguirre-Liguori, et al., Mol Ecol. 26 (2017) 1-15

T4	T5	T6	T7	T8	T9	T10	T11	P12	P13	P14	P15	P16	P17	P18	P19	PC1 (37)	PC2 (37)	PC1 (28)	PC2 (28)	31 pops
206.367	34.519	10.548	23.971	24.945	23.372	25.321	21.056	1364.045	331.830	3.087	118.971	972.839	13.759	497.377	35.047	4.355	0.285	4.820	1.384	1
205.644	34.047	10.448	23.599	24.837	23.053	25.075	20.873	1317.697	320.519	3.175	118.749	941.777	13.331	473.651	36.495	4.134	0.144	4.606	1.140	1
233.151	35.414	10.239	25.175	24.918	21.808	25.440	20.349	675.557	156.045	4.347	101.809	416.551	17.085	310.210	38.057	0.959	2.996	-	-	1
238.086	35.995	10.423	25.572	25.400	22.155	25.907	20.714	682.849	157.637	4.005	102.410	420.661	16.494	311.964	37.729	1.224	3.302	1.853	3.053	1
238.983	35.986	10.571	25.415	25.609	22.205	26.057	20.876	738.427	171.608	4.113	104.142	455.022	16.601	329.420	40.176	1.527	3.230	-	-	1
235.942	35.592	10.519	25.073	25.361	22.013	25.804	20.693	740.045	172.952	4.439	104.411	458.095	16.914	330.039	40.506	1.423	3.006	-	-	1
235.618	35.579	10.513	25.066	25.337	22.006	25.784	20.679	734.593	171.479	4.420	104.189	454.445	16.909	328.134	40.324	1.396	3.003	-	-	1
227.316	33.852	9.291	24.561	23.722	20.600	24.188	19.219	760.502	176.674	5.821	101.833	466.997	20.206	338.783	44.646	0.299	2.285	-	-	1
229.865	33.176	8.216	24.960	22.937	19.697	23.374	18.319	815.198	188.903	6.098	101.617	497.678	21.541	360.050	48.529	-0.193	2.137	0.495	2.279	1
192.854	30.819	7.864	22.955	20.727	18.389	21.498	17.262	1012.019	233.490	4.218	105.811	644.019	17.694	414.542	54.944	0.706	0.167	1.303	0.646	1
231.614	33.094	7.889	25.205	22.741	19.475	23.181	18.074	819.144	188.908	6.311	101.025	498.918	22.088	363.424	49.159	-0.428	2.182	-	-	1
257.535	32.691	5.572	27.119	21.684	17.819	21.974	16.267	940.305	242.776	4.410	114.279	633.399	17.316	469.991	35.890	-0.236	1.973	0.278	2.985	1
238.090	30.449	5.703	24.746	19.831	15.981	20.564	15.164	727.508	170.266	6.853	104.511	451.995	22.821	332.009	28.529	-2.426	0.822	-1.825	1.308	1
253.165	30.741	5.014	25.727	19.835	17.193	20.636	14.895	789.763	201.543	5.299	111.429	515.254	20.061	375.788	25.580	-1.692	1.169	-1.176	2.008	1
241.115	30.412	5.964	24.448	19.255	15.981	20.630	15.150	680.928	159.657	7.077	104.556	425.960	23.081	307.942	27.899	-2.601	0.816	-1.996	1.238	1
242.842	30.408	5.489	24.919	19.305	15.800	20.576	15.068	728.776	170.326	6.809	105.900	461.866	21.694	326.980	26.795	-2.442	0.855	-1.861	1.411	1
238.332	30.636	6.158	24.478	20.099	16.284	20.876	15.466	696.795	171.272	6.218	106.476	442.283	22.717	163.772	28.591	-2.237	0.774	-1.625	1.099	1
241.546	29.657	5.192	24.465	18.649	15.222	19.954	14.458	740.553	177.345	6.820	107.040	471.032	22.012	336.753	27.127	-2.606	0.488	-2.037	1.151	1
242.469	30.353	5.648	24.705	19.199	15.808	20.529	15.018	694.791	163.874	6.795	106.056	441.665	21.462	312.334	26.043	-2.493	0.839	-1.915	1.348	1
230.434	30.246	6.017	24.230	19.717	16.095	20.540	15.280	717.861	180.345	6.007	107.647	459.722	22.509	168.658	28.450	-2.181	0.386	-1.585	0.764	1
220.442	27.626	5.269	22.357	17.647	15.523	18.549	13.511	811.587	226.200	4.146	115.996	551.026	17.622	180.845	27.229	-1.470	-0.997	-1.003	-0.122	

135.934	36.709	16.031	20.679	27.126	26.227	28.136	25.267	1222.486	280.635	2.278	115.663	779.730	11.176	522.656	22.489	6.455	-0.587	-	-	1
115.154	33.917	15.088	18.829	25.147	24.024	25.861	23.635	1735.022	422.632	2.674	115.973	1114.778	10.844	694.853	42.817	7.469	-3.008	7.896	-1.927	1
178.309	35.208	13.426	21.782	24.673	22.919	26.192	21.943	920.151	192.702	1.804	110.779	574.864	8.827	72.578	15.108	3.398	-0.191	3.858	-0.500	1
144.225	32.906	12.834	20.072	23.616	21.960	24.555	21.493	1128.321	239.692	3.340	108.679	709.021	17.080	265.037	33.678	3.313	-1.546	3.860	-1.518	1
135.740	31.051	12.084	18.967	22.415	20.717	23.228	20.346	1267.627	284.819	3.947	111.407	811.513	19.625	536.144	35.236	3.324	-2.752	3.825	-2.163	1
133.928	30.300	11.536	18.764	21.657	20.021	22.479	19.621	1229.858	276.068	4.222	110.888	785.210	19.606	524.994	34.068	2.793	-2.997	-	-	1
158.959	32.030	10.228	21.803	21.107	19.459	22.526	18.846	895.311	204.098	4.890	105.438	546.245	17.529	225.392	26.828	0.704	-1.013	1.256	-0.983	1
157.987	32.563	13.240	19.323	22.408	21.511	24.407	20.528	998.874	214.882	2.846	111.201	629.606	10.775	65.951	22.137	3.037	-1.783	3.515	-1.897	1
154.742	31.216	12.684	18.532	21.635	20.619	23.469	19.660	1057.233	228.399	3.695	108.988	662.900	12.845	79.477	26.113	2.575	-2.423	3.084	-2.428	1
162.361	30.976	10.398	20.578	20.390	19.284	22.285	18.458	1091.934	224.316	4.651	108.222	663.638	16.490	81.051	26.633	1.122	-2.132	-	-	1
221.025	27.699	2.457	25.242	17.454	13.056	17.924	13.056	583.956	114.476	5.140	88.296	327.673	20.502	270.714	20.502	-4.728	-0.216	-4.160	-0.203	
186.291	26.610	4.019	22.591	16.882	14.136	17.539	13.353	701.641	137.246	5.628	92.350	396.270	23.216	210.707	24.878	-3.956	-1.797	-3.368	-1.770	
197.759	26.802	2.277	24.525	16.588	13.318	17.026	12.749	825.896	161.684	6.605	92.057	452.554	24.410	382.989	27.195	-4.525	-1.453	-3.936	-1.065	
185.809	25.456	1.642	23.813	15.575	11.908	15.982	11.908	576.431	109.843	4.300	80.268	276.876	22.887	269.256	22.887	-5.686	-2.021	-5.113	-2.301	
161.224	24.050	2.923	21.127	15.017	12.552	15.509	11.981	897.474	171.495	6.998	94.649	505.847	26.777	276.808	31.306	-4.472	-3.721	-3.870	-3.356	
194.045	24.546	0.964	23.582	15.003	11.632	15.359	11.083	824.212	171.904	8.955	94.272	481.704	31.521	382.459	34.073	-5.842	-2.222	-5.180	-1.580	1
281.503	31.765	5.571	26.194	21.180	16.177	21.795	15.462	635.817	146.802	6.085	103.511	394.008	19.183	293.069	22.636					
263.913	29.601	4.480	25.121	18.987	14.586	19.618	13.709	427.071	80.242	6.559	82.624	225.658	21.890	193.210	24.045					

S2 Table. List of the 18 phenotypic traits measured and estimates of narrow-sense heritabilities (h2, with standard deviation s.d).

Mother plant data were available only for 3 grain-related traits.

Trait	Description	Mean h2 (s.d)	Mean h2 (s.d) w/ mother plant data	Mean h2 (s.d.) by population										
				P1a	P9a	P10a	M6a	M8a	P2b	P3b	P5b	P6b	P9b	M7b
Plant architecture														
PLPlant Height	Length of the main tiller from the stem base to the tip of the primary tassel (m)	0.335 (0.220)	NA	0.2 (0.263)	0.278 (0.266)	0.198 (0.214)	0.28 (0.257)	0.739 (0.253)	0.216 (0.258)	0.12 (0.132)	0.307 (0.292)	0.206 (0.255)	0.779 (0.209)	0.367 (0.264)
HLE Height of the Lowest Ear	Length of the primary tiller from the stem base to the lowest ear insertion (m)	0.310 (0.220)	NA	0.111 (0.168)	0.161 (0.186)	0.126 (0.164)	0.252 (0.28)	0.774 (0.225)	0.155 (0.203)	0.696 (0.262)	0.354 (0.319)	0.377 (0.319)	0.455 (0.312)	0.813 (0.174)
HHE Height of the Highest Ear	Length of the primary tiller from the stem base to the highest ear insertion (m)	0.389 (0.265)	NA	0.119 (0.191)	0.139 (0.158)	0.629 (0.275)	0.182 (0.202)	0.136 (0.161)	0.358 (0.318)	0.236 (0.253)	0.146 (0.182)	0.484 (0.329)	0.23 (0.263)	0.757 (0.211)
Til number of Tillers	Number of tillers	0.362 (0.334)	NA	0.014 (0.051)	0.19 (0.257)	0.016 (0.081)	0.783 (0.321)	0.484 (0.385)	0.288 (0.361)	0.182 (0.313)	0.215 (0.304)	0.071 (0.19)	0.876 (0.17)	0.859 (0.253)
LBr number of Lateral Branches	Number of lateral branches on the primary tiller	0.427 (0.326)	NA	0.685 (0.319)	0.548 (0.392)	0.368 (0.34)	0.828 (0.225)	0.11 (0.194)	0.815 (0.279)	0.085 (0.161)	0.315 (0.318)	0.829 (0.233)	0.025 (0.052)	0.093 (0.186)
NoE number of Nodes with Ears	Number of nodes with ears on the primary tiller	0.207 (0.167)	NA	0.054 (0.092)	0.063 (0.14)	0.037 (0.073)	0.206 (0.32)	0.182 (0.218)	0.146 (0.257)	0.098 (0.19)	0.383 (0.384)	0.179 (0.278)	0.582 (0.362)	0.344 (0.288)
Leaf features														
LeL Leaf Length	Length of one intermediate leaf on the primary tiller (cm)	0.282 (0.151)	NA	0.206 (0.209)	0.24 (0.204)	0.369 (0.229)	0.648 (0.23)	0.442 (0.252)	0.136 (0.184)	0.183 (0.208)	0.168 (0.176)	0.21 (0.183)	0.226 (0.2)	0.274 (0.214)
LeW Leaf Width	Width of one intermediate leaf on the primary tiller (cm)	0.192 (0.158)	NA	0.27 (0.364)	0.124 (0.202)	0.074 (0.129)	0.588 (0.368)	0.324 (0.329)	0.082 (0.203)	0.047 (0.091)	0.18 (0.219)	0.215 (0.244)	0.127 (0.195)	0.083 (0.135)
LeC Leaf Colour	Colour of leaves on the whole plant on a qualitative scale (1-4)	0.154 (0.109)	NA	0.416 (0.383)	0.148 (0.229)	0.121 (0.2)	0.073 (0.16)	0.1 (0.129)	0.272 (0.366)	0.089 (0.12)	0.121 (0.196)	0.218 (0.277)	0.087 (0.126)	0.047 (0.088)
Reproduction														
FFT Female Flowering Time	Number of days from field planting* to first visible silks	0.664 (0.254)	NA	0.819 (0.314)	0.777 (0.263)	0.938 (0.128)	0.338 (0.376)	0.43 (0.354)	0.794 (0.354)	0.14 (0.212)	0.721 (0.383)	0.789 (0.283)	0.651 (0.37)	0.911 (0.17)
MFT Male Flowering Time	Number of days from field planting* to anther dehiscence	0.486 (0.346)	NA	0.456 (0.443)	0.13 (0.164)	0.912 (0.173)	0.714 (0.308)	0.59 (0.41)	0.886 (0.216)	0.076 (0.136)	0.678 (0.37)	0.785 (0.331)	0.01 (0.039)	0.105 (0.216)
TBr Tassel Branching	Number of branches in the tassel of the primary tiller	0.150 (0.162)	NA	0.194 (0.299)	0.026 (0.064)	0.138 (0.252)	0.4 (0.351)	0.479 (0.341)	0.231 (0.316)	0.007 (0.022)	0.029 (0.058)	0.071 (0.149)	0.047 (0.097)	0.024 (0.068)
Grain features														
Gr number of Grains per ear	Number of grains per ear based on 5 ears for teosintes	0.434 (0.313)	NA	0.129 (0.233)	0.026 (0.047)	0.452 (0.332)	0.415 (0.376)	0.525 (0.348)	0.189 (0.339)	0.795 (0.233)	0.898 (0.153)	0.112 (0.187)	0.877 (0.163)	0.352 (0.373)
GrL Grain Length	Average length of grains as measured on 10 mature grains (mm)	0.482 (0.252)	0.631 (0.246)	0.634 (0.356)	0.06 (0.085)	0.809 (0.221)	0.662 (0.282)	0.369 (0.383)	0.546 (0.381)	0.851 (0.175)	0.126 (0.179)	0.402 (0.35)	0.397 (0.322)	0.441 (0.351)
GrWi Grain Width	Average width of grains based as measured on 10 mature grains (mm)	0.207 (0.155)	0.274 (0.160)	0.093 (0.18)	0.127 (0.166)	0.273 (0.283)	0.448 (0.307)	0.345 (0.341)	0.187 (0.242)	0.222 (0.265)	0.022 (0.04)	0.45 (0.361)	0.06 (0.092)	0.051 (0.087)
GrWe Grain Weight	Average grain weight based on 50 mature grains (g)	0.422 (0.272)	0.511 (0.043)	0.322 (0.365)	0.537 (0.305)	0.68 (0.297)	0.778 (0.254)	0.791 (0.238)	0.358 (0.399)	0.133 (0.232)	0.173 (0.247)	0.112 (0.21)	0.118 (0.218)	0.643 (0.295)
GrC Grain Colour	Average colour intensity of mature grains on a qualitative scale (1-6)	0.332 (0.265)	NA	0.613 (0.364)	0.311 (0.292)	0.136 (0.228)	0.162 (0.229)	0.798 (0.242)	0.212 (0.266)	0.094 (0.124)	0.688 (0.315)	0.021 (0.038)	0.455 (0.307)	0.159 (0.263)
Stomata features														
StD Stomata Density	Density of stomata based on 9 image measurements on a single leaf (mm^2)	0.302 (0.276)	NA	0.081 (0.235)	0.038 (0.115)	0.066 (0.167)	0.402 (0.403)	0.072 (0.151)	0.495 (0.459)	0.639 (0.408)	0.183 (0.293)	0.607 (0.388)	0.013 (0.038)	0.731 (0.352)

*field planting performed on 11 day old seedlings

S3 Table. Significance of main effects for each trait as determined by models M1 and M3.
P-values of Wald tests are indicated, in red for P-values<0.05 for fixed effects and their interactions.

Model	Trait	Field	Year	Pop	Year:Field	Pop:Year	Pop:Field
M1	PL	5.31E-99	3.36E-18	4.85E-33	1.20E-23	3.97E-02	5.32E-12
M1	HLE	7.89E-170	1.79E-03	1.94E-22	1.39E-37	3.79E-02	1.64E-05
M1	HHE	9.24E-151	1.11E-09	3.03E-26	1.38E-32	4.18E-02	3.63E-16
M1	Til	2.40E-09	3.25E-05	5.47E-53	5.06E-17	2.59E-02	2.84E-09
M1	LBr	2.82E-02	2.57E-18	3.52E-37	3.42E-23	2.28E-03	8.06E-07
M1	NoE	6.35E-01	1.01E-04	1.97E-10	1.59E-01	1.23E-05	6.19E-09
M1	LeL	2.18E-75	5.64E-11	6.52E-27	2.62E-05	7.03E-18	3.04E-16
M1	LeW	4.91E-01	0.00E+00	1.21E-38	1.95E-06	2.32E-40	2.96E-03
M1	LeC	1.65E-01	1.44E-09	5.09E-08	2.30E-01	3.94E-01	1.00E-02
M1	FFT	6.37E-07	2.76E-33	1.12E-72	4.15E-09	6.82E-17	1.60E-13
M1	MFT	1.14E-02	1.19E-17	1.37E-73	7.98E-01	2.57E-11	7.50E-09
M1	TBr	1.94E-04	2.42E-10	3.02E-21	1.00E+00	4.85E-30	6.15E-03
M1	Gr	7.61E-02	5.18E-07	1.08E-11	5.92E-01	8.10E-04	2.32E-01
M1	GrL	3.83E-01	2.44E-13	9.02E-56	2.56E-03	1.19E-07	1.87E-03
M1	GrWi	5.34E-01	1.34E-03	3.94E-40	4.69E-05	6.87E-05	7.78E-03
M1	GrWe	8.08E-01	1.21E-01	1.81E-56	1.10E-02	4.92E-01	2.03E-18
M1	GrC	1.08E-02	3.66E-24	3.12E-12	6.29E-01	4.17E-02	1.42E-17
M1	StD	2.52E-04	1.02E-23	1.36E-24	4.52E-12	1.08E-02	8.50E-03

Model	Trait	Field	Year	Alt	Year:Field	Alt:Year	Alt:Field
M3	PL	1.37E-95	3.92E-16	3.17E-09	1.64E-26	2.05E-02	6.44E-10
M3	HLE	1.49E-167	1.96E-03	7.22E-03	1.34E-44	1.17E-02	2.19E-07
M3	HHE	6.97E-146	6.39E-09	8.27E-03	1.54E-35	3.23E-01	3.78E-14
M3	Til	7.42E-09	3.07E-05	2.61E-31	8.11E-20	7.43E-01	1.94E-09
M3	LBr	4.67E-02	4.87E-17	2.73E-20	1.20E-24	4.12E-01	8.53E-01
M3	NoE	8.45E-01	5.55E-04	7.80E-01	2.32E-02	8.72E-06	1.00E-06
M3	LeL	1.04E-69	6.83E-10	7.55E-10	1.72E-07	1.93E-05	1.95E-11
M3	LeW	4.74E-01	0.00E+00	6.57E-39	1.40E-07	7.93E-43	3.22E-06
M3	LeC	1.72E-01	3.23E-09	7.48E-05	4.62E-02	9.41E-02	5.56E-01
M3	FFT	3.95E-08	5.37E-30	7.50E-43	2.03E-10	5.78E-05	6.33E-04
M3	MFT	1.66E-03	3.93E-16	7.69E-42	1.76E-01	1.86E-07	1.49E-02
M3	TBr	2.90E-04	1.01E-10	6.04E-04	5.51E-01	2.57E-30	6.59E-05
M3	Gr	9.50E-02	1.02E-06	3.40E-12	6.55E-01	7.26E-01	6.25E-01
M3	GrL	4.53E-01	5.79E-13	3.25E-13	3.68E-03	4.67E-07	1.36E-05
M3	GrWi	4.28E-01	1.57E-03	2.86E-26	9.89E-05	1.92E-05	7.23E-03
M3	GrWe	9.60E-01	1.21E-01	9.15E-38	2.59E-01	9.01E-01	2.06E-18
M3	GrC	8.79E-03	9.98E-25	1.07E-09	6.62E-01	9.07E-01	4.74E-21
M3	StD	4.08E-05	6.29E-25	2.25E-25	2.64E-14	9.68E-05	6.92E-04

S4 Table. Description of 46 SSRs and genotyping success rate.

SSRs were multiplexed by groups of 3 (Trio) except for one (phi046).

Chromosome, Motif, Size range, the number of alleles and the success rate of genotyping among all individuals are given.

SSRs with success rate <40.7% were discarded from analyzes (*).

SSR	Trio	Chromosome	Motif	Size range	#alleles	success rate (%)
phi046	1	3	ACGC	62-66	8	86.3
phi427434	2	2	ACC	123-144	11	74.9
phi112		7	AG	125-163	19	77.6
umc1496		5	GCA	135-164	22	71.2
phi331888	3	5	AAG	124-138	13	87.9
phi029		8	AG/AGCG	145-165	23	80.7
phi031*		6	GTAC	186-228	NA	0.00
phi121*	4	8	CCG	97-106	NA	0.00
phi127		2	AGAC	97-132	14	85.5
phi065		9	CACTT	130-149	6	91.1
phi059	5	10	ACC	115-159	10	74.9
phi109188		5	AAAG	146-177	18	82.4
phi402893		2	AGC	207-243	17	75.1
umc1319	6	10	ACC	113-125	6	94.1
phi308707		1	AGC	114-132	8	84.3
phi064		1	ATCC	73-112	19	75.8
phi084	7	10	GAA	149-160	10	85.7
phi104127		3	ACCG	156-168	11	75.6
phi453121*		3	ACC	207-228	NA	0.00
phi108411*	8	9	AGCT	117-139	NA	0.00
phi034		7	CCT	119-145	16	84.7
phi330507		5	CCG	132-142	6	76.4
phi062	9	10	ACG	150-179	11	90.7
phi024		5	CCT	161-176	12	87.1
phi032		9	AAAG	232-240	13	79.1
phi002	10	1	AACG	71-75	8	91.9
phi089		6	ATGC	85-94	12	92.0
phi014		8	GGC	148-169	10	90.3
phi078*	11	6	AAAG	122-214	NA	0.00
phi116		7	ACTG/ACG	140-175	20	89.5
phi033		9	AAG	234-261	14	88.4
phi109275	12	1	AGCT	120-140	19	63.2
umc1133		6	ATAC	86-103	17	69.8
phi233376		8	CCG	137-157	18	63.0
phi448880	13	9	AAG	171-186	6	93.1
dupssr34		4	TTG	114-176	39	88.2
phi213984*		4	ACC	284-302	NA	0.00
phi102228	14	3	AAGC	123-135	14	85.9
phi072*		4	AAAC	132-266	NA	0.00
phi051		7	AGG	137-147	13	66.8
phi053	15	3	ATAC	167-213	19	95.1
phi115		8	AT/ATAC	290-310	2	97.0
phi227562		1	ACC	307-332	13	82.2
phi335539*	16	1	CCG	91-98	10	40.7
phi308090		4	AGC	210-223	11	82.2
phi389203		6	AGC	301-313	11	92.2

S5 Table. Characteristics, association with phenotypes and correlation with environment of 218 outlier SNPs.

Chromosome, position, NonGenic: NG (>1kb from a gene) or Genic: G (<1kb on the two sides of the gene) status, Position to the nearest gene(s) (inside, 1Kb upstream or downstream), and corresponding Gene Name are indicated, as well as location within inversions. Minor allele frequency (MAF) in the association panel is also shown. Note that SNP_1 to SNP_124 were detected by a combination of methods [53], SNP_125 to SNP_141 were detected on another sample of 49 populations [89], and SNP_142 to SNP_218 were detected solely from Fst-scans. SNPs whose frequencies are correlated to PCenv1 and PCenv2 are shown in the corresponding columns. The subset of 171 SNPs used in the association study are marked in bold. We highlighted those associated with the different traits at 10% FDR threshold. Pink for those associated in the K5 model ("5"), blue for those associated to the 11 population model ("11") and purple for their intersections ("5-11"). SNPs included in the multiple SNP models, with the 11 populations of origin as structure control, are marked in bold black boxes.

SNP	Chr	Position	Gene Name	Status	Gene function (name)	Inv	PCenv1	PCenv2	MAF	PL	HLE	HHE	Til	LBr	NoE	LeL	LeW	LeC	FFT	MFT	TBr	Gr	GrL	GrWi	GrWe	GrC	StD
SNP_1	1	50033864	GRMZM2G124532	G	phytochrome B1 (phyB1)				11.12	5		5					5				5	5	5-11	5			
SNP_2	1	51320026	GRMZM2G084940	G	inverstase alkaline neutral 7 (invan7)				8.12	5	5	5					5				5	5			5		
SNP_3	1	70582465	GRMZM2G524711	G		Inv1n	1	1	35.55	5		5	5				5		5	5				5			
SNP_4	1	103827827	GRMZM5G866843	G	similar to RCD one2 (sro2)	Inv1n	1		37.29	5		5							5								
SNP_5	1	168414611	GRMZM2G018472	G	bHLH-transcription factor 26 (bhlh26)				49.81	5	5	5										5			5		
SNP_6	1	180944170	GRMZM2G098494	G	Na+/H+ antiporter 7 (nhx7)				33.95																		5
SNP_7	1	195761492	GRMZM5G872204	G	Discolored-paralog2 (dsc2)				20.9				5						5-11				5				
SNP_8	1	197386628	GRMZM2G159581	G																							
SNP_9	1	198735122	GRMZM2G055968	G					10.56	5		5									5	5	5	5	5		
SNP_10	1	249318279	GRMZM2G103843	G					12.67												5						
SNP_11	1	252393890		NG																							
SNP_12	1	262457397	GRMZM2G351304	G																							
SNP_13	2	41485594	GRMZM2G414537/GRMZM2G414620	G					18.58														5				
SNP_14	2	143673112	GRMZM2G042897	G				1	49.18																		
SNP_15	2	193191211	GRMZM2G158809	G	phosphatidylethanolamine-binding protein 18 (pebp18)				23.44										5-11	5-11							
SNP_16	3	21318315		NG																							
SNP_17	3	201798588	GRMZM2G171373	G	hexokinase6 (hex6)				27.06	5																	
SNP_18	3	216419176	GRMZM2G429899	G	shrunken2 (sh2)				20.18																		
SNP_19	4	3150731		NG			1		24.59																		
SNP_20	4	18573737	GRMZM2G440208	G	phosphogluconate dehydrogenase 3 (pgd3)				15.32																		
SNP_21	4	33118370	GRMZM2G104907/GRMZM2G104942	G			1		34.19																		
SNP_22	4	41373553	GRMZM2G138060	G	sugary1 (su1)		1		21.08			5							5	5	5				5		5
SNP_23	4	46065852		NG																							
SNP_24	5	1358397	GRMZM2G138676	G					19.91	5		5							5	5							
SNP_25	5	6381066	GRMZM2G141903/GRMZM2G141940	G					23.19	5-11		5-11															
SNP_26	5	39893932	GRMZM2G012958	G					41.69	5	5	5	5														

SNP_27	5	42301877	GRMZM2G035405	G	ARF-transcription factor 18 (arftf18)				9.93	5	5	5										5		5							
SNP_28	5	64116218	GRMZM2G145088	G					13.28	5											11	5									
SNP_29	5	64775301	GRMZM2G060583	G					37.62				5		5																
SNP_30	5	65125729	GRMZM2G011998	G					16.39			5-11																			
SNP_31	5	92182318	GRMZM2G179454	G																											
SNP_32	5	148413102		NG			1		22.68			5																			
SNP_33	5	215830190	GRMZM5G837869	G	cytochrome P450 13 (cyp13)				19.05																						
SNP_34	6	51228486		NG					47.44				5				5							5							
SNP_35	6	87120463	GRMZM2G027995	G	eukaryotic initiation factor4a (Eif4A)				11.72																						
SNP_36	6	88593564	GRMZM2G113244	G																											
SNP_37	6	104503351	GRMZM2G159744	G					7.72																						
SNP_38	7	143113709	GRMZM2G179024	G					47.48																5						
SNP_39	7	174983719	GRMZM2G005732	G			1		40.88				5									5	5								
SNP_40	8	4957045	GRMZM2G475017	G	C3H-transcription factor 323 (c3h23)				48	5														5							
SNP_41	8	40263865	GRMZM2G132903	G	enoyl-CoA hydratase1 (ech1)		1		34.62																						
SNP_42	8	44447408		NG																											
SNP_43	8	47311640		NG																											
SNP_44	8	51741986		NG																											
SNP_45	8	56453845	GRMZM2G001755	G			1		32.83															5							
SNP_46	8	124654224	GRMZM2G058745	G	hexokinase3 (hex3)																										
SNP_47	8	132878592	GRMZM2G056014	G	calcium pump1 (cap1)		1		12.98									5	5					5							
SNP_48	8	136050565	GRMZM2G113078/GRMZM2G088132	G					21.12																						
SNP_49	9	124233766	GRMZM2G171428	G	WRKY-transcription factor 61 (wrky61)	Inv9e	1		6.85	5		5				5		5	5					5	5	5		5			
SNP_50	9	126940455	GRMZM5G806784	G		Inv9e	1	1	22.64							5		5	5	5				5	5			5			
SNP_51	9	136573380	GRMZM2G306482	G		Inv9e	1	1	10.29			5				5		5	5					5	5	5		5			
SNP_52	9	138613085	GRMZM2G095727	G		Inv9e	1	1	8.1	5		5				5		5	5					5	5	5		5			
SNP_53	9	145528034	GRMZM2G528283	G				1	23.5	5								5	5					5				5			
SNP_54	1	168286704		NG			1		35.73										5												
SNP_55	1	168348856		NG																											
SNP_56	1	179343553	GRMZM2G131245	G																											
SNP_57	1	236806700	GRMZM5G836588	G			1	1	25.83																						
SNP_58	3	9584508	GRMZM2G363535	G			1	1	33.24		5																				
SNP_59	3	62335914		NG					7.01															5						
SNP_60	3	64509459		NG																											
SNP_61	3	109095362		NG				1	6.96			5															5				
SNP_62	3	111561600	GRMZM2G411536	G				1	15.11															5	5						
SNP_63	4	47564793		NG																											

SNP_64	4	170388650		NG		Inv4m	1		38.38											5	5							
SNP_65	4	170711055	GRMZM2G473411	G		Inv4m	1		26.73																			5
SNP_66	4	170783306		NG		Inv4m																							
SNP_67	4	171111841		NG		Inv4m	1		28.29		5																		
SNP_68	4	173995149	AC186287.7_FG005	G		Inv4m	1		27.64																				
SNP_69	4	177881170	GRMZM2G300624	G		Inv4m	1		27.23			5																	
SNP_70	4	180504078	GRMZM2G000764	G		Inv4m	1		25.28																				
SNP_71	4	181223878	GRMZM2G021912	G		Inv4m	1		28.14																				
SNP_72	4	225087669		NG																									
SNP_73	5	101940822		NG																									
SNP_74	5	102660991		NG			1		41.66			5																	
SNP_75	5	103126178		NG																									
SNP_76	5	103151251		NG																									
SNP_77	5	103575632		NG																									
SNP_78	5	103652454		NG			1		39.22			5																	
SNP_79	5	104653139		NG																									
SNP_80	5	107146543	GRMZM2G015033	G																									
SNP_81	5	112610568		NG					8.2													5							
SNP_82	5	112693361		NG				1	5.63													5		5					
SNP_83	5	148344999		NG			1		26.16																				
SNP_84	7	150171759	GRMZM2G316191	G	PHD-transcription factor 26 (phd26)																								
SNP_85	8	18204	GRMZM2G122853	G																									
SNP_86	8	21240		NG			1		27.89															5		5			
SNP_87	8	228284		NG			1		25.34								5												
SNP_88	8	260479		NG																									
SNP_89	8	356976	GRMZM2G085038	G																									
SNP_90	8	559686		NG			1	1	19.3																				
SNP_91	8	618589	GRMZM2G077233	G			1		19.53																				
SNP_92	8	825774	GRMZM2G400489	G	MYB-related-transcription factor 2 (mybr2)		1		24.35										5	5									
SNP_93	8	828512	GRMZM2G400489	G	MYB-related-transcription factor 2 (mybr2)																								
SNP_94	8	939485	GRMZM2G387528	G																									
SNP_95	8	1343730	GRMZM2G140885	G			1		48.74																				
SNP_96	8	1372690	GRMZM2G042636	G			1		26.26																				
SNP_97	8	2289581	GRMZM2G120876	G			1		26.61	5									5	5									
SNP_98	8	2315211	GRMZM2G120857	G	isopropylmalate dehydrogenase4 (imd4)		1		24.18										5	5									
SNP_99	8	2350251	GRMZM2G027728	G	ribosomal protein L30 (rpl30)				41.74									11											
SNP_100	8	3576402	GRMZM2G044963	G			1		32.61				5																

SNP_101	8	3979145	GRMZM2G139952	G			1		20.99																				
SNP_102	8	4515370	GRMZM2G022711	G			1		37.68																				
SNP_103	8	4598922		NG			1		27.68										5	5									
SNP_104	8	4616707	GRMZM2G029731	G			1		36.95																				
SNP_105	8	4801331	GRMZM2G049429	G			1		25.54																				
SNP_106	8	4843324		NG			1		27.91																				
SNP_107	8	5915456	GRMZM2G024151	G			1		25.87																				
SNP_108	8	6809333	GRMZM2G056145	G			1		28.41																				
SNP_109	8	8452261		NG																									
SNP_110	8	13716313		NG					41.15																				
SNP_111	8	17372541		NG			1		38.91																				
SNP_112	8	28124172	GRMZM2G075222/GRMZM2G075250	G			1		22.84																				
SNP_113	8	28525840	GRMZM2G018228	G			1		30.17																				
SNP_114	8	28699032	GRMZM2G379286	G			1		23.55										5	5									
SNP_115	8	28763790	GRMZM2G107205	G			1		24.02										5	5									
SNP_116	8	28852684	GRMZM2G122003	G			1		23.45										5										
SNP_117	8	29016489	GRMZM2G393347	G			1		24.04										5	5									
SNP_118	8	32570959	GRMZM2G075992	G			1		41.12				11						5	5			5						
SNP_119	8	43099181		NG				1	11.11	5		5																	
SNP_120	8	52086769		NG				1	10.59	5		5											5						
SNP_121	8	55936851	GRMZM2G131055/GRMZM2G131074	G			1		40.87				5										5						
SNP_122	8	67729118	GRMZM2G043764/GRMZM2G345039	G	RNase PH homolog (rph1)			1	7.88	5		5							5	5				5					
SNP_123	9	130990858	GRMZM2G084525	G		Inv9e		1	6.27	5-11													5		5				
SNP_124	9	135144815	GRMZM2G121456	G	virescent30 (v30)	Inv9e		1	10.62	5		5							5-11	5			5	5	5				
SNP_125	1	165413025		NG					25.26				5						5	5					5				
SNP_126	1	266029571		NG			1		24.77												5				5				
SNP_127	3	84903125	GRMZM2G112579	G																									
SNP_128	3	111866508	GRMZM2G135322	G				1	15.86										5				5						
SNP_129	4	45663714		NG					16.54										5	5					5				5
SNP_130	4	233810694	GRMZM2G063909	G			1		40.85	5	5	5								5		5	5						
SNP_131	5	100073232		NG			1		43.46														5						
SNP_132	8	45263545	GRMZM2G142870	G	multidrug resistance associated protein6 (mrpa6)		1		11.65										5	5			5		5-11				
SNP_133	8	68313700	GRMZM2G168654	G					18.89	5		5													5				
SNP_134	8	72300837	GRMZM2G403797	G			1		26.36				5										5						
SNP_135	9	119304019	GRMZM2G098079	G		Inv9e																							
SNP_136	9	126017135	GRMZM2G170843	G		Inv9e	1	1	8.39	5		5						11	5	5			5	5	5-11		5		
SNP_137	9	136302086	GRMZM2G115825	G		Inv9e	1	1	5.97	5		5					5		5	5			5	5	5		5		
SNP_138	9	139382337	GRMZM2G166767	G		Inv9e																							
SNP_139	9	140781294	GRMZM2G083402	G		Inv9e																							

SNP_140	9	141470887	GRMZM2G146313	G		Inv9e																							
SNP_141	9	145765201	GRMZM2G328268	G				1	17.3	5		5									5	5			5				
SNP_142	1	49815938	GRMZM2G013639	G					10.9	5		5					5					5		5	5	5			
SNP_143	1	54922664		NG					44.39	5		5																
SNP_144	1	137979331		NG																									
SNP_145	1	227989956		NG			1		34.92																				
SNP_146	1	228005584		NG																									
SNP_147	1	236818610	GRMZM5G836588	G					10.12												5				5				
SNP_148	1	243023543	GRMZM2G125762	G					47.94	5			5				5-11									11			
SNP_149	1	258688386		NG					41.07								5-11					5							
SNP_150	1	268560405	GRMZM2G000622	G			1		37.64				5												5				
SNP_151	2	12442449	GRMZM2G313020	G					28.04				5																
SNP_152	2	31448551	GRMZM2G041876	G																									
SNP_153	2	74084625	GRMZM2G026918	G				1	35.43	5	5	5																	5
SNP_154	2	83127028		NG			1		24.06												5								
SNP_155	2	133085160		NG					16.96															5					
SNP_156	2	146125013	GRMZM2G076415	G					45.87	5		5				5-11													
SNP_157	2	163413504		NG					24.72	5		5	5									5			5	5	5-11		
SNP_158	2	189968292	GRMZM2G041633	G					45.98	5	5	5										5							
SNP_159	2	217634370	GRMZM2G109425	G					30.02																5	5	5		
SNP_160	2	230520562	GRMZM2G158872	G					10.19												5	5			5		5		
SNP_161	3	8606769	AC232289.2_FG005	G					42.27	5		5																	
SNP_162	3	14375497	GRMZM5G816228	G																									
SNP_163	3	110514686	GRMZM2G421029	G																									
SNP_164	3	112158024		NG			1		35.61	5																			
SNP_165	3	113076088	GRMZM2G369203	G					49.26	5		5																	
SNP_166	3	123867416	GRMZM2G073750	G	ARF-transcription factor 9 (arftf9)				23.45																				
SNP_167	3	162828569	GRMZM2G133173	G					21.22			5																	
SNP_168	3	177124534	GRMZM2G396231	G					35.41																				
SNP_169	3	219383329	GRMZM2G061890	G			1		39.8	5		5																	5
SNP_170	3	226255690	GRMZM2G030596	G					38.8																5				
SNP_171	3	231056268		NG					33.53																				
SNP_172	4	5666978	GRMZM2G169160	G	seryl-tRNA synthetase (ser2)				30.73																				
SNP_173	4	9126461		NG				1	34.52	5		5	5	5			5				5	5		5	5	5	5		
SNP_174	4	41405883		NG			1		25.11	5		5																	
SNP_175	4	51902320		NG			1		29.68	5																			
SNP_176	4	78668023	GRMZM2G415219	G					37.55																				5
SNP_177	4	122884528		NG				1	45.35																				
SNP_178	4	195551119		NG					41.27																				
SNP_179	5	20080442	GRMZM2G156748	G					30.45																	5	5-11		

SNP_180	5	77196738		NG					28.29	5		5												5	5			
SNP_181	5	87715294		NG			1		28.01																				
SNP_182	5	98391122		NG				1	32.77																				
SNP_183	5	113828969		NG			1		30.75																		5		
SNP_184	5	115978816	GRMZM2G452633	G					35.98												5		5						
SNP_185	5	137582803		NG			1		34.51														5						
SNP_186	5	148348395		NG			1		24.66																				
SNP_187	5	184701046	GRMZM2G560811	G					19.69												5								
SNP_188	5	193653062	GRMZM2G005996	G			1		21.67	5									5										
SNP_189	6	43975209	GRMZM2G043995	G					41.68				5										5						
SNP_190	6	45242004		NG			1		20.58											5									
SNP_191	6	52199461		NG					36.7																				
SNP_192	6	54776449	GRMZM2G134214	G	PHD-transcription factor 32 (phd32)		1		26.76														5						
SNP_193	6	157274662		NG					32.2																				
SNP_194	6	169128855		NG																									
SNP_195	7	37836093	GRMZM2G071877	G			1		31.77		5	5							5	5	5	5							
SNP_196	7	51505486		NG			1		37.61				5						5	5			5		5				
SNP_197	7	71077552		NG																									
SNP_198	7	75676145		NG																									
SNP_199	7	84801067		NG					39.36	5		5																	
SNP_200	7	116544684		NG					34.11	5		5																	
SNP_201	7	139116154	GRMZM2G081048/GRMZM2G080871	G			1		43.66																				
SNP_202	7	144729861		NG																									
SNP_203	7	156573041	GRMZM2G153162	G					47.78														5						
SNP_204	9	12269764	GRMZM2G436583	G		Inv9d	1		24.66	5							5-11										5		
SNP_205	9	51050862	GRMZM2G100103	G	poor homologous synapsis1 (phs1)				42.84																				
SNP_206	9	119766875		NG		Inv9e	1	1	8.2	5		5					5	11	5	5			5	5	5-11		5		
SNP_207	9	120626342		NG		Inv9e	1	1	8.62	5		5					5	11	5	5			5	5	5		5		
SNP_208	9	125690268	GRMZM2G174286	G		Inv9e	1	1	10.43										5	5		5	5	5					
SNP_209	9	127497303	GRMZM2G117746	G		Inv9e	1	1	15.12	5		5			5		5		5	5			5				5		
SNP_210	9	134996351	GRMZM2G092174	G	phytochromeB2 (phyB2)	Inv9e	1		41.1	5									5-11	5-11			5						
SNP_211	9	135082028	GRMZM2G090869	G		Inv9e	1	1	7.78	5		5					5		5	5			5	5	5-11		5		
SNP_212	9	135153323	GRMZM2G121303	G		Inv9e																							
SNP_213	9	135258226	GRMZM5G875502	G	enhancer of zeste2 (mez2)	Inv9e	1	1	8.15	5		5					5		5	5			5	5	5		5		
SNP_214	9	138324711	GRMZM2G145451	G		Inv9e		1	32.81														5	5					
SNP_215	9	141331982	GRMZM2G481069	G		Inv9e	1	1	7.83	5		5					5		5	5			5	5	5-11		5		
SNP_216	10	54940038	GRMZM2G101900	G																									
SNP_217	10	101049847	GRMZM2G159330	G																									
SNP_218	10	118507744	GRMZM2G437119	G					31.58	5																			

S6 Table. Number of individuals used to test associations between 171 SNPs and 18 phenotypes.

SNP	PL	HLE	HHE	Til	LBr	NoE	LeL	LeW	LeC	FFT	MFT	TBr	Gr	GrL	GrWi	GrWe	GrC	StD
SNP_1	1045	1025	1018	1033	1025	1034	1046	1046	908	835	837	1042	918	1015	1015	989	997	602
SNP_2	1067	1047	1040	1055	1047	1056	1068	1068	926	851	853	1063	938	1038	1038	1012	1020	612
SNP_3	1054	1035	1028	1042	1035	1043	1055	1055	913	842	844	1050	925	1025	1025	1000	1007	605
SNP_4	1055	1035	1028	1043	1035	1044	1056	1056	917	841	843	1051	927	1026	1026	1000	1008	606
SNP_5	990	971	964	978	971	979	990	990	853	785	787	986	869	965	965	940	947	565
SNP_6	1043	1024	1017	1031	1024	1032	1044	1044	906	832	834	1039	917	1015	1015	989	997	600
SNP_7	1058	1038	1031	1046	1038	1047	1059	1059	917	843	845	1054	929	1028	1028	1003	1011	606
SNP_9	1059	1040	1033	1047	1039	1048	1060	1060	919	844	847	1055	931	1030	1030	1004	1012	608
SNP_10	1065	1045	1039	1053	1045	1054	1066	1066	926	848	850	1061	936	1035	1035	1010	1018	607
SNP_13	1026	1006	1000	1015	1007	1016	1027	1027	889	810	812	1022	899	997	997	971	979	583
SNP_14	1044	1025	1020	1033	1024	1033	1045	1045	906	832	834	1040	915	1015	1015	991	997	599
SNP_15	1056	1036	1029	1044	1037	1045	1057	1057	917	845	847	1053	928	1028	1028	1002	1010	605
SNP_17	1047	1027	1020	1035	1027	1036	1048	1048	910	836	838	1043	920	1019	1019	993	1001	598
SNP_18	1052	1034	1026	1040	1032	1041	1053	1053	917	836	838	1048	927	1023	1023	997	1005	601
SNP_19	1051	1031	1024	1039	1031	1040	1052	1052	914	837	839	1047	923	1022	1022	997	1004	602
SNP_20	575	569	565	565	555	564	576	576	514	425	427	571	508	546	546	543	545	264
SNP_21	1026	1007	1000	1015	1006	1015	1027	1027	892	813	815	1022	902	997	997	973	980	589
SNP_22	962	943	936	950	945	952	963	963	837	766	768	960	843	935	935	909	917	561
SNP_24	1019	1001	994	1008	999	1008	1020	1020	885	820	822	1015	893	992	992	967	975	579
SNP_25	1037	1018	1011	1025	1017	1026	1038	1038	901	822	824	1033	909	1006	1006	982	988	594
SNP_26	1059	1039	1033	1048	1039	1049	1060	1060	920	845	847	1056	931	1030	1030	1004	1012	606
SNP_27	1047	1028	1021	1036	1027	1036	1047	1047	908	834	836	1043	921	1019	1019	993	1001	599
SNP_28	1056	1036	1029	1044	1036	1045	1057	1057	915	842	844	1052	930	1028	1028	1003	1010	607
SNP_29	1052	1032	1025	1040	1033	1041	1053	1053	915	838	840	1048	924	1023	1023	997	1005	603
SNP_30	1040	1020	1013	1029	1020	1030	1041	1041	903	829	831	1036	912	1012	1012	986	994	594
SNP_32	1010	991	984	999	991	999	1011	1011	877	804	806	1006	888	983	983	958	966	575
SNP_33	1055	1035	1028	1043	1035	1044	1056	1056	916	839	841	1051	926	1025	1025	1000	1008	601
SNP_34	943	926	919	931	925	932	943	943	817	733	735	939	826	918	918	894	901	536

SNP	PL	HLE	HHE	Til	LBr	NoE	LeL	LeW	LeC	FFT	MFT	TBr	Gr	GrL	GrWi	GrWe	GrC	StD
SNP_35	1039	1019	1012	1028	1019	1028	1040	1040	901	835	837	1035	910	1010	1010	984	992	593
SNP_37	1044	1024	1017	1032	1024	1033	1045	1045	910	839	841	1040	916	1015	1015	989	998	601
SNP_38	1050	1030	1023	1039	1030	1039	1051	1051	911	837	839	1046	922	1020	1020	994	1002	601
SNP_39	1046	1026	1019	1035	1026	1035	1047	1047	908	833	835	1042	919	1016	1016	991	998	598
SNP_40	1046	1026	1019	1034	1026	1035	1047	1047	908	830	832	1042	918	1017	1017	993	1000	595
SNP_41	1032	1012	1005	1021	1012	1021	1033	1033	896	819	821	1028	908	1003	1003	978	985	592
SNP_45	913	895	889	903	896	904	914	914	785	719	721	909	795	886	886	861	869	527
SNP_47	1063	1043	1036	1051	1043	1052	1064	1064	923	848	850	1059	934	1034	1034	1008	1016	612
SNP_48	1051	1031	1024	1039	1032	1040	1052	1052	913	839	841	1047	924	1022	1022	996	1004	602
SNP_49	1057	1037	1030	1046	1037	1046	1058	1058	920	842	844	1053	930	1028	1028	1002	1010	605
SNP_50	1040	1021	1014	1028	1020	1029	1041	1041	911	826	828	1036	918	1012	1012	988	994	587
SNP_51	1064	1044	1037	1052	1044	1053	1065	1065	925	848	850	1060	935	1035	1035	1009	1017	609
SNP_52	1059	1039	1032	1048	1039	1048	1060	1060	919	844	846	1055	929	1029	1029	1003	1011	607
SNP_53	967	950	944	956	947	958	968	968	834	780	781	963	844	939	939	913	921	554
SNP_54	961	943	937	950	943	952	962	962	830	760	763	958	841	935	935	912	919	553
SNP_57	1053	1033	1026	1041	1033	1042	1054	1054	916	840	842	1049	924	1024	1024	998	1006	602
SNP_58	1015	995	988	1003	995	1004	1016	1016	880	814	816	1011	891	987	987	961	969	584
SNP_59	1031	1012	1005	1019	1011	1020	1032	1032	893	820	822	1027	906	1002	1002	978	986	593
SNP_61	1053	1033	1026	1041	1033	1042	1054	1054	913	838	839	1049	925	1024	1024	998	1006	608
SNP_62	1057	1037	1030	1045	1037	1046	1058	1058	921	842	844	1053	928	1028	1028	1002	1010	607
SNP_64	998	978	971	986	978	987	999	999	870	796	798	994	873	969	969	944	952	573
SNP_65	1040	1021	1014	1029	1020	1029	1041	1041	905	825	827	1036	916	1014	1014	989	997	594
SNP_67	978	959	953	967	959	968	979	979	849	775	776	975	856	953	953	929	936	566
SNP_68	1053	1033	1026	1041	1033	1042	1054	1054	915	837	839	1049	927	1025	1025	999	1007	600
SNP_69	994	975	968	981	978	984	994	994	862	795	798	991	870	967	967	941	950	568
SNP_70	1046	1026	1019	1034	1026	1035	1047	1047	914	830	832	1042	920	1018	1018	993	1000	596
SNP_71	1049	1030	1023	1037	1029	1038	1050	1050	913	834	836	1045	922	1022	1022	996	1004	598
SNP_74	1047	1027	1020	1035	1027	1036	1048	1048	908	835	836	1043	919	1018	1018	993	1000	606
SNP_78	987	968	961	976	968	976	988	988	857	799	800	983	860	957	957	933	939	569
SNP_81	1041	1021	1015	1029	1023	1031	1042	1042	902	826	828	1037	915	1015	1015	989	997	601
SNP_82	1011	991	985	999	991	1001	1012	1012	876	809	810	1007	890	982	982	961	967	579

SNP	PL	HLE	HHE	Til	LBr	NoE	LeL	LeW	LeC	FFT	MFT	TBr	Gr	GrL	GrWi	GrWe	GrC	StD
SNP_83	1054	1034	1027	1042	1034	1043	1055	1055	915	841	843	1050	926	1025	1025	999	1008	603
SNP_86	992	974	967	981	972	981	993	993	858	802	804	988	867	963	963	938	946	561
SNP_87	979	960	953	966	959	968	979	979	847	790	792	975	853	949	949	925	932	554
SNP_90	1017	998	991	1007	997	1006	1018	1018	883	807	809	1013	888	988	988	962	971	586
SNP_91	1053	1033	1026	1041	1033	1042	1054	1054	916	840	842	1049	925	1024	1024	999	1006	606
SNP_92	1054	1034	1027	1042	1034	1043	1055	1055	915	841	843	1050	927	1025	1025	999	1007	603
SNP_95	1037	1018	1011	1025	1018	1026	1038	1038	899	824	826	1033	910	1009	1009	984	991	595
SNP_96	1049	1029	1022	1037	1029	1038	1050	1050	914	836	838	1045	920	1020	1020	994	1002	599
SNP_97	1027	1007	1000	1014	1008	1016	1027	1027	889	824	826	1023	902	1000	1000	976	985	585
SNP_98	1049	1029	1022	1037	1029	1038	1050	1050	908	833	835	1045	922	1020	1020	994	1002	604
SNP_99	1033	1014	1007	1021	1013	1022	1034	1034	899	820	822	1029	908	1004	1004	980	988	589
SNP_100	1048	1029	1022	1036	1028	1037	1049	1049	912	836	838	1044	922	1019	1019	993	1001	602
SNP_101	1030	1010	1003	1018	1010	1019	1031	1031	896	821	823	1026	905	1001	1001	975	983	586
SNP_102	1049	1029	1023	1038	1029	1039	1050	1050	909	836	838	1045	923	1020	1020	994	1002	602
SNP_103	1006	986	981	996	987	996	1007	1007	874	799	800	1002	882	978	978	952	960	583
SNP_104	1050	1031	1024	1038	1031	1039	1051	1051	910	839	841	1046	923	1023	1023	997	1005	600
SNP_105	1016	996	990	1005	998	1006	1017	1017	876	802	804	1013	889	988	988	963	970	581
SNP_106	1043	1023	1016	1032	1023	1032	1044	1044	903	831	833	1039	918	1015	1015	990	997	595
SNP_107	1059	1039	1032	1047	1039	1048	1060	1060	920	844	846	1055	930	1029	1029	1003	1011	603
SNP_108	1032	1013	1007	1021	1016	1023	1033	1033	898	829	831	1028	906	1005	1005	980	987	594
SNP_110	887	868	864	878	870	880	888	888	764	703	705	883	778	865	865	840	850	521
SNP_111	1004	985	979	993	985	993	1005	1005	872	792	794	1001	876	975	975	951	957	581
SNP_112	1048	1029	1022	1036	1028	1037	1049	1049	913	833	834	1044	919	1019	1019	993	1001	598
SNP_113	1028	1008	1001	1016	1008	1017	1028	1028	893	814	816	1024	904	1003	1003	977	985	583
SNP_114	1052	1032	1025	1040	1032	1041	1053	1053	916	839	841	1048	924	1024	1024	998	1006	603
SNP_115	1062	1042	1035	1049	1042	1051	1062	1062	924	845	847	1058	933	1032	1032	1006	1014	605
SNP_116	1055	1035	1028	1043	1035	1044	1056	1056	918	840	842	1051	928	1026	1026	1001	1009	604
SNP_117	1038	1018	1011	1026	1018	1027	1039	1039	899	824	826	1034	915	1010	1010	985	992	597
SNP_118	1037	1017	1010	1026	1018	1027	1038	1038	903	828	830	1033	910	1008	1008	982	991	596
SNP_119	608	602	599	598	590	597	609	609	549	455	457	604	542	581	581	578	580	281
SNP_120	1036	1016	1009	1024	1016	1025	1037	1037	898	821	823	1032	910	1008	1008	982	990	591

SNP	PL	HLE	HHE	Til	LBr	NoE	LeL	LeW	LeC	FFT	MFT	TBr	Gr	GrL	GrWi	GrWe	GrC	StD
SNP_121	1056	1036	1029	1044	1036	1045	1057	1057	919	840	842	1052	929	1027	1027	1001	1009	603
SNP_122	1059	1039	1032	1047	1039	1048	1060	1060	921	843	845	1055	932	1030	1030	1005	1013	605
SNP_123	1057	1037	1030	1045	1037	1046	1058	1058	916	842	844	1053	929	1029	1029	1003	1011	605
SNP_124	1039	1019	1012	1027	1019	1028	1040	1040	901	825	827	1035	912	1010	1010	985	993	599
SNP_125	1057	1037	1030	1046	1037	1046	1058	1058	918	844	845	1053	928	1028	1028	1002	1010	608
SNP_126	1037	1019	1013	1026	1017	1026	1038	1038	900	825	827	1033	912	1009	1009	983	992	592
SNP_128	1050	1032	1025	1038	1030	1039	1051	1051	914	835	837	1046	923	1021	1021	996	1003	604
SNP_129	1044	1024	1017	1032	1025	1033	1045	1045	908	835	837	1041	917	1015	1015	989	997	601
SNP_130	1060	1041	1034	1048	1041	1049	1061	1061	923	846	848	1056	932	1031	1031	1005	1013	605
SNP_131	1055	1035	1028	1043	1035	1044	1056	1056	915	842	844	1051	927	1025	1025	999	1008	604
SNP_132	1053	1033	1027	1041	1033	1043	1054	1054	914	837	839	1049	929	1026	1026	1001	1009	601
SNP_133	1005	986	979	993	985	994	1006	1006	869	792	794	1001	881	977	977	951	960	576
SNP_134	908	889	882	898	889	897	909	909	777	702	702	904	791	881	881	858	866	523
SNP_136	1056	1036	1029	1044	1036	1045	1057	1057	917	843	845	1052	928	1027	1027	1002	1009	606
SNP_137	1022	1002	995	1011	1003	1011	1023	1023	886	807	809	1018	906	995	995	971	978	586
SNP_141	1052	1032	1025	1040	1032	1041	1053	1053	913	840	842	1048	925	1023	1023	998	1005	603
SNP_142	1053	1033	1027	1041	1033	1043	1054	1054	912	836	839	1049	924	1023	1023	998	1005	602
SNP_143	1043	1023	1016	1031	1024	1032	1044	1044	905	837	839	1039	915	1014	1014	989	997	598
SNP_145	1039	1019	1012	1027	1021	1028	1040	1040	904	829	831	1035	912	1011	1011	985	993	596
SNP_147	1046	1026	1020	1036	1027	1036	1047	1047	909	835	837	1043	919	1017	1017	992	999	600
SNP_148	1015	996	989	1004	995	1004	1016	1016	883	803	805	1011	890	986	986	960	969	578
SNP_149	957	939	933	947	941	949	958	958	827	761	763	955	841	933	933	910	916	560
SNP_150	1023	1004	997	1011	1004	1012	1024	1024	893	816	818	1019	901	996	996	971	978	587
SNP_151	1038	1019	1013	1026	1018	1027	1039	1039	899	828	830	1034	910	1009	1009	985	993	590
SNP_153	1038	1019	1013	1027	1019	1029	1039	1039	901	826	828	1035	913	1012	1012	987	994	596
SNP_154	1051	1031	1024	1039	1031	1040	1052	1052	913	838	839	1047	923	1021	1021	995	1004	599
SNP_155	1041	1021	1014	1029	1023	1030	1042	1042	905	832	834	1037	916	1015	1015	989	997	599
SNP_156	1008	989	982	996	989	997	1009	1009	873	805	807	1004	885	981	981	958	963	581
SNP_157	993	974	967	981	974	983	994	994	856	786	787	989	869	965	965	939	947	566
SNP_158	1040	1021	1014	1028	1020	1029	1041	1041	904	827	829	1036	916	1011	1011	988	995	587
SNP_159	1012	992	986	1002	992	1001	1013	1013	879	805	807	1008	885	982	982	958	965	584

SNP	PL	HLE	HHE	Til	LBr	NoE	LeL	LeW	LeC	FFT	MFT	TBr	Gr	GrL	GrWi	GrWe	GrC	StD
SNP_160	1041	1021	1014	1029	1021	1030	1042	1042	904	834	836	1037	915	1012	1012	988	995	596
SNP_161	1026	1007	1000	1014	1007	1016	1027	1027	890	820	821	1023	901	999	999	973	981	592
SNP_164	1032	1012	1005	1020	1013	1021	1033	1033	895	818	820	1028	907	1005	1005	979	987	592
SNP_165	1058	1038	1031	1046	1038	1047	1059	1059	919	842	844	1054	930	1029	1029	1003	1011	605
SNP_166	1057	1037	1030	1045	1038	1046	1058	1058	918	843	845	1053	928	1028	1028	1002	1010	605
SNP_167	1051	1031	1024	1040	1031	1040	1052	1052	913	837	839	1047	923	1023	1023	997	1005	602
SNP_168	1057	1037	1031	1045	1037	1047	1058	1058	920	843	845	1053	929	1029	1029	1003	1012	604
SNP_169	1019	1000	993	1007	1000	1009	1020	1020	883	814	816	1015	895	991	991	965	973	587
SNP_170	1021	1003	996	1010	1003	1012	1022	1022	885	815	817	1017	897	994	994	968	976	589
SNP_171	1034	1014	1007	1022	1014	1023	1035	1035	898	825	827	1030	906	1005	1005	979	987	595
SNP_172	1007	988	982	997	989	997	1008	1008	875	806	808	1003	884	980	980	956	963	581
SNP_173	1039	1019	1012	1027	1019	1028	1040	1040	899	827	829	1036	913	1010	1010	987	992	596
SNP_174	1029	1010	1003	1017	1009	1018	1030	1030	898	816	817	1025	904	1000	1000	975	985	587
SNP_175	1004	984	977	993	984	993	1005	1005	871	792	794	1000	880	976	976	950	960	579
SNP_176	1013	993	988	1002	994	1003	1014	1014	876	802	803	1009	888	985	985	959	968	582
SNP_177	1066	1046	1039	1054	1046	1055	1067	1067	926	849	851	1062	937	1036	1036	1010	1018	610
SNP_178	1041	1022	1015	1029	1021	1030	1042	1042	902	832	833	1037	915	1013	1013	987	995	603
SNP_179	941	921	914	930	923	930	942	942	803	734	736	937	822	911	911	886	895	544
SNP_180	1031	1012	1007	1020	1012	1021	1032	1032	899	824	826	1028	906	1004	1004	978	986	581
SNP_181	1039	1019	1012	1027	1019	1028	1040	1040	905	830	832	1035	911	1009	1009	983	991	593
SNP_182	1042	1022	1015	1031	1022	1031	1043	1043	905	828	830	1038	915	1014	1014	988	996	600
SNP_183	1048	1028	1021	1036	1028	1037	1049	1049	908	833	834	1044	920	1019	1019	993	1002	597
SNP_184	1052	1033	1026	1040	1032	1041	1053	1053	916	839	841	1048	929	1025	1025	1000	1007	597
SNP_185	1002	983	976	991	983	992	1003	1003	873	796	798	999	882	976	976	952	959	577
SNP_186	1051	1031	1024	1040	1031	1040	1052	1052	915	837	839	1047	923	1023	1023	997	1005	603
SNP_187	1041	1023	1016	1029	1021	1030	1042	1042	903	829	831	1037	913	1013	1013	988	995	596
SNP_188	1055	1035	1028	1043	1035	1044	1056	1056	916	842	844	1051	927	1026	1026	1000	1008	603
SNP_189	1053	1033	1026	1042	1033	1042	1054	1054	914	838	840	1049	927	1023	1023	998	1006	603
SNP_190	1051	1031	1024	1039	1031	1040	1052	1052	912	835	837	1047	924	1023	1023	997	1006	601
SNP_191	1055	1035	1028	1043	1035	1044	1056	1056	918	840	842	1051	926	1025	1025	1000	1007	602
SNP_192	1043	1024	1018	1031	1023	1032	1044	1044	908	828	830	1039	916	1014	1014	989	996	600

SNP	PL	HLE	HHE	Til	LBr	NoE	LeL	LeW	LeC	FFT	MFT	TBr	Gr	GrL	GrWi	GrWe	GrC	StD
SNP_193	1025	1005	998	1014	1005	1014	1026	1026	886	814	815	1022	898	996	996	970	978	590
SNP_195	1051	1031	1025	1039	1031	1041	1052	1052	912	837	839	1047	924	1024	1024	998	1006	602
SNP_196	1004	984	977	993	984	993	1005	1005	867	800	802	1000	875	974	974	948	956	583
SNP_199	1019	999	993	1008	999	1008	1020	1020	883	811	813	1015	897	991	991	966	974	585
SNP_200	1039	1019	1013	1027	1019	1028	1040	1040	903	834	836	1035	912	1008	1008	982	990	590
SNP_201	1033	1013	1006	1021	1013	1022	1034	1034	895	822	823	1029	906	1004	1004	979	987	596
SNP_203	1057	1037	1030	1045	1038	1046	1058	1058	919	845	847	1053	929	1028	1028	1002	1010	607
SNP_204	1051	1031	1025	1039	1031	1041	1052	1052	913	840	842	1047	925	1025	1025	999	1007	604
SNP_205	602	596	592	591	582	591	603	603	544	449	451	598	538	575	575	572	574	275
SNP_206	1054	1034	1027	1042	1035	1043	1055	1055	916	843	845	1050	927	1024	1024	999	1006	602
SNP_207	1032	1012	1006	1021	1014	1021	1033	1033	896	817	819	1028	906	1003	1003	978	985	588
SNP_208	610	604	600	600	590	599	611	611	549	456	458	606	545	583	583	580	582	283
SNP_209	1016	997	990	1005	997	1005	1017	1017	883	812	815	1012	891	988	988	963	970	586
SNP_210	1043	1023	1016	1031	1023	1032	1044	1044	907	830	832	1039	918	1016	1016	990	999	597
SNP_211	1054	1034	1027	1042	1035	1043	1055	1055	915	841	843	1050	928	1025	1025	999	1007	605
SNP_213	1026	1006	999	1014	1006	1015	1026	1026	893	816	818	1023	901	996	996	970	978	585
SNP_214	1047	1027	1020	1035	1028	1036	1048	1048	908	836	838	1043	918	1017	1017	991	999	599
SNP_215	1042	1022	1015	1030	1022	1032	1043	1043	907	829	831	1038	914	1012	1012	987	994	600
SNP_218	1054	1034	1027	1042	1035	1043	1055	1055	914	843	845	1050	928	1027	1027	1001	1009	609

S7 Table. Additive and dominance effects of SNPs associated to traits after the 11 population structure correction.

SNP	Trait	Additive effect	Dominance effect
SNP_25	PL	0.062	0.125
SNP_123	PL	-0.168	0.129
SNP_25	HHE	0.063	0.138
SNP_30	HHE	-0.135	0.106
SNP_156	LeL	-0.023	-0.027
SNP_148	LeW	0.233	0.17
SNP_149	LeW	0.226	0.25
SNP_204	LeW	-0.146	0.343
SNP_99	LeC	0.102	-0.143
SNP_136	LeC	0.822	1.139
SNP_206	LeC	0.896	0.984
SNP_207	LeC	0.969	0.794
SNP_7	FFT	4.243	-4.788
SNP_15	FFT	-1.485	4.877
SNP_124	FFT	1.803	-4.858
SNP_210	FFT	0.161	-3.364
SNP_15	MFT	-1.337	5.543
SNP_28	MFT	5.676	5.579
SNP_210	MFT	0.911	-4.038
SNP_118	Til	-1.175	-1.586
SNP_1	GrL	0.052	0.339
SNP_157	GrL	-0.132	-0.203
SNP_148	GrWe	0.099	0.316
SNP_157	GrWe	-0.249	-0.038
SNP_179	GrWe	0.172	-0.226
SNP_132	GrWe	0.489	-0.296
SNP_136	GrWe	0.762	1.768
SNP_206	GrWe	0.849	1.494
SNP_211	GrWe	1.161	2.075
SNP_215	GrWe	1.195	1.923

ANNEX I. Stomata identification

Stomata detection in Teosinte

Microscopic imaging

Leaf samples were stored at 4°C during the timeframe of imaging. From each sample one leaf (of three per bag) was used. From the leaf one 5mm disc was cut for microscopic analysis. The disc was cut from the middle (vertically) of the leaf. The leaf discs were put into 80-well glass bottom plates and pressed against the well bottom using a custom-made, spring-mounted stamp array. The samples were stained with Calcofluor + 10% KOH (1:1) in order to increase cell wall fluorescence. Microscopic images of the leaf discs were taken in high throughput using the Opera High Content Screening System. A 20x water-immersion objective was used. A 488nm laser was used for excitation and fluorescence was captured at 520nm. For each disc, 9 images of pre-defined locations were taken, hereinafter referred to as image "fields". Each field represents an area of 0.15mm². In each of these locations 10 images were taken in a stack along the z-Axis to counter the height variability on the leaf surface. This resulted in 9 stacks per sample and a total of 24,480 stacks.

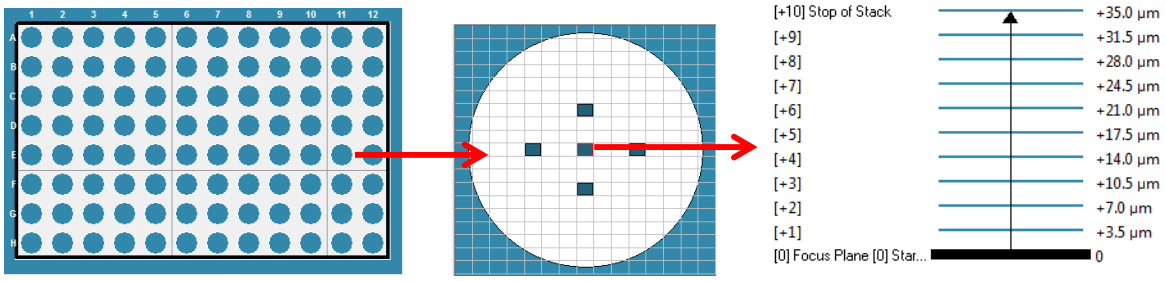


Figure 1: Scheme of the different levels of image acquisition. First picture shows the 96-well plate (first and last column could not be used for technical reason). Second picture shows a single well with 5 image field selected for acquisition. Last picture shows the z-stack of images for the central image field.

Image Analysis

The image analysis algorithms were implemented in Matlab. For stomata detection, the image "stacks" were collapsed into single 2D images by maximum intensity projection, and then saved in bitmap format. Additionally, a second composite image with enhanced cell walls was created for each stack. Therefor each layer was filtered to reduce background signal and increase contrast and was excluded from the stack if the cell wall to background ratio was too low.

Detection of stomata

A large fraction of images were not suited for stomata analysis due to disturbing factors caused by the nature of the samples, e.g. leaf veins, molding or surface height variation. Therefore, images were automatically selected based on the median brightness of 9 image blocks: The median brightness of at least 7 blocks had to be greater than 80. The cutoff for this was set based on two sets of manually selected good quality images and bad quality images, respectively. In high quality images stomata appeared as black holes in a white surface of epidermal cells (Figure 3, top-left). Thus, for initial object detection a simple intensity threshold was used: all pixels lower than 30 were set to one and all others set to 0, thus creating a binary image with white pixels (1) as foreground object and black pixels (0) as background (Figure 3, top-right). Then a number of filters were applied to these preliminary objects: First, very small objects (<100px) were removed. Then adjacent objects were merged by dilation followed by erosion in order to connect the two holes that form one stoma. Then objects that were too small (<300px) or too large (<3000px) for stomata were removed (Figure 3, bottom-left). The median size of the remaining preliminary stomata was calculated. In order to be considered as true stomata objects had to meet the following requirements: First, the object's area should be greater than the median area - 500px and smaller than the median area +1200px, but at least 600px. The ellipse representing the object should have a major axis shorter than 3000px and its eccentricity should be smaller than 0.92. All objects passing this filter were considered to be true stomata. As a final check of the detection quality the image was separated in 16 blocks. Only if in at least 13 of these blocks stomata were found the detection was considered successful (Figure 3, bottom-right). Otherwise the results of the image were discarded. This is due to the fact that stomata are generally distributed evenly throughout the image and if this is not the case it is likely because of out of focus areas or disturbing objects in the image. Because area of the image was known (0.15mm²), the stomata counts were converted to stomatal density (stomata/mm²). Then for each sample the median and standard deviation of all measured fields was calculated. In order to test the accuracy of the algorithm, stomata were counted manually for 54 random samples (median of 9 images per sample). These manual counts were then tested for correlation with the automatic measurements. The correlation coefficient R^2 is 0.82, indicating good correlation between the two methods.

Detection of cells

For detection of cells the image with emphasized cell walls was used (Figure 4, top-left). First, image quality was checked using a score for binary thresholding. If this score was smaller than 0.75 the image was discarded. Otherwise, a Canny edge detection function was applied, resulting in a binary image with edges (high contrast areas) marked as white single pixel lines (Figure 4, top-right). This means that each cell wall surrounded by a double line. In order to merge these lines the image was dilated and holes within the cell wall were removed. Then the cell wall was thinned to an equal thickness (Figure 4, bottom-left). The image was inverted so cells became foreground objects. Cells were filtered to be larger than 2500px and smaller than 10,000px. The objects that passed this filter were considered to be true cells (Figure 4, bottom-right). Due to disturbances in the cell wall intensity cells were not closed in all parts of the image and a reliable edge connection algorithm could not be developed. Therefore, cell density was estimated from only the cells that could be detected in the image. For all detected cells the total area was calculated. The number of cells was then divided by the total area to obtain an estimate of cell density in the image. In order to test the accuracy of the algorithm, cells were counted manually on 53 random samples (median of 9 images per sample). These manual counts were then tested for correlation with the automatic measurements. The correlation coefficient R^2 is 0.81, indicating good correlation between the two methods.

Output

For reliable detection of stomata and cells high image quality was crucial. Due to the nature of the samples and the high throughput imaging approach this could not always be achieved. Therefore stomatal density could only be measured in 59% of the 2800 samples. In 41% of these 1670 samples cell density was successfully estimated. Results were saved as a table in csv format with the following columns: Sample ID, Median stomatal density [stomata/mm²], Std. dev. of stomatal density [stomata/mm²], Manual control of stomatal density [stomata/mm²] Number of analyzed fields, Median cell density [cells/mm²] and Manual control of cell density [cells/mm²].

Figures

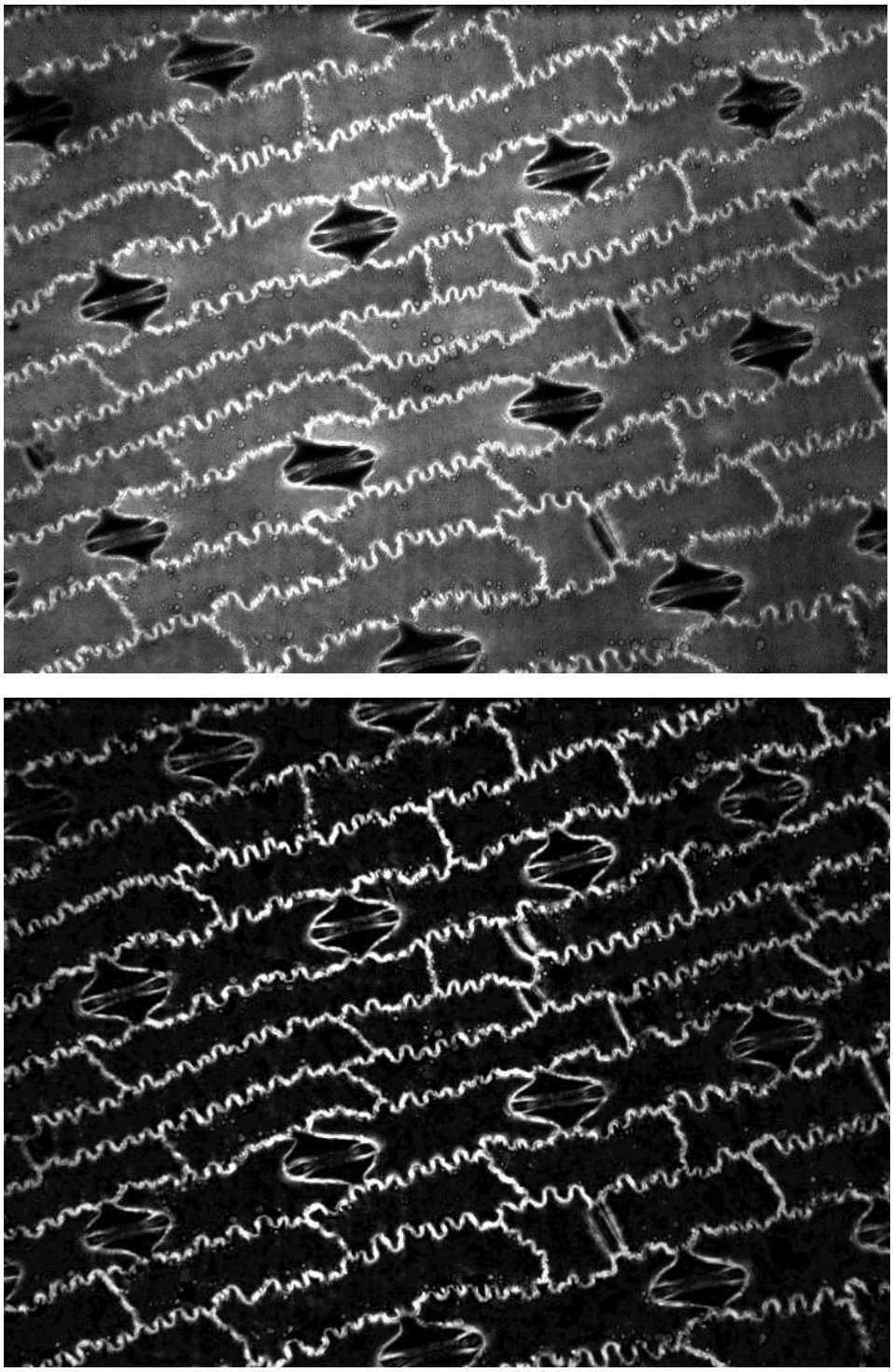


Figure 2: Comparison between image for stomata detection (top) and image for cell detection with emphasized cell walls (bottom).

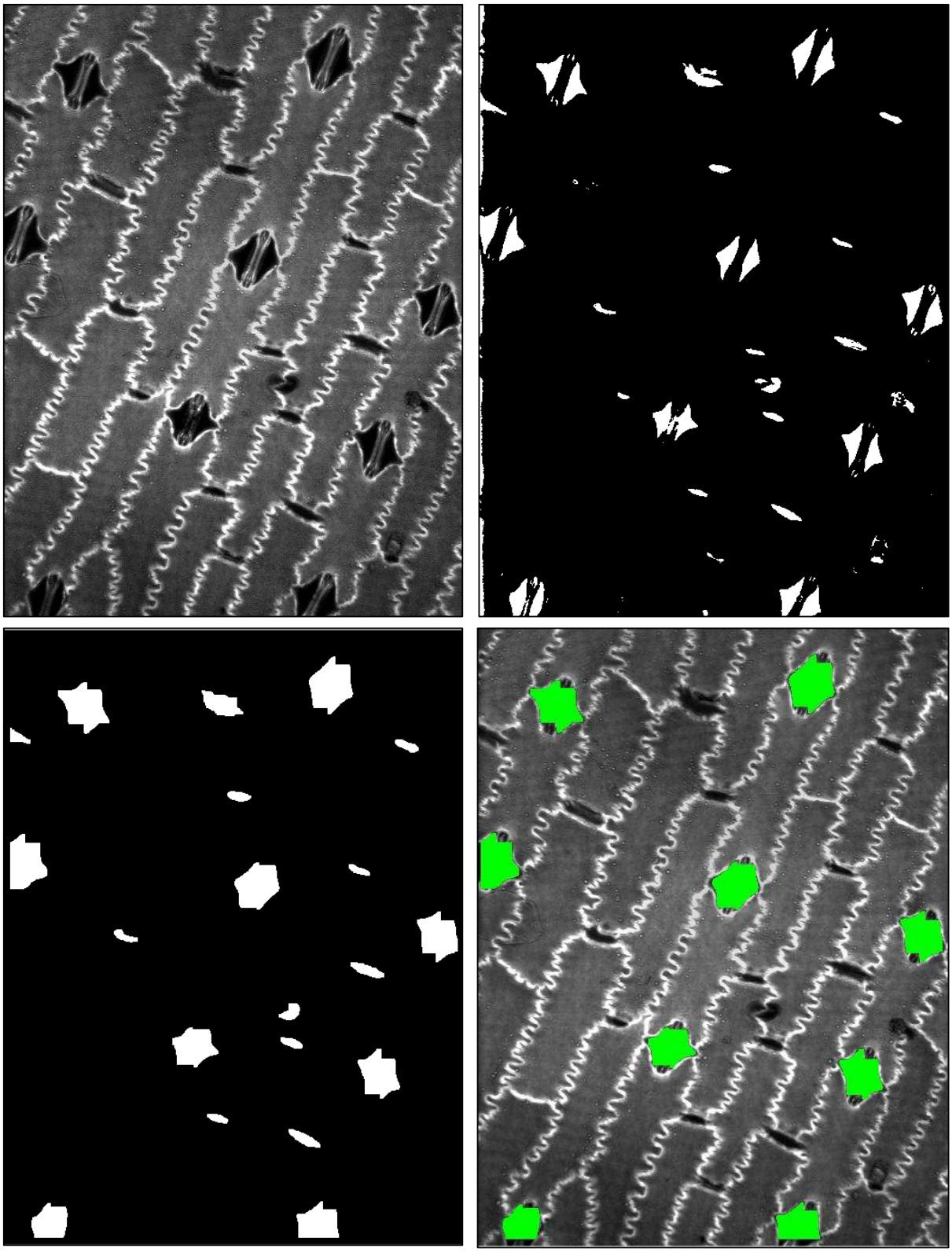


Figure 3: Stages of stomata detection: top-left: original image; top-right: initial binary image; bottom-left: merged objects after first filter; bottom-right: overlay of image and detected stomata (green).

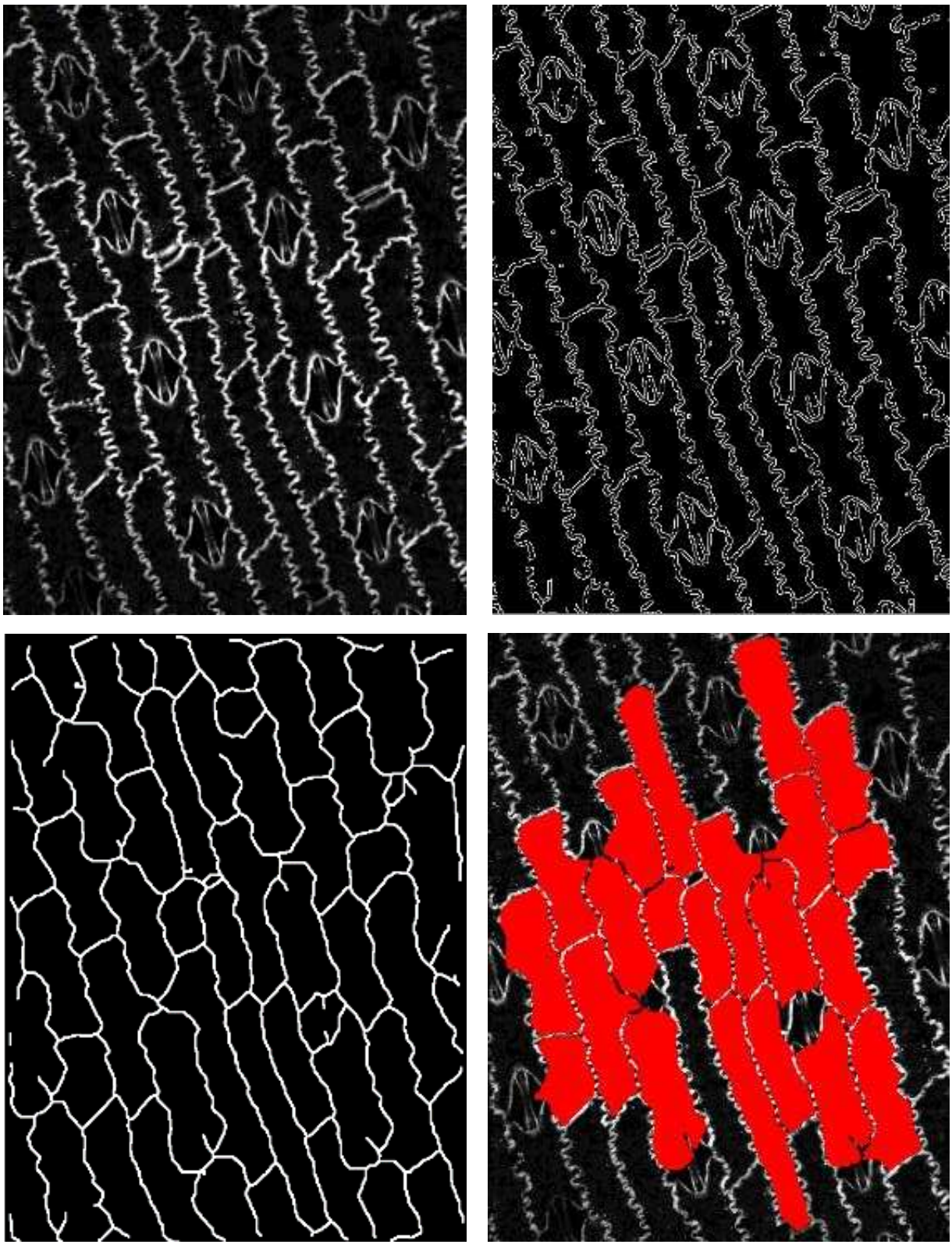


Figure 4: Stages of cell detection: top-left: original image with emphasized cell walls; top-right: edge detection; bottom-left: detected cell walls; bottom-right: overlay of image and detected cells (red).

III. CHAPTER 2 : PATTERNS OF ABUNDANCE AND ADAPTATION ASSOCIATED WITH TRANSPOSABLE ELEMENTS IN TEOSINTE GENOMES

This chapter deals with transposable element (TE) content and variation in teosintes. Its principal aim is to determine TE candidates for local adaptation. To do so, I retrieved raw high-throughput sequencing (HTS) genomic data previously generated on six teosinte populations. I first tested several software to call the insertions, and to estimate their frequencies in the six populations. A major difficulty came from the fact that the HTS data were obtained for 20 individuals but at very low sequencing depth (around 1x per diploid individual). TE insertions discovery was first guided by those present in the B73 maize reference genome annotation. I used PopoolationTE2 software and performed feature adjustments. I performed these analyses on maize B73 version AGPv2, but later reran them on maize B73 genome and annotation AGPv4, that was made publicly available in the course of my PhD.

I subsequently studied the variation of TE insertions in positions different from those in the B73 annotation, hereafter *de novo* TE insertions. I identified these TEs from short read mapping using maize TE sequence library generated from the maize genome annotation AGPv4 by implementing the software Tlex-de-novo. This software is still in course of development by A-S. Fiston-Lavier. I interacted closely with her, through testing and discussing the package.

I characterized and analyzed both reference and *de novo* TE insertion data that I generated. I further applied a series of filters and tests to choose a handful of adaptive candidates. My initial objective was to genotype TE candidates by PCR on the entire panel of 37 teosinte populations spanning both altitudinal gradients as well as on the genetic association panel that encompasses all teosinte individuals from the common garden experiments (in a similar way as was performed for SNP data in Chapter 1). Nevertheless, a series of technical, mainly bioinformatic, difficulties impeded that I arrive to the candidate list with enough time to perform the necessary experiments and analyses. TE genotyping of candidate insertions is foreseen to commence in the immediate future to complete the paper draft that I have included here as Chapter 2.

In a complementary approach, I wished to investigate the role on teosinte adaptation of TE insertions known to have phenotypic effect on maize. To this end I chose three bibliographic candidates and genotyped their polymorphism on the genetic association panel of teosinte populations as well as populations sampled along the altitudinal gradients. I further tested these insertions for association to common garden measured traits as well as correlation with environmental variables. These results are included and discussed in the paper draft.

Patterns of abundance and adaptation associated with transposable elements in teosinte genomes

Authors: Natalia Elena Martínez-Ainsworth[1], Clémentine Vitte[1], Jean-Tristan Brandenburg[2], Hélène Corti[1], Yves Vigouroux[4], Anna-Sophie Fiston-Lavier[3], Domenica Manicacci[1] and Maud Tenaillon[1]

[1] : Génétique Quantitative et Evolution – Le Moulon, Institut National de la Recherche Agronomique, Université Paris-Sud, Centre National de la Recherche Scientifique, AgroParisTech, Université Paris-Saclay, France

[2] : Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa.

[3] : Université de Montpellier, Institut de Recherche pour le développement, UMR Diversité, Adaptation et Développement des plantes, Montpellier, France

[4] : Institut des Sciences de l'Evolution de Montpellier (UMR 5553, CNRS-UM-IRD-EPHE), Université de Montpellier, Place Eugène Bataillon, Montpellier, France

Keywords: transposable element frequencies, de novo insertions, reference insertions, altitudinal adaptation, association mapping, epistasis.

III.1 INTRODUCTION

Transposable elements (TEs) are selfish genetic features that have or have had the capacity to move between different genomic regions. Plant genomes are loaded with TE copies, yet the fraction of the genome they occupy is extremely variable among species ranging, from 1% in *Utricularia gibba* (Ibarra-Laclette, Lyons et al. 2013) to 85% in maize (Schnable, Ware et al. 2009). TE content results from a balance between their transposition rate and their removal by recombination (Vitte and Panaud 2005; Bennetzen 2007; Hawkins, Proulx et al. 2009). But because TEs contribute to functional variation by causing structural rearrangements, gene disruption, and perturbation of gene expression, natural selection is also central to their evolution within their host genomes (Tenaillon, Hollister et al. 2010). This is illustrated by the continuous arms race at play between TE proliferation and TE epigenetic surveillance mechanisms evolved by their hosts to counteract their damaging effects (Lisch and Slotkin 2011).

Population level studies are essential to better characterize the role of natural selection in the evolution of TE content. To date, the contours of the key evolutionary factors governing TE frequencies and fate in plant genomes have been analyzed within the *Arabidopsis* genus (Stuart, Eichten et al. 2016), with the limitations of using a species with a compact genome. Most attention has focused on direct and indirect deleterious effects of TEs and their host's purging efficacy. Hence, in *Arabidopsis thaliana*, the density of methylated TEs correlates negatively with neighboring genes' expression, pointing to a possible methylation fitness cost; an observation consistent with the correlation between the age of TEs and their degree of methylation as well as their distance from genes (Hollister and Gaut 2009). More recently, the analysis of recent transposition events in different accessions of *A. thaliana* revealed that while accession-specific insertions were found equally distributed across the genome, shared insertions were less abundant in gene-rich regions, thus suggesting a purge by purifying selection in these regions (Quadrana, Silveira et al. 2016). Purifying selection being positively linked to the effective population size (N_e), has also been proposed as the main driver of *Ac*-like elements content between the selfer species *A. thaliana* (low N_e, high *Ac* content) and its outcrosser congener *A. lyrata*, (high N_e, low *Ac* content and high segregation frequencies) (Wright, Le et al. 2001). This pattern of stronger purifying selection acting against TE insertions in *A. lyrata* whose N_e is greater, extends to other TE families (Lockton and Gaut 2010).

An often-neglected aspect of TE evolution is the immediate fitness advantages they may confer to their hosts. Polymorphism of TE indels is indeed a major source of phenotypic variation among individuals. Most striking examples of TE benefits have been observed between wild and domesticated forms or among domesticated forms of various origins, where insertions have generated alleles with large effects on phenotypes that have been exploited by human selection (for a review, (Vitte, Fustier et al. 2014). For example, the insertion of a retrotransposon in a MADS-box transcription factor has conferred parthenocarpic fruit development in apple cultivars (Yao, Dong et al. 2012). Likewise, white-skin grape cultivars have been derived from red-skin cultivars through selection of a retrotransposon-induced mutation blocking the expression of a *Myb*-factor regulating anthocyanin biosynthesis (Kobayashi, Goto-Yamamoto et al. 2004). These examples are however still sporadic and their discovery has been guided by the observation of drastic phenotypic changes and top-down approaches. TEs contribution to more subtle, polygenic adaptation has therefore yet to be assessed. Interestingly, results in *A. thaliana* indicate that TE variants tend not to be in linkage disequilibrium with nearby single nucleotide polymorphisms (SNPs), suggesting that they constitute a distinct source of genetic diversity (Stuart, Eichten et al. 2016).

Here, we undertook a first characterization of TE content and putative adaptive insertions in teosintes. Teosintes are the closest wild relatives of maize, a crop where TEs were first discovered and constitute an overwhelming ~85% of the genome (Schnable, Ware et al. 2009). The maize genome is derived from an ancient paleopolyploid ancestor resulting from two whole genome duplication events about 5 and 12 million yeas ago (Blanc 2004; Swigoňová, Lai et al. 2004) as well as a fattening of the genome due to the insertion of TEs within the last 3 million years. These TEs include Miniature Inverted Transposable Elements - MITEs (Zerjal, Joets et al. 2009), but mostly Long Terminal Repeat (LTR) retrotransposons (Sanmiguel, Gaut et al. 1998) that currently occupy over 75% of the maize reference genome assembly (Baucom, Estill et al. 2009). Most maize TEs therefore correspond to insertions predating domestication (Wang and Dooner 2006; Baucom, Estill et al. 2009; Stitzer, Anderson et al. 2019), implying that TEs in maize are most likely a subset of those found in wild teosintes. Extensive variation between maize accessions exists (Wang and Dooner 2006; Chia, Song et al. 2012), in particular TE content has been shown to vary considerably among maize lines (Springer, Anderson et al. 2018). For instance, TE content genome-wide comparison of four maize lines found 1.6 Gb of variable TE sequences with approximately 20% of genome differences between any two genome pairs due to non-shared TEs. Polymorphic TE insertions encompass over 2,000 genes, highlighting TE potential phenotypic effects (Anderson, Stitzer et al.). Maize TE transcription in B73 has been found to be restricted to a small percentage

(~15% as estimated from mappable RNAseq transcripts) of TE families and is highly variable among tissues, with considerable differences among maize lines (Anderson, Stitzer et al. 2019). Interestingly, transcription of a handful of TE families is up-regulated upon abiotic stress and this correlates with up-regulation of nearby genes, suggesting that TEs play a role as potential enhancers of stress-response genes (Makarevitch, Waters et al. 2015)

More specific examples of adaptive insertions in maize include the iconic insertion of a 4.9kb *hopscotch* retrotransposon in the regulatory region of the *teosinte branched 1* (*Tb1*) transcription factor (hereafter *Tb1-ins*). *Tb1-ins* enhances *tb1* expression that confers apical dominance to maize (Studer, Zhao et al. 2011). It is present at low frequencies in teosintes and therefore predates maize domestication (Studer, Zhao et al. 2011). Intriguingly, while *Tb1-ins* has a drastic effect when inserted into a maize inbred background (Lukens and Doebley 1999), it has been reported that no measurable effect on tillering were observed in a sole teosinte population grown in artificial conditions (greenhouse) (Vann, Kono et al. 2015). Another TE insertion with notable effects in maize is a 143 bp miniature inverted-repeat transposable element (MITE) found in the *Vegetative to generative transition 1* (*Vgt1*) regulatory region (hereafter *Vgt1-ins*). *Vgt1* cis-regulates the *Ap2*-like transcription factor *ZmRap2.7* localized 70kb downstream (Salvi, Sponza et al. 2007). Maize plants carrying *Vgt1-ins* display lower *ZmRap2.7* transcription and early flowering (Salvi, Sponza et al. 2007). Polymorphism of absence/presence of *Vgt1-ins* associates with flowering time variation in maize landraces supporting its role in altitudinal and latitudinal adaptation (Ducrocq, Madur et al. 2008). Although flowering time is a highly complex trait with other QTLs having been found to associate with earlier flowering time in northern latitude maize (Salvi, Corneti et al. 2011), it has been shown that the heavy stable methylation found on MITE insertion at the *Vgt1* locus likely affects *ZmRap2.7* transcription abundances (Castelletti, Tuberosa et al. 2014). Finally, a CACTA-like transposable element inserted in the promoter of a maize CCT-domain (*CO, CO-LIKE and TIMING OF CAB1*-domain) containing gene (hereafter *ZmCCT-ins*) was found in temperate maize where it confers adaptation to long-day length by attenuating photoperiod sensitivity. *ZmCCT-ins* suppresses the expression of the *ZmCCT* gene under long days which provokes an up-regulation of the *Zea centroradialis8* (*ZCN8*) floral activator (Yang, Li et al. 2013). Interestingly, the absence of *ZmCCT-ins* in a sample of 12 teosinte accessions suggests that this insertion had occurred after maize domestication (Yang, Li et al. 2013).

Because TEs are present in multiple, and often truncated copies, their discovery from short-read re-sequencing is a daunting task (Rech, Bogaerts-Márquez et al. 2019). Recent years have seen the development of a flurry of bioinformatics tools for detecting polymorphism of TE insertions

(reviewed in (Goerner-Potvin and Bourque 2018). Most of these use repositories of TE sequences, which are either built from a vast range of organisms or contain species-specific TEs only (Goerner-Potvin and Bourque 2018). They typically combine mapping information of short-reads to TE repositories and to genome assemblies masked for TEs. Successful applications have uncovered evidence of positive selection at insertions close to genes involved in response to stress, behavior and development in the model species *Drosophila melanogaster* (Rech, Bogaerts-Márquez et al. 2019). Likewise, detection of insertions absent from the reference genome in 28 *D. melanogaster* European populations, has pointed to 17 insertions with repeatable correlations between allele frequencies and geographical/temporal variables across the European and American continent (Lerat, Goubert et al. 2019). Frequency patterns of TEs in the Asian tiger mosquito have also pointed to their adaptive role in the recent colonization of temperate environments (Goubert, Henri et al. 2017). Application of these tools on much larger and complex plant genomes however poses practical and conceptual challenges.

In the present study, we adapted existing pipelines to characterize TE content and frequencies from pooled sequencing data of teosinte populations. We characterized the TE content and polymorphism by presence-absence of insertions that were either present (*reference*) or absent (*de novo*) from the B73 maize reference genome in four teosinte populations, two lowlands from the subspecies *Zea mays* ssp. *parviglumis* and two highlands *Zea mays* ssp. *mexicana*. We addressed three main questions: How does TE content differ among populations? Can we identify candidate insertions for altitudinal adaptation? What is the geographical pattern of variation and phenotypic effects of the insertions *Tb1-ins*, *Vgt1-ins*, and *ZmCCT-ins* in teosinte populations?

III.2 Material and methods

III.2.1 Plant material and sequencing

We used whole genome paired-end sequencing data (2 x 100 bp) of pooled individuals from four teosinte populations (Fustier, Brandenburg et al. 2017). Pools consisted of 20 individuals per population. These populations represented elevation extremes with two lowland populations from the subspecies *Zea mays* ssp. *parviglumis* (thereafter *parviglumis*) and two highland populations from the subspecies ssp. *mexicana* (thereafter *mexicana*) (Supp. Table S1). We used phenotypic (1125 plants) and neutral SSR (1664) genotyping data for an association panel comprising eleven teosinte populations as described in Fustier *et al*. (2019). This association panel was previously

evaluated for 18 phenotypic traits in two common gardens at two mid-elevation locations during two consecutive years. For each population, half-sib seeds collected from the eleven populations were sown in a four-block randomized design in each location and year. The 18 phenotypic traits measured included plant architecture, reproduction and physiology (Supp. Table S2). We used available DNAs from additional populations sampled along two elevation gradients (Fustier, Martínez-Ainsworth et al. 2019) to genotype TE insertions (see below) for a total of 17 populations and 20 individuals per population (Supp. Table S1).

III.2.2 Estimating content and frequency of *reference* insertions

As a first approach, we characterized in our sample of four populations, the insertion polymorphisms of transposable elements (TEs) insertions *present* in the B73 maize reference genome, that we refer to as *reference* insertions. To this purpose, we used PopoolationTE2 (Kofler, Gómez-Sánchez et al. 2016), a pipeline designed to handle pooled sequencing data to estimate population frequency of *reference* insertions. Briefly, PopoolationTE2 determines TE insertion frequencies from resequencing data by identifying and then quantifying at any given TE insertion point (1) the presence of TE insertion from read-pairs for which one read uniquely aligns to the non-TE region flanking the insertion point in the reference genome, while the other read aligns to a TE sequence from a TE annotation database; and (2) the absence of TE insertion from read-pairs mapping at a distance predicted in the absence of the TE insertion in the reference genome. PopoolationTE2 therefore relies on the parallel use of a reference genome sequence masked for TE insertions, and a TE annotation database.

The B73 v4 genome sequence was masked for TEs using the annotation file provided by Michelle Stitzer. As for TE database, we used intact TE sequences (i.e. catalog of TE sequences of all insertions found in the genome) from B73 v4 maize TE annotation database (Jiao, Peluso et al. 2017) that was recently updated (Stitzer, Anderson et al. 2019). The TE database included 341,241 elements, with retrotransposons strongly represented by LTR-retrotransposons (42%) in contrast to non-LTR retrotransposons (0.4%) and DNA transposons mainly represented by TIR transposons (51%) and to a lesser extent Helitrons (6.6%), Table 1 in (Stitzer, Anderson et al. 2019). The TE database encompassed 13 superfamilies and 27,301 families with highly variable number of copies among them. Superfamilies are listed with their common name and the number of families they harbor shown in parentheses, for DNA transposons: DHH or Helitrons (1,722), DTA or hAT (275), DTC or CACTA (73), DTH or Pif/Harbinger (358), DTM or Mutator (67), DTT or Tc1/Mariner (269), DTX or Unknown TIR (76), and for retrotransposons: RLC or Copia/Ty1 (2,788), RLG or

Gypsy/Ty3 (7,719), RLX or Unknown LTR (13,290), RIT or RTE (2), RIL or L1 (29) and RST or SINE (533). Sequencing reads of each population were aligned independently to the masked genome sequence and to the TE annotation database using the *bwa* aligner with default parameters of option *bwasw* (Li, Ruan et al. 2008)which uses the Smith-Waterman algorithm allowing partial mapping of reads which is adequate to include reads that may be spanning a TE insertion site. Mapped reads were restored into pairs using the *se2pe* function of PopoolationTE2. From read pairs, we generated a physical pileup (ppileup) file that summarized for every site in the genome, absence or presence of insertions of each individual TE copy. When generating the ppileup we used the option *homogenize-pairs* which enabled the use of identical number of mapped paired-end reads for all samples by subsampling the smallest number of informative pairs among the samples.

We estimated TE insertion frequencies for a curated set of *reference* TE insertions by combining the functions *identifySignatures*, *frequency* and *pairupSignatures* functions with default parameters. We recovered these signatures using the *joint mode* to estimate frequency of all insertions for which there was sufficient coverage to determine presence/absence in the four teosinte populations. Only TE insertions that presented both forward and reverse insertion signatures were kept for further analysis. Because we determined frequencies from individual TE copies, we imposed that the coordinates of the recovered TE insertions (from the TE annotation) were comprised between the start and stop positions of the corresponding TE insertions on the maize reference genome.

From TE frequencies of *reference* insertions, we determined the relative proportion of TE superfamilies and families in the four teosinte populations. We also determined subset of TE insertions from the maize reference genome that we used to interrogate teosinte genomes. Indeed, PopoolationTE2 was originally designed for *Drosophila melanogaster*, a genome with relatively poor TE content, and low amount of nested insertions. It therefore focuses on identifying single (i.e. non-nested) TE insertions. In the TE-rich maize genome where large blocks of nested TEs are the rule rather than the exception, PopoolationTE2 retrieved non-nested TEs as well as the outermost TEs of these blocks (thereafter nonTE-flanked TEs).

III.2.3 Discovery and frequency estimate of *de novo* TE insertions

As a second and complementary approach, we investigated frequencies at TE insertions not present in the maize reference genome. We refer to these insertions as *de novo* insertions relative to the reference genome. Note that we inferred *de novo* insertions from the maize TE annotation database, thus allowing for interrogating teosinte new TE insertions of known B73 TEs, but not

insertions from new non-B73 TEs. To do so, we used *T-lex-de-novo*, a software that searches for TE insertions in alternative locations than those from a reference genome (Kelley, Peyton et al. 2014). *T-lex-de-novo* uses read-pair information, alignment to a reference genomic sequence and alignment to a TE database to retrieve two sources of evidence of TE insertions (Supp. Figure S1): (i) One End Anchored (OEA) read evidence, where one read aligns over its entire length to the reference genome, while its mate aligns to a sequence of the TE annotation database; (ii) Clipped Read (CR) evidence where part of one read aligns to the reference genome and the soft-clipped part of that read together with its mate align to a sequence of the TE database. To recover reliable *de novo* TE insertions in each population separately, we devised a number of stringent filters (Supp. Figure S1). First, we retained insertions supported by at least two independent read-pairs recovering the exact same insertion point. Second, we discarded *de novo* insertions separated by less than 150bp, considering that we did not have the power to distinguish between one or two independent insertions at this distance. From this set of insertions, we determined TE content in each teosinte population and compared the relative proportion of TE superfamilies and families in across populations. We also computed genomic landscapes of *de novo* insertions for each superfamily using 100kb bins along each of the 10 maize chromosomes.

We further retrieved insertions that were present in all four populations. We restricted population frequency estimates of *de novo* insertions to CR evidence only as it provides an insertion point (Supp. Figure S1). Because the insertion point may vary slightly from one population to another due to small insertion-deletions, we defined an insertion zone (±20 bp around the insertion point that corresponds to the detection resolution in *T-lex-de-novo*). We required both, clipped reads to have at least 5 bp mapped to the reference genome (with no insertions or deletions), and traversing reads to overlap the insertion point over at least 10 bp with no insertions, deletions or softly clipped edges (Figure. S1). In addition, following Fustier et al. (2017), we required that the local depth within a 100 bp-window surrounding the insertion point ranged between 12 and 50 reads (12x-50x depth). We finally estimated insertion frequencies for this curated set of *de novo* insertions. For each TE insertion, the frequency was the ratio of the number of clipped reads in the insertion zone divided by the sum of the number of clipped- and twice the number of traversing-reads spanning the insertion zone (Figure. S1). Because clipped reads may indicate presence on both sides of the insertion point, we counted both reads of pairs encompassing one traversing read.

III.2.4 Detection and genotyping of candidate TE insertions.

We aimed at recovering TE candidate insertions for altitudinal adaptation. We therefore seek TE insertions whose frequency where highly differentiated between the lowland and highland population of each gradient. We computed pairwise F_{ST} values between the lowland and highland populations of each gradient both for *reference* and *de novo* TE insertions. We retrieved candidate insertions by selecting the 5% highest F_{ST} values in both gradients, whose frequencies changed in the same direction (increase or decrease) in the two gradients. We inspected the genomic context of candidate insertions: distance and identity of the closest upstream and downstream genes.

We used three well-characterized maize insertions previously described in the literature for their phenotypic effects, to perform PCR-assays both on the entire sample of 17 populations and on the association mapping panel developed by Fustier et al., (2019) (Fustier, Martínez-Ainsworth et al. 2019). These three maize insertions are: the *Hopscotch* insertion into the *cis*-regulatory region of the *tb1* gene (*Tb1-ins*); the MITE insertion into a conserved non-coding sequence of the *Vgt1* locus, that regulates the *Rap2.7* gene (*Vgt1-ins*); and the CACTA-like insertion into the promoter of the *ZmCCT* gene (*ZmCCT-ins*). These insertions affect branching, flowering time and photoperiod sensitivity, respectively.

For PCR assays, DNA was extracted from leaf tissue (Fustier et al. 2017) and was PCR-amplified in 20 µl reaction mix containing 1X Taq buffer (Promega), 0.2 mM of dNTP, 0.8 µM of each primer, 2 units of home made Taq polymerase and additional $MgCl_2$ at various concentrations (Supp. Table S3). We used previously published primers to genotype insertions at *Tb1*, *Vgt1* and *ZmCCT* (Salvi, Sponza et al. 2007; Yang, Li et al. 2013; Vann, Kono et al. 2015), with minor modifications for *Vgt1* including a home-designed reverse primer (Supp. Table S3). For the short MITE we used two primers, located on each side of the TE, whereas longer CACTA and Copia TEs required a combination of three primers (located on each side of the TE, plus one located inside the TE sequence) to verify the presence/absence of the insertion while controlling for PCR failure. We used the following general conditions for amplification with varying number of cycles (N), annealing temperature (Tm), duration (Dex) and temperature (Tex) extension: 94°C for 4 min, N cycles of 94°C for 30 s, Tm°C for 30 s, and Text °C for Dex, with a final extension of 72 °C for 10 min. PCR products were visualized on a 1% agarose gel and scored for presence/absence of insertions based on band size. Primers, detailed protocols ($MgCl_2$, Tm, number of cycles, Tex, Dex), expected bands and sizes for presence/absence of insertions are presented in Supp. Table S4.

III.2.5 Geographical distribution of candidate TE insertions, and association with environment and phenotypes.

We obtained frequency estimates of the maize TE insertions (*Tb1-ins, Vgt1-ins* and *ZmCCT-ins*) for 31 populations (Supp. Table S1). We plotted homozygote and heterozygote frequencies of *Tb1-ins* and *Vgt1-ins* maize TE insertions for the 11 common garden populations on a geographical map (Figure 6-A, Supp. Figure S8-A), as well as a scatter plot of the population frequencies of the insertion for all 11 populations genotyped (Figure 6-B, Supp. Figure S8-B). To investigate co-variation of insertion frequencies with environmental variables, we used BayEnv2 (Coop, Witonsky et al. 2010). We used a covariance matrix of relatedness between 28 populations previously computed from SNP data as well as environmental data summarized in the form of the first principle component on the same set of populations (Fustier, Martínez-Ainsworth et al. 2019) as well as for each of the 19 climatic layers separately (Cuervo-Robayo, Téllez-Valdés et al. 2014). For each maize insertion, we tested whether its frequency among populations was determined primarily by the covariance matrix, or by a combination of the covariance matrix and the principal component best summarized by altitude or alternatively one of the 19 climatic layers.

To test association between genotyped maize TE insertions and phenotypic measurements, we performed association mapping analyzes following the restricted maximum likelihood mixed model proposed in Fustier et al., (2019). The model controls for neutral genetic structuring using downstream analyzes (genetic cluster assignment and kinship matrices) from 38 genotyped simple sequence repeat (SSR) (Fustier, Martínez-Ainsworth et al. 2019) for the same association panel. Maize TE insertion genotypes were coded as homozygous for presence, homozygous for absence or heterozygous. The model employed describes each observed phenotypic measurement as the response variable Y explained by a series of fixed and random factors as,

$$Y_{ijklom} = \mu + \alpha_i + \beta_j + \theta_{ij} + \gamma_{k/ij} + \sum^{4}_{n=1} b_n \cdot C^n_{ijkolm} + \zeta_o + u_{ijklm} + \varepsilon_{ijklom} \qquad \text{(M1)}$$

where μ is the total mean, α_i is the fixed year effect of the experiments (i = 2013, 2014), β_j is the fixed experimental field effect (j being the experimental location, *SENGUA*, *CEBAJ)*, θ_{ij} is the year by field interaction, $\gamma_{k/ij}$ is the fixed block effect (k = 1, 2, 3, 4) nested within the year-by-field combination, b_n is the fixed effect of the structure covariate C^n (n = number of STRUCTURE groups -1) with membership values for the C^n covariates calculated at the individual level, ζ is the fixed TE insertion presence/absence factor effect with one level for each of the three genotypes (o=1, 2, 1; with o=2 for heterozygous individuals), u_{ijklom} is the random genetic effect of each individual and ε_{ijklm} is the individual residue. We assumed that the vector of u_{ijklom} followed a (0,K

σ^2u) distribution, where K is the inversed kinship matrix. A second version of the model exchanged the fixed effect structure covariate for the strict belonging of each individual to its sampled population of origin, thus 11 populations. Both models were tested with ASReml v.3.0 (Butler, Cullis et al. 2007) software package in R and run for each TE independently with each fixed effect tested through a Wald test. For each phenotype, significant effects of the maize TE insertions were obtained from their Wald statistics p-values.

III.3 RESULTS

III.3.1 TE content across teosinte populations

In order to characterize TE content in four teosinte populations, we used two different tools: one that discovered a subset of *reference* insertions present in the maize reference genome, and one that discovered *de novo* insertions absent from the reference genome, yet characterized. Those tools provide different information that are hardly comparable. On one hand, PopoolationTE2 interrogated only the subset of nonTE-flanked *reference* insertions for which all locations were covered by reads in all four populations. The superfamily and family relative proportions of nonTE-flanked *reference* insertions differed from the maize TE genome-wide content (Supp. Figure S2). On the other hand, T-lex-de-novo detected genomic insertions flanked by low-copy DNA and absent from the reference genome. Per population, we retrieved between 17,173 and 17,724 *reference* insertions (Supp. Figure S3), and between 198,869 and 277,617 *de novo* insertions (Figure 1). Because we did not filter for coverage of locations in all populations for *de novo* insertions, we recovered a much greater proportion of them.

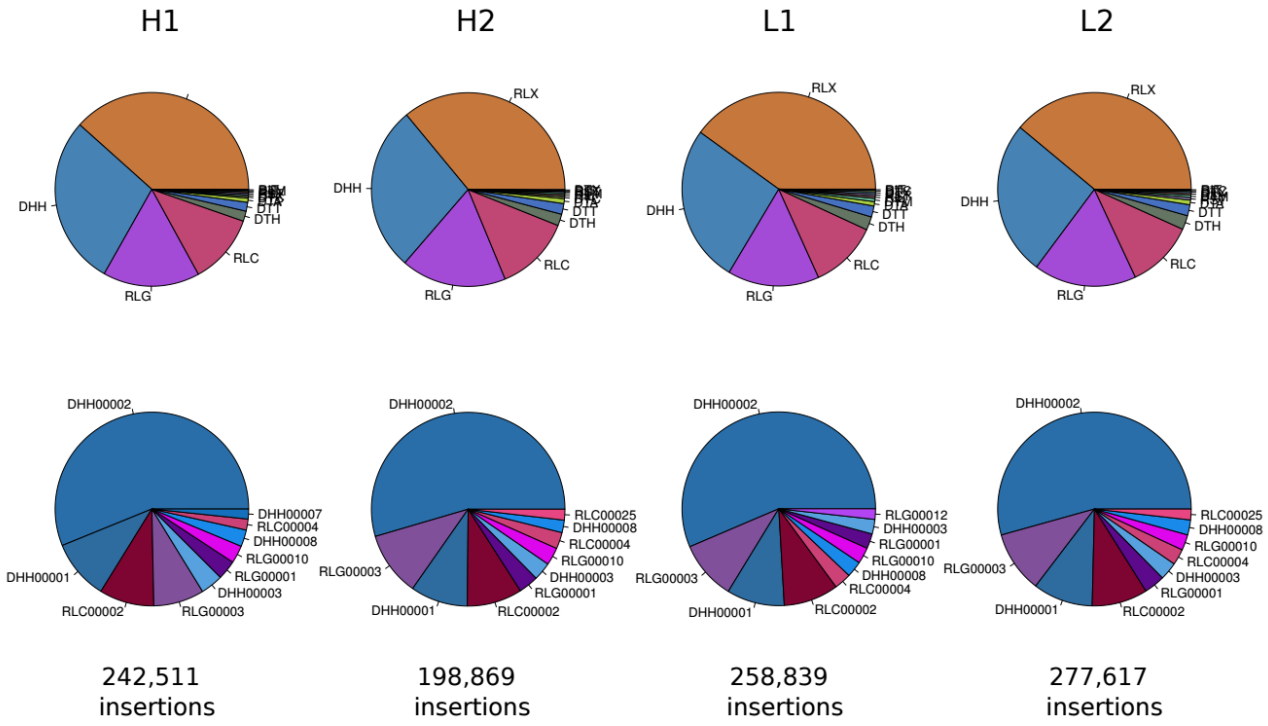


Fig 1: Superfamily and ten most abundant family pie charts for *de novo* TE insertions in four HTS populations. Total number of elements found per population are indicated below each pair of pie charts.

Both *reference* and *de novo* revealed similar profiles of relative proportions at the superfamily and family level among the four teosinte populations (Supp. Figure S3 and Figure 1). We confirmed this pattern at the family level, with highly significant population pairwise correlation of TE abundance (r>0.99) across both *reference* and *de novo* families (Supp. Figure S4 and Supp. Figure S5). Interestingly, we found an enrichment of DNA elements discovered both at the superfamily and family level for *reference* insertions (Fig S3), when compared to the *reference* insertions surveyed from the maize genome (nonTE-flanked insertions, Supp. Figure S2-B). There was indeed a greater proportion of DTT and DTH superfamilies and a smaller proportion of RLC. In fact, the patterns of abundance of superfamilies resembled more the one obtained when extracting from the maize insertions surveyed, the ones that were single (Supp. Figure S2-D) as opposed to the ones that had elements nested within them (Supp. Figure S2-C). As for *de novo* insertions, we observed an opposite pattern with greater proportion of retroelements (Figure 1) with respect to the elements therein searched (Supp. Figure S2-A).

We described genomic landscapes of *de novo* insertions for the four most abundant superfamilies. The patterns differed among superfamilies, with RLG present uniformly along the genome, RLC and DHH exhibiting an insertion/retention landscape consistent with depletion in the pericentromeric regions in contrast to RLX where pericentromeric regions displayed a higher abundance of insertions (Figure 2 and Supp. Figure S6). Altogether these genomic patterns further confirmed similarities across populations.

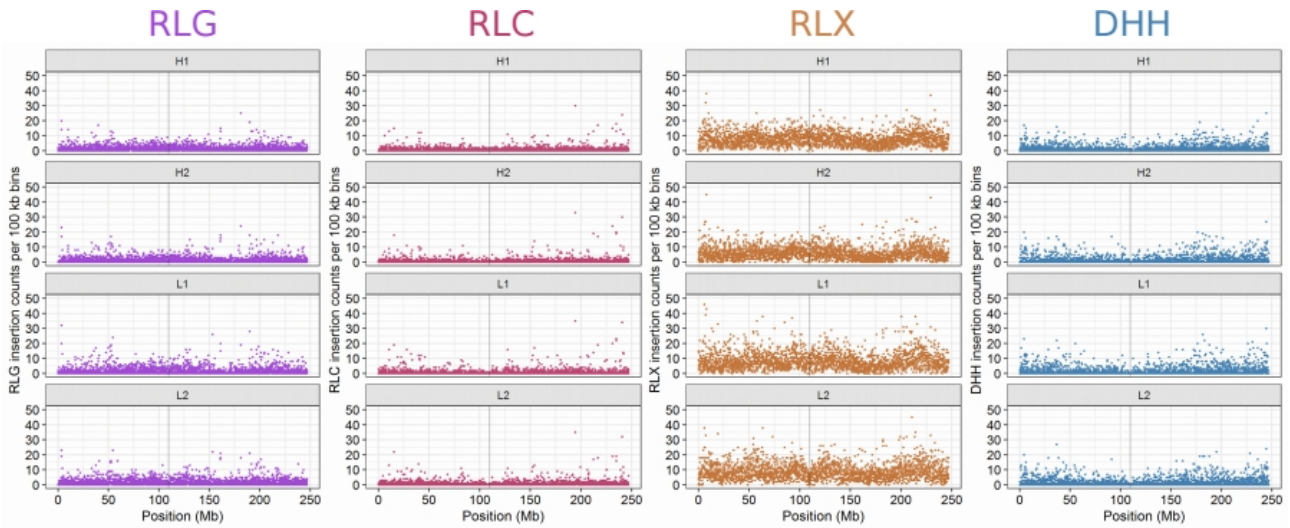


Fig. 2: Genomic landscape of *de novo* insertions of all superfamilies along chromosome 4. Each point represents the amount of TE insertions present in 100kb bins along each chromosome.

III.3.2 Selection of candidate insertions

Prior to determine insertion frequencies, we obtained curated sets of *reference* and *de novo* insertions. Our set of curated *reference* insertions encompassed 18,127 elements for which we described patterns of shared and unique insertions across the four populations (Figure 3B). A majority of them (15,976) were found in all four populations while insertions unique to a population represented a small fraction (0.003%). When discarding insertions common to all four populations, we observed that the two highland populations had more insertions in common (1,156) than either of them with the two lowlands (H1-L1: 993, H1-L2: 655, H2-L1: 1000, H2-L2: 711) or the two lowlands between them (720). This observation was in line with the greater genetic proximity of the two highlands as previously assessed using a set of 1000 neutral single nucleotide polymorphisms (Figure 3A). As for the *de novo* insertions, our filters reduced the curated dataset that we used to estimate population frequency to 1,818 insertions. Because we filtered on presence of insertions in all four populations as a first step (OEA+CR evidence), the pattern of shared versus unique insertions could not be interpreted in terms of genetic proximity.

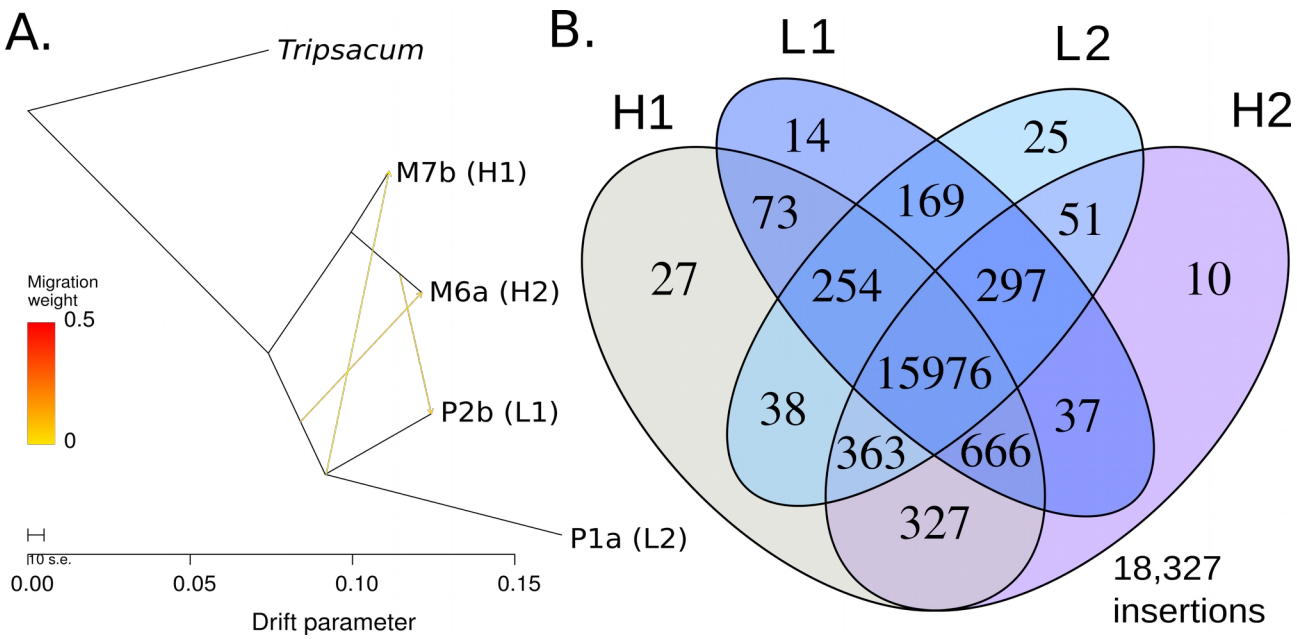


Fig. 3: Population ancestral graph among HTS populations (A) and number of *reference* insertions unique or shared among teosinte populations (B). The inference of population history(A) was obtained with though a TreeMix (v.1.11) analyzes on 19,000 SNPs of the MaizeSNP50 Genotyping BeadChip data (Aguirre-Liguori, Tenaillon et al. 2017) for these populations which were shared with the *Tripsacum dactyloides* outgroup. Yellow arrows indicate evidence for shallow gene flow between some of the populations. The venn diagram (B) was constructed for a total of 18,127 insertions whose coordinates were covered in all four populations.

From frequencies of curated TEs, we identified 120 *reference* and eight *de novo* candidate insertions (Supp. Table S3). The number of *reference* insertions therefore exceeded what was expected (0.0025 x 18,327 ≈ 38). Among *reference* insertions present in the 5% tail of highest F_{ST} values between lowland and highland populations of the two gradients, we observed an enrichment for insertions with the same directionality (120 among 152, Figure 4A) that is, insertions whose frequency increased/decreased along elevation in both gradients (χ^2=119.89, p-value<2.2 10^{-16}). On the contrary, the number of candidate *de novo* insertions was close to expectation (0.0025 x 1838=5), and we found no specific enrichment for insertions sharing the same directionality within the 5% outliers (χ^2=1.007, p-value=0.3154) (Supp. Figure S8). These observations suggested that the set of *reference* candidate insertions was likely to contain true positives.

For the set of *reference* candidate insertions, we further investigated the relative contribution of different superfamilies with respect to the curated *reference* elements set from where they were taken. To do so, we grouped the less numerous superfamilies together (DTX, RST, RIT, RIL, DTC, DTM and DTA). Interestingly, we found a traceable superfamily influence (χ^2= 47.975 p-value=1.195 10^{-08}) with noticeably more RLG and DTH but less DTT and RLX elements among *reference* candidate insertions than expected (Figure 4B). We next inquired whether *reference* candidate insertions were found more often among TEs annotated as single or those which

contained at least one nested element or fragment. We observed that our *reference* candidates (Figure 4C) were more often single (χ^2=12.137, p-value=4.94 10^{-4}) (Supp. Table S3). Moreover, besides eleven *reference* candidates falling inside genes, we detected a highly skewed distribution towards *reference* candidates inserting in the 5' of genes when compared to all reference insertions for which F_{ST} values were available (χ^2 = 4.679, p-value = 0.03). Note that the distance to genes was also slightly smaller for *reference* candidates than for all other *reference* insertions from the scatterplot (Figure 4D).

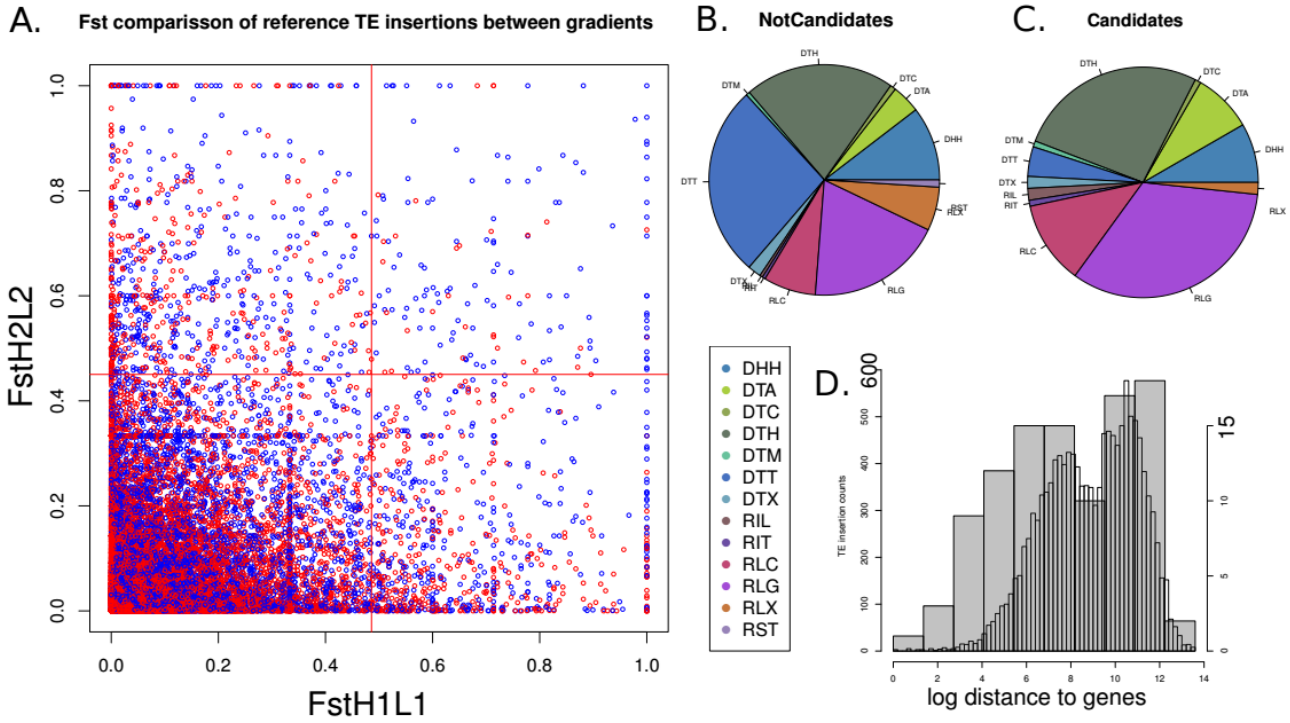


Fig. 4: *Reference* insertions scatterplot of pairwise F_{st} values per gradient (H1L1/H2L2) (A), with pie charts depicting superfamily occupation by *reference* non-candidate insertions (B) and candidate insertions (C) and overlapping histograms of log-distance (bp) to the closet gene for candidate and non-candidate insertions. The F_{st} scatterplot shows the 5% outlier threshold for each gradient (red lines) and color of dots indicate parallel (blue) or opposite (red) frequency clines in the two highland-lowland comparisons. *Reference* candidate insertions (120) are the blue dots in the upper right square. All other dots of the scatterplot were considered as our set of non-candidate insertions (18,207). Histogram for candidate insertions is shown with gray bars with counts indicated on the right y-axis, and histogram for non-candidate insertions is shown in white bars with counts along the left y-axis.

III.3.3 Association mapping

From screening efforts of three maize TE insertions, only two were effectively found in our teosinte populations. *Tb1-ins* was present at low frequency among 31 populations ranging between 0 and 16% with two exceptions of *parviglumis* lowland populations on gradient A (P1a and P3a) reaching 33% and 38% presence respectively (Table 1; Supp. Figure S9-C). Among the 11 populations of the association panel, one *parviglumis* population from gradient A (P10a) as well as

one from gradient B (P2b) showed considerable presence of this insertion at heterozygous state (Supp. Figure S9-A). *Vgt1-ins* exhibited higher frequency along gradient B (P9b, P8b and P1b) albeit important also at gradient A (P9a, M7a and P9a) (Table 1; Figure 5-D). Frequencies among the 31 screened populations were highly variable, reaching values of up to 64% with no clear altitudinal pattern (Figure 5-D). The association panel displayed notable proportion of *Vgt1-ins* at homozygous state among intermediate altitude populations with slightly more moderate values in highland populations (Figure 5-A). Average day-length differed by only less than an hour between the northernmost and southernmost of the 11 populations of the association panel, taking as example the growing season of the year 2013 at which common gardens were grown (Figure 5-C). Finally, *ZmCCT-ins* was absent from all 11 teosinte populations, whilst we consistently found it in maize B73 controls.

In order to perform a Bayesian estimation of the correlation between insertion frequencies and environmental variables taking into account genetic correlation among 22 populations, we employed a 5% threshold calculated on 1000 neutral SNPs genotyped for the same populations. This stringent threshold rendered no detectable association of neither *Tb1-ins* nor *Vgt1-ins* to the first principal component. Instead, *Vgt1-ins* was found associated to three environmental layers when tested independently: bios15 (precipitation seasonality), bios02 (mean diurnal range) and bios06 (minimum temperature of the coldest month)

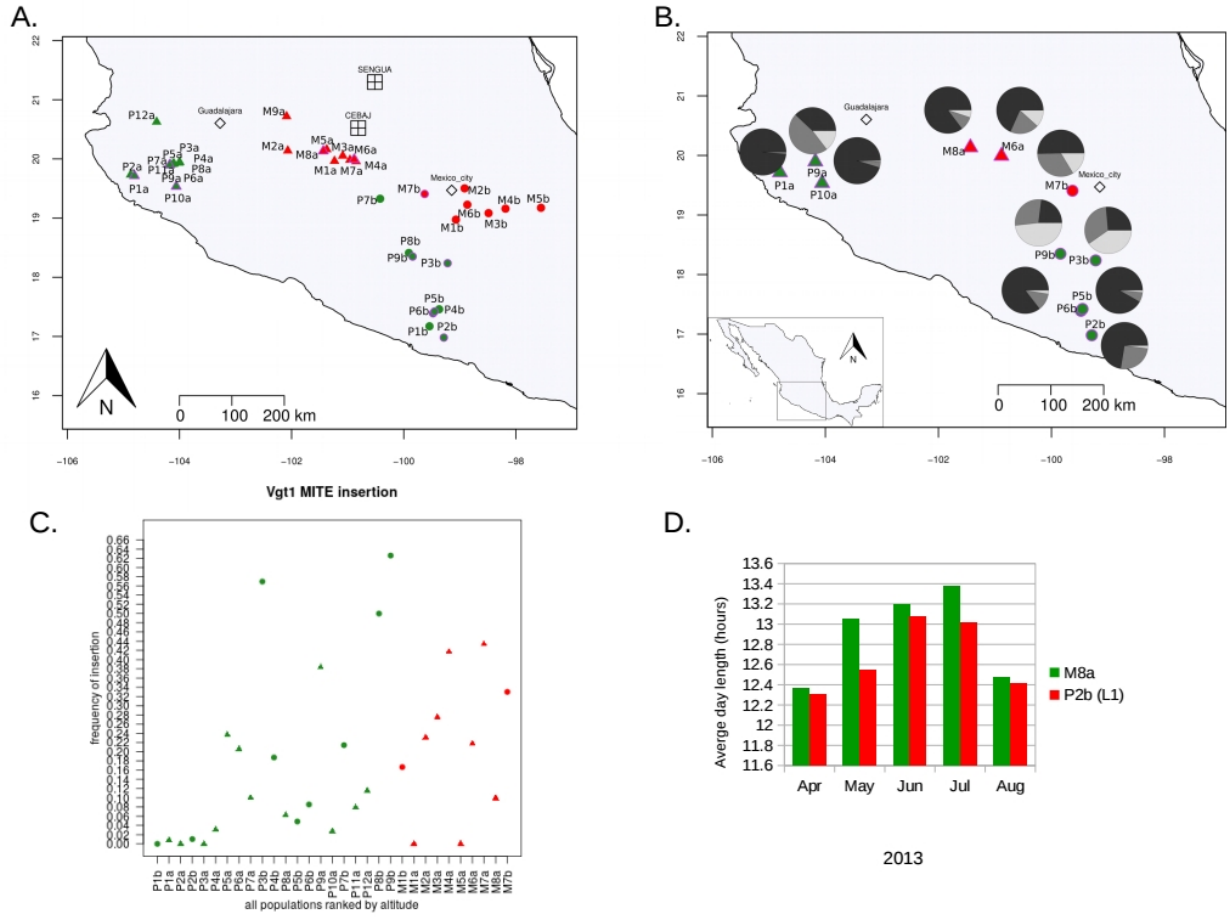


Fig. 5: Geographic localization of the entire set of 11 teosinte populations (A), geographical distribution of *Vgt1-ins* frequencies over the association panel (B) frequencies of Vgt1-ins for all 31 populations ranked by altitude (C), and average day-length from April to August 2013 for the northernmost and southernmost populations of the association panel (D). Colors in A indicate subspecies (*parviglumis*=green, *mexicana*=red) and shapes relate to gradients (gradient 1 = circles, gradient 2 = triangles). Pie plots in B indicate the proportion of individuals homozygotes for presence (light gray), absence (black) and heterozygotes (mild gray).

Finally, when correcting for neutral structure on 11 populations in the genetic association mixed model, we found that *Vgt1-ins* was strongly associated to male flowering time, and to a lesser yet significant extent (<0.05) to female flowering time, leaf width and grain coloration (Supp. Table S5). Interestingly, two different traits were recovered with the five genetic groups correction: number of tillers and grain length (Supp. Table S5). We found that *Tb1-ins* strongly associates to female flowering time, plant height, height of the lowest ear, and less strongly to male flowering time for both the five group and the 11 population models. Additionally, this polymorphism also associated to leaf length and number of grains with the five groups model, whereas to grain length, grain weight and stomata density with the 11 populations model (Supp. Table S5). We further examined models for *Tb1-ins* including a genotype by population interaction, and found that with the *k=5* model all but two traits had a significant interaction, which was reduced to five traits under

the 11 populations model, among them female and male flowering time, while conserving all previously significantly associated traits plus number of grains.

III.4 DISCUSSION

Transposable elements (TEs) are a conspicuous feature of plant genomes (Lisch and Slotkin 2011). While TE content has been described in many crops for which reference genomes are available (reviewed in (Vitte, Fustier et al. 2014), much less is known about their wild relatives. Crops have derived recently from their wild relatives, and most TE families, for instance in maize, are inactive (Feschotte, Jiang et al. 2002). We therefore expect no recent TE bursts nor major difference in TE content between wild and domesticated forms. At the population level, however, domestication may have affected TE frequencies through domestication bottlenecks and selection, as suggested ifor sunflower (*Helianthus annuus*) (Mascagni, Barghini et al. 2015). How different is TE content across wild populations and how does it compare with crop TE content? How powerful are current bio-informatic tools to screen TE insertions and detect plausible candidates from population resequencing data? Are the adaptive insertions detected in crops also contributing to trait variation in wild populations? These are some of the questions that we addressed in the closest wild relatives of maize, the teosintes.

III.4.1 TE content does not differ among teosinte populations.

In order to decipher TE content, we employed two tools, one that detected *reference* insertions present in the maize genome, and one that detected *de novo* insertions absent in the maize genome, albeit described in the repertoire of maize TEs. The four maize lines B73, W22, Mo17 and PH207 (Anderson, Stitzer et al. 2019) show comparable genomic structural annotations and similar global TE contents. Likewise, we found similar *reference* TE contents amongst the four teosinte populations (Supp. Figure S3). As expected, pattern of population genetic proximity revealed by TEs was similar to the one described for neutral SNPs (Fustier et al 2017), with both *mexicana* populations (H1 and H2) showing higher genetic proximity (Figure 3-A), thus sharing more common *reference* insertions than with either *parviglumis* populations (Figure 3-B). Patterns of the relative contribution of superfamilies and families corresponded to the ones described for typable single insertions in the maize reference genome (Supp. Figure S2-C). This observation indicated that discovery of *reference* insertions was strongly biased towards the detection of DNA elements, which are less abundant in plant genomes, albeit more often found as single elements when compared with RNA elements (Supp. Figure S2-C,D).

As opposed to *reference* insertions, *de novo* insertions revealed a TE content that differed markedly from patterns described for the maize genome. RLX and DHH indeed occupied the majority of all elements (Figure 1, Figure S2-A). However, landscapes of *de novo* insertions along the chromosomes corroborated patterns found in maize, indicating a robust detection. For instance, we found three general patterns that comply with the superfamily-level landscapes described for the B73 v.4 TE annotation (Stitzer, Anderson et al. 2019). These are (1) an enrichment of DHH insertions towards chromosome arms; (2) pericentromeric reduction for RLC elements; and (3) presence of RLG elements at pericentromeric regions (Figure 2, Supp. Figure S6). Our landscapes pertaining RLX reflected the higher dynamism of these elements as registered by chromosome, with for instance high pericentromeric yet low centromeric occupancy along chromosome 4 (Figure 2) unlike the rather flat distribution observed for chromosome 1. On chromosome 4 we observed an additional increment of RLX along the right end was not quite as clear in B73 v.4 TE annotation. Since it has been observed that within TE superfamilies, each family can vary greatly in their profiles, we inquired the reference chromosomal distribution of the five-top annotated RLX families on chromosome 4 yet found only moderate support for our chromosome 4 swelling. This could arguably reflect the fact that RLX elements are not strongly dominated by any family, with elements we tested ranging only up 269 copies and many families (11,418) represented by only one element, in contrast to RLG for example, where the most abundant families contribute 15,303 and 12,093 elements and only 1,800 families are unitary.

As for the biases towards detecting more RLX and DHH in *de novo* insertions, this may have several non-exclusive origins. These two superfamilies contain rather long elements with old insertion age (Stitzer, Anderson et al. 2019) which could in principal make them easier to detect when filtering for shared and sufficiently frequent elements between maize and its wild ancestor. Unclassified LTR retrotransposons (RLX) superfamily is composed by highly divergent elements, many of which are non coding or truncated (Baucom, Estill et al. 2009), possibly making it easier to pinpoint through the blating algorithm we employed, not so likely for highly similar and repetitive copies inherited from RLG for example. In addition, RLX are known to be enriched in maize LTR retrotransposon methylation spreading (Baucom, Estill et al. 2009). As for helitrons (DHH), these elements have been reported to vary greatly among individual maize plants (Messing and Dooner 2006), thereby perhaps boosting the per population identification of these elements. Also, helitrons have been found to responsible of collinearity shuffling at the maize *bz* locus (Lai, Li et al. 2005), so if such behavior indeed occurs genome-wide it may help explain the enrichment of these elements found for *de novo* positions. Helitrons may carry fragments of genes and can sometimes

produce new combined transcripts, in maize non-autonomous helitrons have been found to contain coding sequence's from different host genes (Lai, Li et al. 2005; Morgante, Brunner et al. 2005) thus participating in gene innovation generating variation that could in turn have been selected. Thus, prodding for adaptive candidates in a helitron-enriched sample of *de novo* TE insertions could likely find new genes or functions.

Even with detection biases, it is to note that all four teosinte populations displayed very similar TE content across families and superfamilies. There are two important consequences of this observation. The first one is that although our populations differs in genome size (H1=6.710, L1=5.991, H2= 6.249, L2 =6.932, estimates from five plants per populations taken from Munoz-Diez et al. (2013) (Diez, Gaut et al. 2013), TE content did not seem to account for those differences. This corroborates previous results showing that chromosomal knobs rather than TEs are the primary determinants of genome size difference within *Zea mays* (Chia, Song et al. 2012; Diez, Meca et al. 2014) (Bilinski, Albert et al. 2018). The second one is that *parviglumis* and *mexicana* share similar TE content, respectively to maize TE insertions. These two subspecies have diverged around 60,000 years ago (Ross-Ibarra, Tenaillon et al. 2009), and most TE insertions may therefore predate that divergence. It is possible, however, that these two species differ from one another at families that have inserted new copies since their divergence yet these families may not have been selected during maize domestication or have not yet been characterized in maize.

III.4.2 Candidate insertions insert more often 5’ of genes.

We attempted to identify both *reference* and *de novo* candidate insertions, with special emphasis on spatially varying selection pressures that could generate polymorphic patterns of positively selected yet not species-fixed TEs (González, Karasov et al. 2010). Because detection tools and subsequent filters necessary to establish curated sets of insertions suffered from strong biases, we were not able to establish a site occupancy frequency spectrum for TE insertions. Such spectrum is informative to estimate the strength of selection acting against TE insertions as has been shown in *Arabidopsis* (Hazzouri, Mohajer et al. 2008; Lockton and Gaut 2010) and *Capsella grandiflora* (Horvath and Slotte 2017) and have even been used to detect insertions undergoing positive selection in *Drosophila* (González, Lenkov et al. 2008). Alternatively, empirical distributions of population summary statistics, such as TE frequencies in combination with low Tajima’s D values along flanking sequences have proven useful to reduce reference and de novo location insertions to a few positively selected candidates in *Drosophila* (Kofler, Betancourt et al. 2012). Given our data’s characteristics and our availability of two pairs of contrasting altitude

populations our choice of empirical population summary statistic was an F_{ST} outlier approach. We observed that while aiming at altitude-related candidate *reference* insertions, such set displayed specific features that indicate that they indeed may be more often involved in adaptive processes. Besides showing consistent patterns between gradients, we additionally observed that they were depauperate of nested TE insertions, and inserted more often in 5' of genes. Although this 5' enrichment observation is a coarse estimate of these TE insertion's significance, we consider it likely reflects a higher potential of generating a phenotypic effect. Studies have indeed reported more climate associated candidate TEs in gene regulatory regions with respect to 'neutral' TE insertions in *Drosophila melanogaster* (González, Karasov et al. 2010). TEs with reported functions in crop genomes are also found more often upstream (considering upstream, 5' and insertions in promoters) of genes they potentially affect (Vitte, Fustier et al. 2014). And rice *mPing* element recent burst preferentially inserted in the 5' gene flanking sequences (Naito, Zhang et al. 2009). In maize, TEs inserted upstream of up-regulated genes in response to stress conditions are themselves also expressed so perhaps acting as local enhancers of such genes expression under stress (Makarevitch, Waters et al. 2015).

Two superfamilies were enriched for *reference* candidates: RLG and DTH, while on the contrary, DTT and RLX exhibited a deficit. Such patterns were not recovered from the list of TEs with reported effects in crops (Vitte, Fustier et al. 2014) where RLC and DTA proved especially bountiful. Otherwise RLG were in fact strongly represented for TEs enriched for nearby genes that were up-regulated under abiotic stress conditions, perhaps indicating context specific action (Makarevitch, Waters et al. 2015), however so were RLX. It has also been reported from maize expression sequence tag (EST) databases, that RLG have on average the most ETS, followed by RLC (Vicient 2010). The noticeable larger proportion of *reference* candidates in RLG elements with respect to the curated set content, seems counter-intuitive to the observation that our candidates were, also closer to genes, since *gypsy* LTR-retrotransposons have been reported to show a negative correlation with gene density and recombination rates in a study of non-redundant TEs described for 81 inbred maize lines resequencing data (Lai, Schnable et al. 2017). The opposite being true for DNA transposons (Lai, Schnable et al. 2017) and in view that datasets enriched for insertions in high recombination regions, they are ideal to search for putatively adaptive insertions (Lerat, Goubert et al. 2019). This supports the adaptive potential of our enriched *reference* candidate DTH (Pif/Harbinger) insertions, a superfamily characterized in the maize B73 v.4 annotation as presenting many copies, small size and somewhat close to genes (Stitzer, Anderson et al. 2019). Furthermore, in accordance to our results, in the TE-rich and larger wheat genome, DTH

showed the highest peaks in gene vicinity, mounting asymmetrically at about 2,000 bp from genes with higher prevalence in the upstream region (Wicker, Gundlach et al. 2018). Unlike our *reference* candidate enrichment results, in wheat DTT was in fact found to form sharp mirrored peaks closer to genes (Wicker, Gundlach et al. 2018). We found particular interest in a *reference* DTH candidate that belongs to the DTH00434 family, found on chromosome 6 at position 114745805 which corresponds in the B73 maize annotation inserts directly inside the Zm00001d037170 gene, a putative bZIP transcription factor superfamily protein (Supp. Table S3). This family stands out by presenting the highest tissue-specific expression, in mature pollen (Stitzer, Anderson et al. 2019), although this gene's expression has been measured in various tissues, it was not measured for mature pollen (Walley, Sartor et al. 2016). This insertion was absent in both our *parviglumis* populations, thus we might want to ask what its frequencies resemble along the gradients.

Besides a per-element assessment, global patterns of epigenetic modifications could also come in handy to further restrict our *reference* candidate insertions list. TEs are known to affect genes nearby epigenetically (Lippman, Gendrel et al. 2004), an interesting subset of candidates could be discerned by inspecting the methylome state of candidate positions on B73 reports (Achour, Joets et al. 2019) as well as distance to maize-teosinte eQTLs described in (Wang, Chen et al. 2018) to assess their potential consequences.

III.4.3 Maize adaptive insertions do not always associate with trait variation in teosintes.

Natural and artificial selection (domestication) are expected to target distinct trait optima (Allaby 2010; Abbo, Pinhasi van-Oss et al. 2014), but see (Yan, Kenchanmane Raju et al. 2019). Allelic variants selected during domestication however, likely have measurable phenotypic effects on wild specimens (Weber, Clark et al. 2007). In order to test this hypothesis, we assessed the phenotypic impacts of *ZmCCT-ins*, *Tb1-ins* and *Vgt1-ins* in teosintes. Despite previous efforts in detecting *ZmCCT-ins* in numerous teosinte entries, authors (Yang, Li et al. 2013) did find only one case of *ZmCCT-ins* presence out of 41 *mexicana* and 38 *parviglumis* accessions, that they ascribed to gene flow from domesticated maize. The present work, that benefits from an explicit teosinte population-level assessment and two orders of magnitude more entries, further supports *ZmCCT-ins* as a post-domestication insertion, since it was absent from all four teosinte samples. Although domestication often operates on standing variation, as in maize (Weber, Clark et al. 2007), posterior breeding seemingly can operate upon new insertion events, as demonstrated for the CACTA insertion that inactivates *Bx12* gene (a benzoxainoid biosynthesis gene related to herbivore defense) (Meihls, Handrick et al. 2013) found only in temperate maize varieties where a clear selection

signature of no segregating sites along 1.3kb is found around the insertion (Wang, Chen et al. 2018). We additionally found a candidate insertion at the ZmCCT locus (Zm00001d024909) on chromosome 10, that unlike *ZmCCT-ins* does not belong to DTC (CACTA) superfamily but to DTH (Pif/Harbinger). Evaluating the effects of such insertion on teosinte flowering time under different day-length experimental conditions could further aid in our understanding of flowering time control in wild maize relatives.

The *Tb1-ins* has notorious impacts on maize plant architecture through downstream regulation of the *tb1* gene. The effect of *Tb1-ins* in teosinte has only recently been analyzed, for instance on one *parviglumis* population grown in greenhouse conditions at high plant density, where authors did not recover an effect of the insertion on tillering index (Vann, Kono et al. 2015). The absence of significant association of this insertion in our teosinte association panel, with traits such as number of lateral branches and number of tillers, falls in accordance to results by Vann *et al*., (2015). The effects of wild and domestication alleles could differ depending on the genetic background, through epistatic effects (Doust, Lukens et al. 2014). Introgression of the teosinte *tb1* allele as homozygous into isogenic maize lines showed that plants were more phenotypically plastic than when homozygous for maize alleles (Lukens and Doebley 1999). Indeed, plants showed more tillering but this effect was by far reduced when plants were grown at high densities, as part of a genome by environment interaction where shade avoidance is accomplished through taller and less bushy plant architecture (Lukens and Doebley 1999). In order to obtain a 'teosinte-like' phenotype by introgressing teosinte alleles into maize background, *tb1* is not enough and another wild locus is needed, QTL-1L from chromosome 3 (Lukens and Doebley 1999). There is evidence that the interaction between teosinte *tb1* alleles introgressed in maize with QTL-1L on chromosome 3 determines lateral flower gender (Lukens and Doebley 1999). While in the maize background both wild loci interact to form the wild phenotype, the domesticated allele at one locus can be sufficient to produce the domesticated phenotype.

A recent study on rice domestication by Wang, *et al*. (2017) found that present day wild rice (which seems to form a hybrid swarm with local domesticated varieties) and varietal accessions harbor the same sequence for the two most important domestication genes, *sh4* (for non shattering) and *PROG1* (for erect growth). However wild rice continues to show a shattering phenotype regardless of the presence of the domesticated *sh4* allele. Since the domesticated allele has been thought to be shared via gene flow, the authors propose that compensatory mechanisms have evolved in wild rice populations. In our case this might not be the only possible explanation, and rather a study by Swanson *et al*. (2016) on maize transcriptome rewiring by domestication could

have part of the answer. When comparing the topologies of co-expression networks and the correlation between edges in maize and teosinte networks, these authors found that the correlation between edges of the networks was lower than expected by chance. Also, there were fewer conserved gene pairs and lower degree of similarity between neighbors surrounding a candidate gene in the maize network with respect to the teosinte network (Swanson-Wagner, Briskine et al. 2012). Query genes highly connected in teosinte loose connections following domestication. Unfortunately the authors could not assess the possible rewiring of *tb1* because in order to have comparable developmental stages they only worked on 8 day seedlings and *tb1* is expressed later in development. Recent maize gene regulatory networks evidences from the maize expression atlas show that, among four explored tissues for transcriptome data, the *tb1* transcription factor is only expressed in shoot apical meristem tissue, perhaps due to different heterochromatin formation and gene accessibility (Huang, Zheng et al. 2018).

Interestingly, we did find *Tb1-ins* associated to traits other than branching. As previously outlined, this could possibly be due to strong pleiotropy at this locus in wild backgrounds. The genetic background influence hypothesis is further supported by the fact that we indeed observed a considerable TE genotype by population interaction effect for the concerned trait in either of our structure-correcting models (five groups or 11 populations). Indicating that the same insertion doesn't show the same effect in all populations, an observation reported in domestication genetics when different populations have different genetic makeup (Stitzer and Ross-Ibarra 2018).

Vgt1-ins geographic distribution patterns among our association mapping panel followed closely neutral structure in five groups reported in (Fustier, Martínez-Ainsworth et al. 2019) perhaps explaining the lack of association with flowering time when correcting for *K=5* structure, whereas when running the mixed model for 11 population neutral structure correction, we effectively found the insertion presence to associate to flowering time. In maize *Vgt1-ins* is involved in adaptation to long-day conditions (Ducrocq, Madur et al. 2008), (Castelletti, Tuberosa et al. 2014). A survey of *Vgt1-in*s presence in 256 maize populations concluded that farmers management has positively selected *Vgt1-ins* along maize's northern American migration to cold temperate environments (Tenaillon and Charcosset 2011). The results also suggested that *Vgt1-ins* may have been involved in the differentiation of maize varieties according to elevation in tropical Central America (Ducrocq, Madur et al. 2008). Here, despite the very small difference in day-length (Figure 5-D), this insertion effectively affects teosinte flowering time as a form of standing variation. However, the insertion frequency's altitudinal pattern was less clear (Figure 5-C).

Note that these *Tb1-ins* and *Vgt1-ins* were not recovered from our set of candidate insertions because they presented altitudinal profiles that did not match our criteria (Figure 5-A,C; Supp. Figure S9), and neither were correlated to environmental PC1 (strongly co-varying with altitude). We did however find precipitation seasonality, the mean diurnal range and the minimum temperature of the coldest month to covary with *Vgt1-ins.* This corroborates the importance of contrasting each environmental variable on its own as posited by Lotterhos *et al.* (2018) (Lotterhos, Yeaman et al. 2018) who comment on the risks of using PCs to account for biologically meaningful variation, which in certain scenarios is best captured by testing environmental variables independently. It is not entirely clear whether and how these variables could act as selective agents on teosinte *Vgt1-ins*. Indeed, although precipitation seasonality has been reported to be more variable and contribute more strongly to *parviglumis* ecological niche model than to that of *mexicana* (Hufford, Martínez-Meyer et al. 2012) and populations with the highest *Vgt1-ins* frequency were indeed *parviglumis* the insertion was also found at non-negligible frequencies in *mexicana* populations.

ACKNOWLEDGEMENTS

We are especially thankful to Margaux-Alison Fustier for helping to coordinate field experiments with Salvador Montes Hernández and María Guadalupe Camarena in the INIFAP facilities in México. Masters student Juliette Aubert contributed to PCR genotyping of *Tb1-ins* and *Vgt1-ins* under the guidance of Hélène Corti. Jonás Aguirre kindly helped with providing SNPchip data and running the TreeMix analysis. Anthony Venon, Mathieu Falque and Fabrice Dumas contributed SSR genotyping data. Agnès Rousselet helped in DNA extractions. Nicolas Bierne shared constructive suggestions on TE screening and analyzes.

FUNDING

This work was supported by two grants overseen by the French National Research Agency (ANR) (Project ANR 12-ADAP-0002-01) to MIT and YV. NEM-A was funded by a Mexican government CONACYT PhD fellowship grant no. 579966/410748. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

III.5 REFERENCES

Abbo, S., R. Pinhasi van-Oss, et al. (2014). "Plant domestication versus crop evolution: A conceptual framework for cereals and grain legumes." Trends in Plant Science **19**: 351-360.

Achour, Z., J. Joets, et al. (2019). "Low temperature triggers genome-wide hypermethylation of transposable elements and centromeres in maize." bioRxiv.

Aguirre-Liguori, J. A., M. I. Tenaillon, et al. (2017). "Connecting genomic patterns of local adaptation and niche suitability in teosintes." Molecular Ecology **26**: 4226-4240.

Allaby, R. (2010). "Integrating the processes in the evolutionary system of domestication." Journal of Experimental Botany **61**: 935-944.

Anderson, S., M. C. Stitzer, et al. (2019). "Transposable elements contribute to dynamic genome content in maize." bioRxiv.

Anderson, S. N., M. C. Stitzer, et al. (2019). "Dynamic patterns of transcript abundance of transposable element families in maize." bioRxiv.

Baucom, R. S., J. C. Estill, et al. (2009). "Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome." PLoS Genetics **5**.

Bennetzen, J. L. (2007). "Patterns in grass genome evolution." Current Opinion in Plant Biology **10**: 176-181.

Bilinski, P., P. S. Albert, et al. (2018). "Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays." PLoS Genetics **14**.

Blanc, G. (2004). "Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes." the Plant Cell Online **16**: 1667-1678.

Butler, D., B. R. Cullis, et al. (2007). "ASReml-R reference manual." Technical Report.

Castelletti, S., R. Tuberosa, et al. (2014). "A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL Vgt1." Genes|Genomes|Genetics **4**: 805-812.

Chia, J. M., C. Song, et al. (2012). "Maize HapMap2 identifies extant variation from a genome in flux." Nature Genetics **44**: 803-807.

Coop, G., D. Witonsky, et al. (2010). "Using environmental correlations to identify loci underlying local adaptation." Genetics **185**: 1411-1423.

Cuervo-Robayo, A. P., O. Téllez-Valdés, et al. (2014). "An update of high-resolution monthly climate surfaces for Mexico." International Journal of Climatology **34**: 2427-2437.

Diez, C. M., B. S. Gaut, et al. (2013). "Genome size variation in wild and cultivated maize along altitudinal gradients." New Phytologist **199**: 264-276.

Diez, C. M., E. Meca, et al. (2014). "Three Groups of Transposable Elements with Contrasting Copy Number Dynamics and Host Responses in the Maize (Zea mays ssp. mays) Genome." PLoS Genetics **10**.

Doust, A. N., L. Lukens, et al. (2014). "Beyond the single gene: How epistasis and gene-by- environment effects in fl uence crop domestication." **111**.

Ducrocq, S., D. Madur, et al. (2008). "Key impact of Vgt1 on flowering time adaptation in maize: Evidence from association mapping and ecogeographical information." Genetics **178**: 2433-2437.

Feschotte, C., N. Jiang, et al. (2002). "Plant Transposable Elements: Where Genetics Meets Genomics." Nature Reviews Genetics **3**: 329-341.

Fustier, M.-A., N. E. Martínez-Ainsworth, et al. (2019). "Common gardens in teosintes reveal the establishment of a syndrome of adaptation to altitude." bioRxiv: 563585.

Fustier, M. A., J. T. Brandenburg, et al. (2017). "Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples." Molecular Ecology **26**: 2738-2756.

Goerner-Potvin, P. and G. Bourque (2018). "Computational tools to unmask transposable elements." Nature Reviews Genetics **19**: 688-704.

González, J., T. L. Karasov, et al. (2010). "Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila." PLoS Genetics **6**: 33-35.

González, J., K. Lenkov, et al. (2008). "High rate of recent transposable element-induced adaptation in Drosophila melanogaster." PLoS Biology **6**: 2109-2129.

Goubert, C., H. Henri, et al. (2017). "High-Throughput Sequencing of Transposable Element Insertions Suggests Adaptive Evolution of the Invasive Asian Tiger Mosquito Towards Temperate Environments." Molecular Ecology: 1-14.

Hawkins, J. S., S. R. Proulx, et al. (2009). "Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants." Proceedings of the National Academy of Sciences **106**: 17811-17816.

Hazzouri, K. M., A. Mohajer, et al. (2008). "Contrasting patterns of transposable-element insertion polymorphism and nucleotide diversity in autotetraploid and allotetraploid Arabidopsis species." Genetics **179**: 581-592.

Hollister, J. D. and B. S. Gaut (2009). "Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression." Genome Research **19**: 1419-1428.

Horvath, R. and T. Slotte (2017). "The role of small RNA-based epigenetic silencing for purifying selection on transposable elements in capsella grandiflora." Genome Biology and Evolution **9**: 2911-2920.

Huang, J., J. Zheng, et al. (2018). "Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize." BMC Plant Biology **18**: 1-14.

Hufford, M. B., E. Martínez-Meyer, et al. (2012). "Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight." PLoS ONE **7**.

Ibarra-Laclette, E., E. Lyons, et al. (2013). "Architecture and evolution of a minute plant genome." Nature **498**: 94-98.

Jiao, Y., P. Peluso, et al. (2017). "Improved maize reference genome with." Nature.

Kelley, J. L., J. T. Peyton, et al. (2014). "Compact genome of the Antarctic midge is likely an adaptation to an extreme environment." Nature Communications **5**: 1-8.

Kobayashi, S., N. Goto-Yamamoto, et al. (2004). "Retrotransposon-Induced Mutations in Grape Skin Color." Science **304**: 982.

Kofler, R., A. J. Betancourt, et al. (2012). "Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster." PLoS Genetics **8**.

Kofler, R., D. Gómez-Sánchez, et al. (2016). "PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq." Molecular Biology and Evolution **33**: 2759-2764.

Lai, J., Y. Li, et al. (2005). "Gene movement by Helitron transposons contributes to the haplotype variability of maize." Proceedings of the National Academy of Sciences **102**(25): 9068-9073.

Lai, X., J. C. Schnable, et al. (2017). "Genome-wide characterization of non-reference transposable element insertion polymorphisms reveals genetic diversity in tropical and temperate maize." BMC Genomics **18**: 1-13.

Lerat, E., C. Goubert, et al. (2019). Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. Molecular Ecology, Blackwell Publishing Ltd.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores Mapping short DNA sequencing reads and calling variants using mapping quality scores." 1851-1858.

Lippman, Z., A. V. Gendrel, et al. (2004). "Role of transposable elements in heterochromatin and epigenetic control." Nature **430**: 471-476.

Lisch, D. and R. K. Slotkin (2011). Strategies for Silencing and Escape. The Ancient Struggle Between Transposable Elements and Their Hosts. International Review of Cell and Molecular Biology, Elsevier Inc. **292:** 119-152.

Lockton, S. and B. S. Gaut (2010). "The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata." BMC evolutionary biology **10**: 10.

Lotterhos, K. E., S. Yeaman, et al. (2018). "Modularity of genes involved in local adaptation to climate despite physical linkage 06 Biological Sciences 0604 Genetics." Genome Biology **19**: 1-24.

Lukens, L. N. and J. Doebley (1999). "Epistatic and environmental interactions for quantitative trait loci involved in maize evolution." Genetical Research **74**: 291-302.

Makarevitch, I., A. J. Waters, et al. (2015). "Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress." PLoS Genetics **11**: 1-15.

Mascagni, F., E. Barghini, et al. (2015). "Repetitive DNA and plant domestication: variation in copy number and proximity to genes of LTR-retrotransposons among wild and cultivated sunflower (Helianthus annuus) genotypes." Genome biology and evolution **7**: 3368-3382.

Meihls, L. N., V. Handrick, et al. (2013). "Natural Variation in Maize Aphid Resistance Is Associated ½AQ2 Glucoside Methyltransferase Activity." The Plant cell **25**: 1-16.

Messing, J. and H. K. Dooner (2006). "Organization and variability of the maize genome." Current Opinion in Plant Biology **9**: 157-163.

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

Naito, K., F. Zhang, et al. (2009). "Unexpected consequences of a sudden and massive transposon amplification on rice gene expression." Nature **461**: 1130-1134.

Quadrana, L., A. B. Silveira, et al. (2016). "The Arabidopsis thaliana mobilome and its impact at the species level." eLife **5**: 1-25.

Rech, G. E., M. Bogaerts-Márquez, et al. (2019). "Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila." PLoS genetics **15**: e1007900.

Ross-Ibarra, J., M. Tenaillon, et al. (2009). "Historical divergence and gene flow in the genus Zea." Genetics **181**: 1399-1413.

Salvi, S., S. Corneti, et al. (2011). "Genetic dissection of maize phenology using an intraspecific introgression library." BMC Plant Biology **11**.

Salvi, S., G. Sponza, et al. (2007). "Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize." Proceedings of the National Academy of Sciences of the United States of America **104**: 11376-11381.

Sanmiguel, P., B. S. Gaut, et al. (1998). The paleontology of intergene retrotransposons of maize.

Schnable, P. S., D. Ware, et al. (2009). "The B73 Maize Genome: Complexity, Diversity, and Dynamics."

Springer, N. M., S. N. Anderson, et al. (2018). "The maize W22 genome provides a foundation for functional genomics and transposon biology." Nature Genetics **50**.

Stitzer, M. C., S. N. Anderson, et al. (2019). "The Genomic Ecosystem of Transposable Elements in Maize." bioRxiv: 559922.

Stitzer, M. C. and J. Ross-Ibarra (2018). "Maize domestication and gene interaction."

Stuart, T., S. R. Eichten, et al. (2016). "Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation." eLife **5**.

Studer, A., Q. Zhao, et al. (2011). "Identification of a functional transposon insertion in the maize domestication gene tb1." Nature Genetics **43**: 1160-1163.

Swanson-Wagner, R., R. Briskine, et al. (2012). "Reshaping of the maize transcriptome by domestication." Pnas **109**: 11878-11883.

Swigoňová, Z., J. Lai, et al. (2004). "Close split of sorghum and maize genome progenitors." Genome Research **14**: 1916-1923.

Tenaillon, M. I. and A. Charcosset (2011). "A European perspective on maize history." Comptes Rendus - Biologies **334**: 221-228.

Tenaillon, M. I., J. D. Hollister, et al. (2010). "A triptych of the evolution of plant transposable elements." Trends in Plant Science **15**: 471-478.

Vann, L., T. Kono, et al. (2015). "Natural variation in teosinte at the domestication locus teosinte branched1 (tb1)." PeerJ **3**: e900.

Vicient, C. M. (2010). "Transcriptional activity of transposable elements in maize." BMC genomics **11**: 601.

Vitte, C., M. A. Fustier, et al. (2014). "The bright side of transposons in crop evolution." Briefings in Functional Genomics and Proteomics **13**: 276-295.

Vitte, C. and O. Panaud (2005). LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. Cytogenetic and Genome Research, S. Karger AG. **110:** 91-107.

Walley, J. W., R. C. Sartor, et al. (2016). "Integration of omic networks in a developmental atlas of maize." Science **353**: 814-818.

Wang, Q. and H. K. Dooner (2006). "Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus." Proceedings of the National Academy of Sciences **103**: 17644-17649.

Wang, X., Q. Chen, et al. (2018). "Genome-wide Analysis of Transcriptional Variability in a Large Maize-Teosinte Population." Molecular Plant **11**: 443-459.

Weber, A., R. M. Clark, et al. (2007). "Major regulatory genes in maize contribute to standing variation in teosinte (Zea mays ssp. parviglumis)." Genetics **177**: 2349-2359.

Wicker, T., H. Gundlach, et al. (2018). "Impact of transposable elements on genome structure and evolution in bread wheat." Genome Biology **19**: 1-18.

Wright, S. I., Q. H. Le, et al. (2001). "Population Dynamics of an Ac-like Transposable Element in Self- and Cross-Pollinating Arabidopsis." Genetics **158**(3): 1279-1288.

Yan, L., S. K. Kenchanmane Raju, et al. (2019). "Parallels between natural selection in the cold-adapted crop-wild relative Tripsacum dactyloides and artificial selection in temperate adapted maize." Plant Journal: 965-977.

Yang, Q., Z. Li, et al. (2013). "CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize." Proceedings of the National Academy of Sciences **110**: 16969-16974.

Yao, J.-L., Y.-H. Dong, et al. (2012). "Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor." Proceedings of the National Academy of Sciences **98**: 1306-1311.

Zerjal, T., J. Joets, et al. (2009). "Contrasting evolutionary patterns and target specificities among three tourist-like MITE families in the maize genome." Plant Molecular Biology **71**: 99-114.

III.6 Supplementary Figures

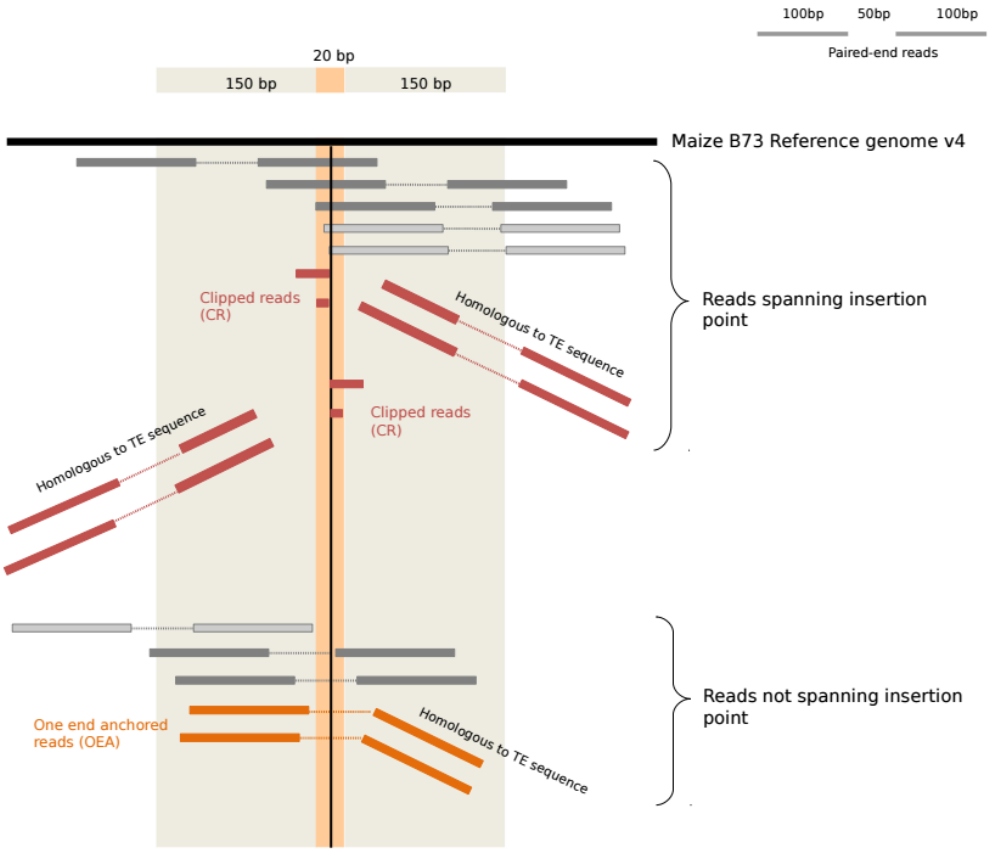


Fig. S1 Detection of *de novo* insertions and population frequency estimates. Tlex *de novo* employs population pooled paired-end reads to call TE *de novo* insertion points. We recovered *de novo* insertions from both One End anchored (OEA) and Clipped reads (CR). The latter defines an insertion point (black vertical line) and zone (20 bp, in yellow). We estimated *de novo* frequency within population from CR reads as the ratio of the number of clipped reads in the insertion zone with at least five bp mapped to the reference genome (here two, in red) over the number of these clipped reads plus the traversing reads spanning the insertion zone with at least 10 bp on each side of the insertion point (here three, in dark gray from the group of reads "spanning the insertion point"). Local depth within the 300-bp window in this example was 7.

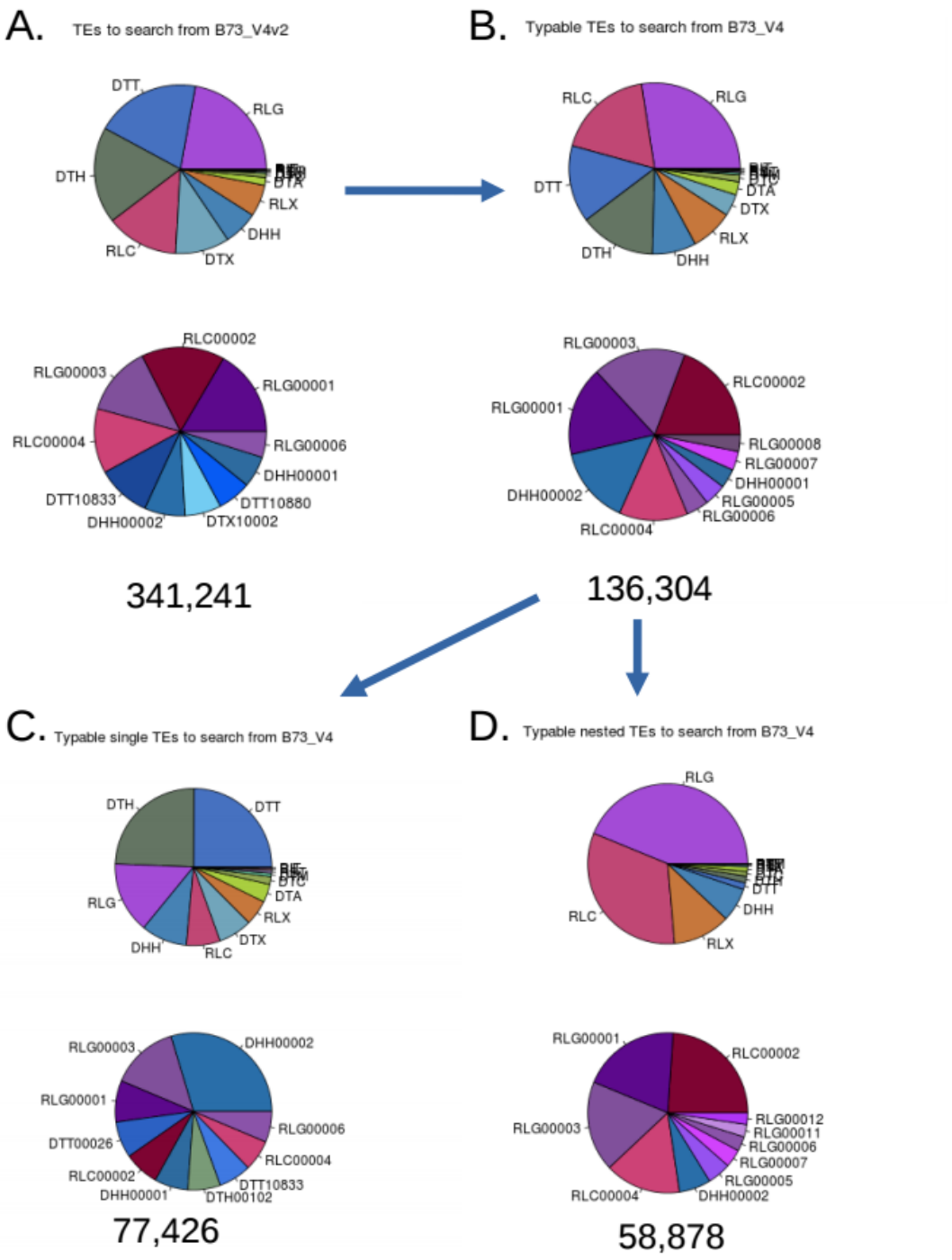


Fig. S2: Superfamily (upper panels) and ten most abundant families (lower panels) pie charts contained in the TE data-base of the reference genome (A), non-nested TEs therein included (B) and their further classification into those that are single (C) and those which themselves include nested TE fragments within them (D). The number of elements are reported below the pie charts.

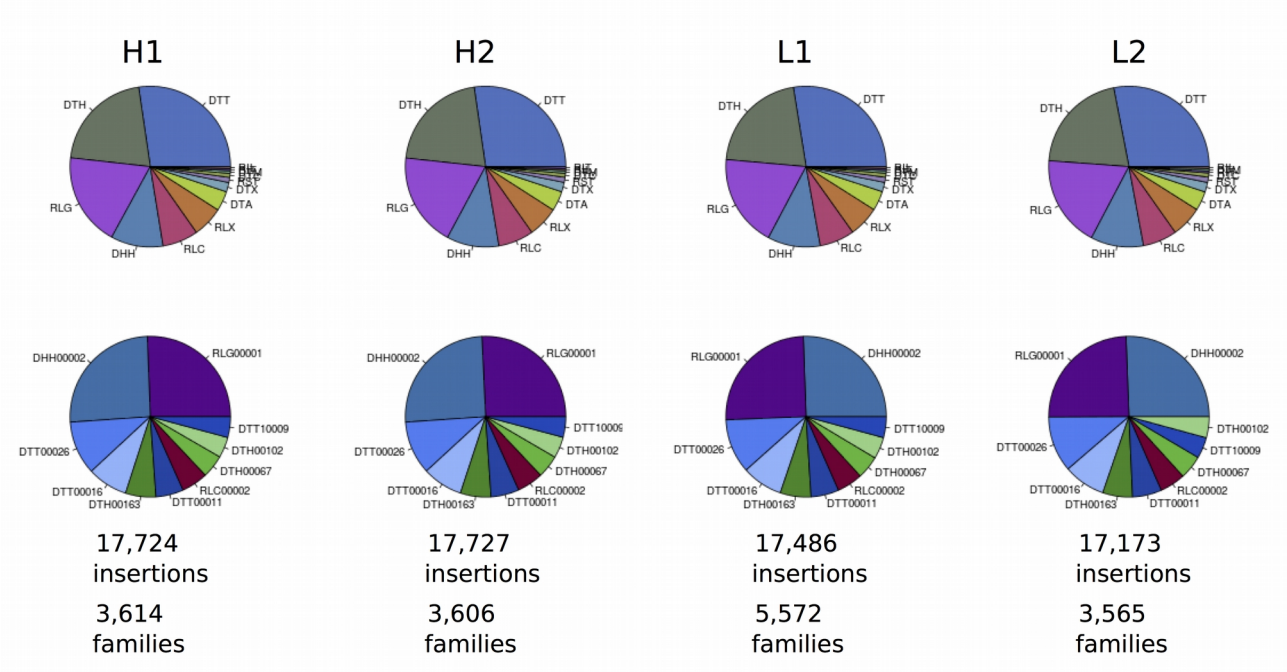


Fig. S3: Superfamily and ten most abundant family pie charts for *reference* TE insertions in four HTS populations. Total number of elements and families found per population are indicated below pie charts. *Reference* TE insertions were searched among the non-nested TEs of the reference genome (Fig. S2 B).

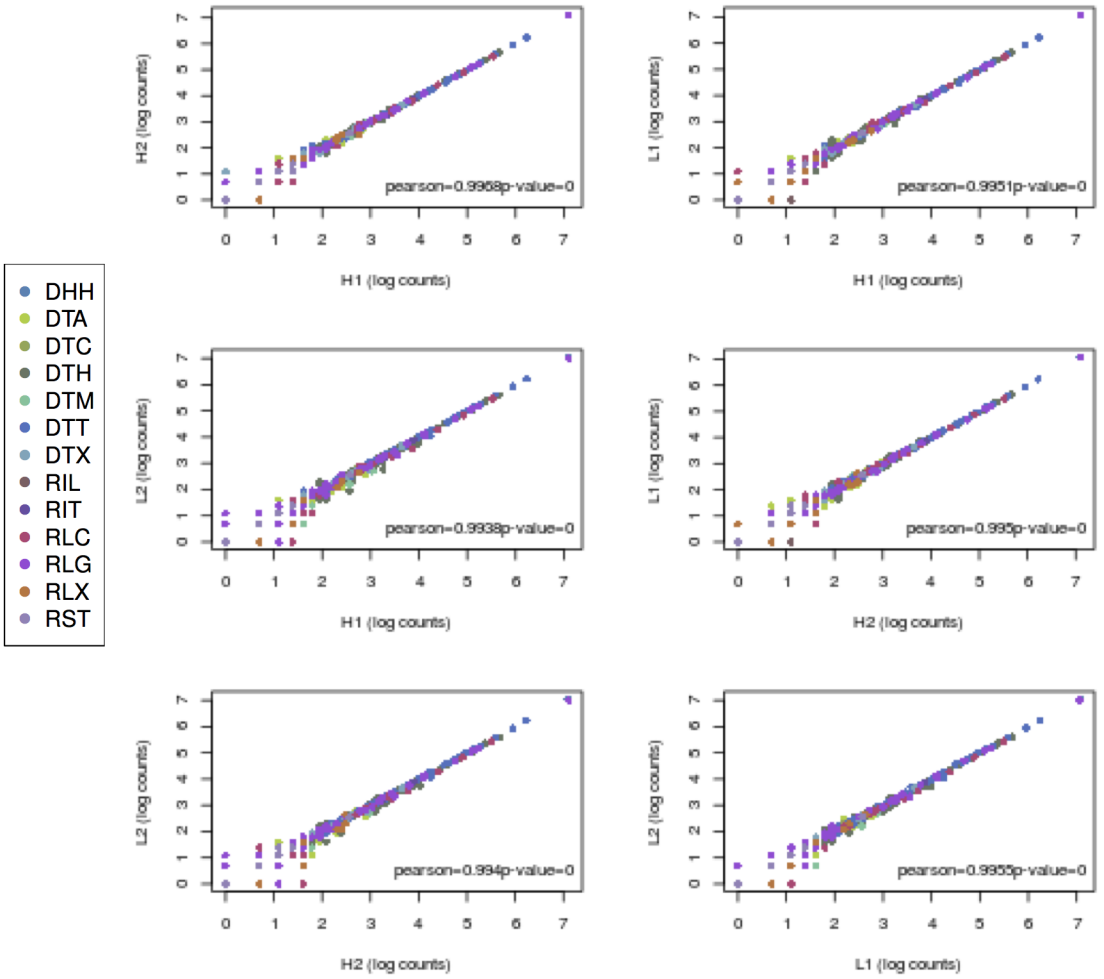


Fig. S4: *Reference* insertions correlation for all TE families obtained for each of the six pairwise comparisons. Log values of counts are plotted. Each pair comprised a different amount of families in common as follows: H1-H2 (3,574), H1-L1 (3,529), H1-L2 (3,521), H2-L1 (3,530), H2-L2 (3,520), L1-L2 (3,507).

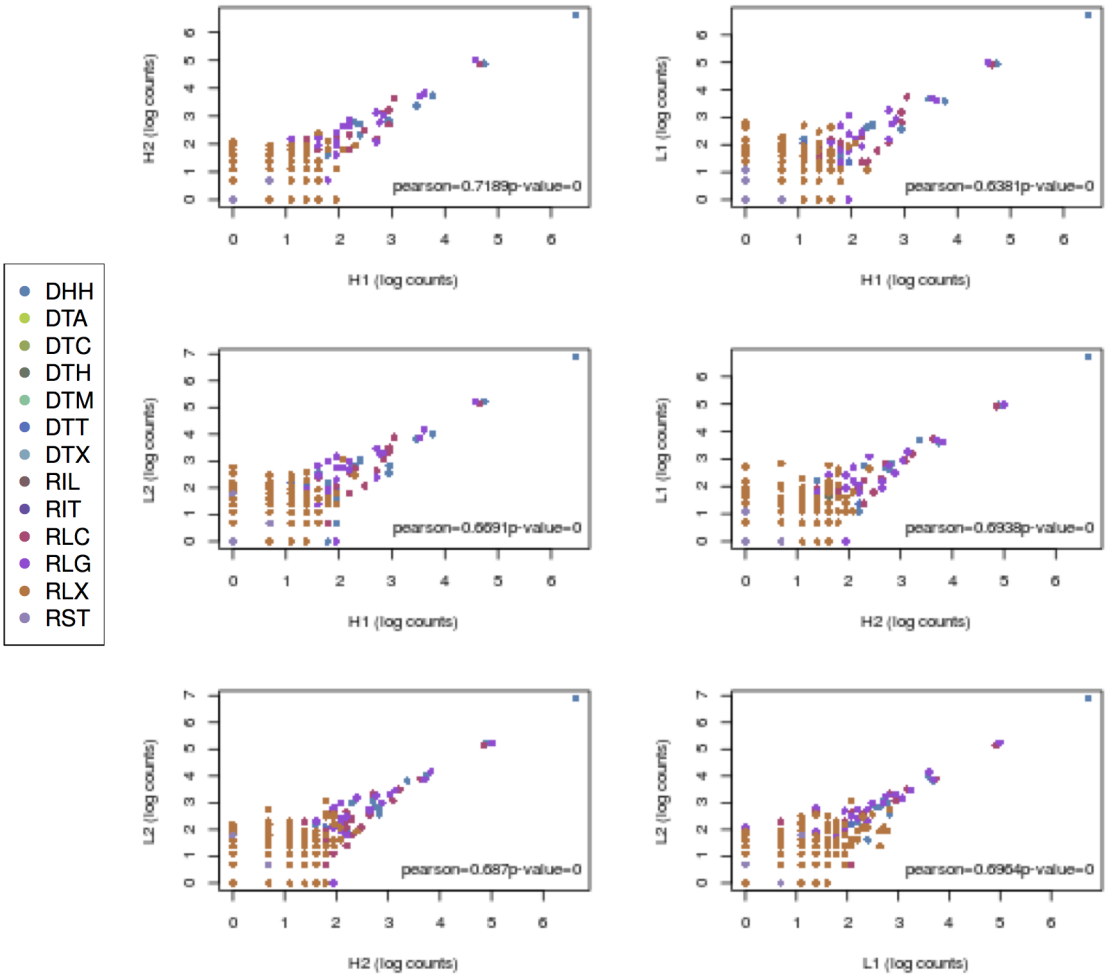


Fig. S5: *de novo* insertions correlation for all TE families obtained for each of the six pairwise comparisons. Log values of counts are plotted.

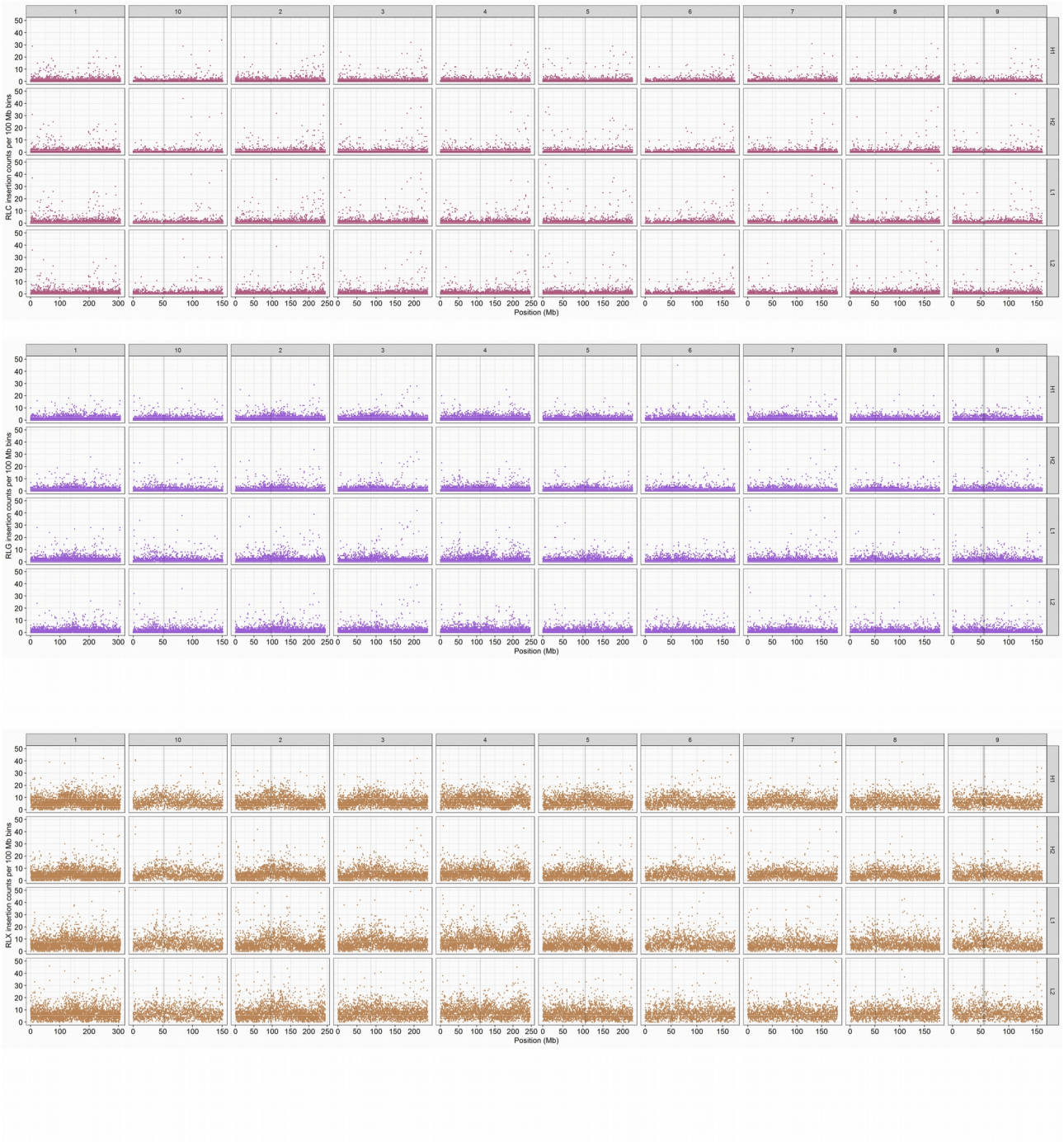


Fig S6: Genomic landscape of *de novo* insertions of all superfamilies along chromosomes 1 to 10. Each point represents the amount of TE insertions present in 100kb bins along each chromosome.

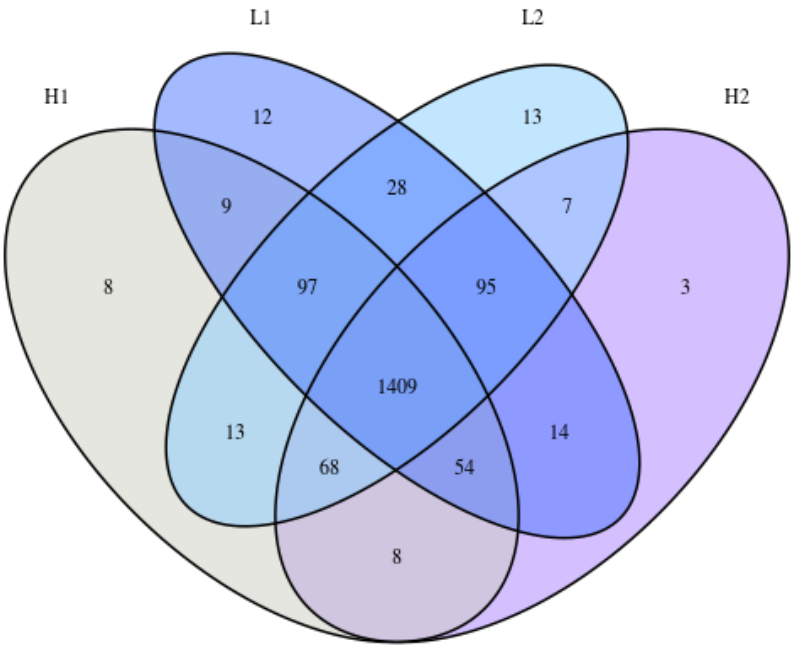


Fig S7: Number of *de novo* insertions unique or shared among HTS teosinte populations. Only insertions with evidence in all four populations were taken (1,838). Since frequencies were calculated after Tlex-de-novo calling, sometimes due to our stringent criteria TEs that had been called present are reported with zero frequency, hence the TEs in the Venn diagram areas other than the four population's intersection.

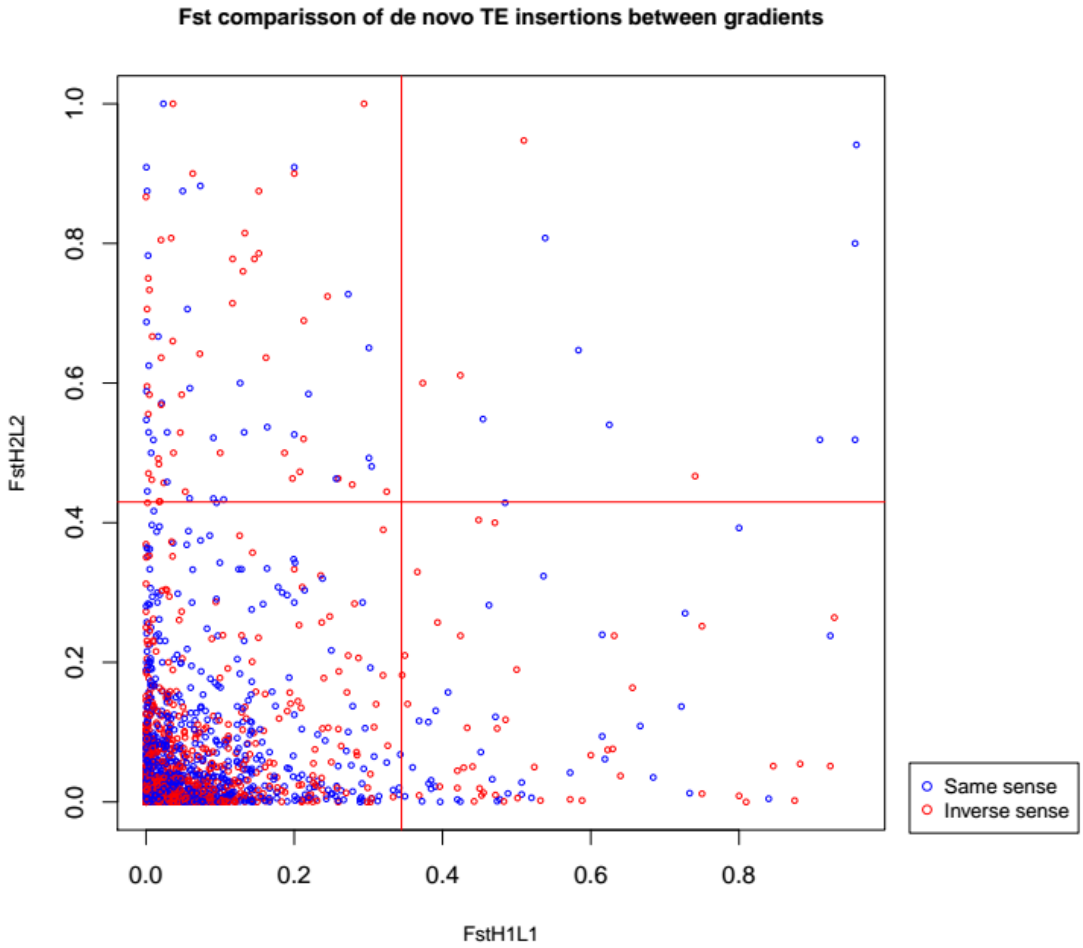


Fig. S8: *de novo* insertions scatterplot on paired F_{ST} values per gradient (H1-L1/H2/L2). The 5% outlier threshold for each gradient is marked and colors indicate parallel or contrasted frequency tendency along highlands and lowlands comparisons.

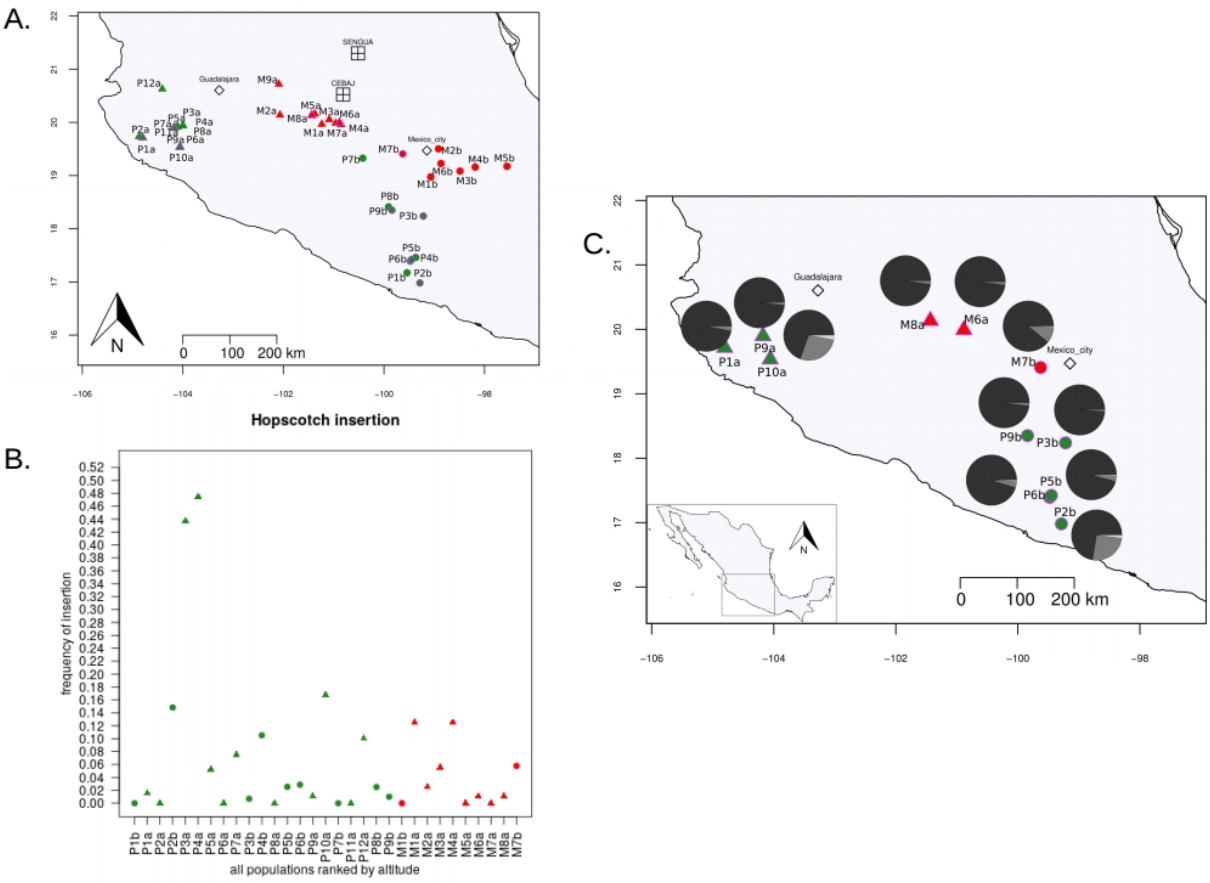


Fig S9: Geographic localization of the entire set of 11 teosinte populations (A), frequencies of *Tb1-ins* for all 31 populations ranked by altitude (B) and geographical distribution of *Tb1-ins* frequencies over the association panel (C). Colors in A indicate subspecies (*parviglumis*=green, *mexicana*=red) and shapes relate to gradients (gradient 1 = circles, gradient 2 = triangles). Pie plots in C indicate the proportion of individuals homozygotes for presence (light gray), absence (black) and heterozygotes (mild gray).

III.7 SUPPLEMENTARY TABLES

S1 Table. Description of 37 teosinte populations, marking sets of populations used for TE identification and association mapping analyses, as well as PCR genotyped insertion frequencies per population.

Population code (parenthesis indicating nomenclature in Fustier et al., 2017), accession names and corresponding references (indicated in the main text) where they are described are indicated along with the subspecies names, geographical coordinates, locality, state and year of collection. colors indicate populations for which different types of data were available, with TE insertion genotyping produced in this study

Population code	Accession names	References	Subspecies	Latitude	Longitude	Altitude	Locality	State	Year of collection	HTS	SSR genotyping	Phenotyping	*Tb1-ins* frequency	number of individuals screened for *Tb1-ins*	*Vgt1-ins* frequency	number of individuals screened for *Vgt1-ins*
Gradient *a*																
P1a (L2)	CM13	[2]	*parviglumis*	19.7154	-104.8035	504	Telpitita	Jalisco	2010				0.030	266	0.015	264
P2a	SMH 577	[2]	*parviglumis*	19.7325	-104.8719	572	El_Llanito	Jalisco	2009				0.000	40	0.000	14
P3a	SMH 570	[1]	*parviglumis*	19.9339	-104.0077	944	La_Labor	Jalisco	2009				0.526	38	0.000	34
P4a	MIT14	[2]	*parviglumis*	19.9458	-103.9958	958	San_Lorenzo	Jalisco	2010				0.604	53	0.061	33
P5a	SMH 569	[2]	*parviglumis*	19.9218	-104.0977	963	Los_Naranjos	Jalisco	2009				0.100	40	0.326	43
P6a	SMH 567	[2]	*parviglumis*	19.9130	-104.1169	976	El_Estanco	Jalisco	2009				0.000	32	0.270	37
P7a	MIT 15	[2]	*parviglumis*	19.9129	-104.1123	976	Ejutla	Jalisco	2010				0.140	43	0.182	22
P8a	SMH 566	[2]	*parviglumis*	19.9027	-104.1572	1140	El_Estanco	Jalisco	2009				0.000	40	0.118	17
P9a	SMH 565	[1]	*parviglumis*	19.8961	-104.1768	1317	Ejutla	Jalisco	2009				0.014	277	0.501	339
P10a	SMH578	[1]	*parviglumis*	19.5354	-104.0583	1369	El Rodeo	Jalisco	2009				0.268	298	0.047	296
P11a	SMH 564	[1]	*parviglumis*	19.9100	-104.1726	1407	Ejutla	Jalisco	2009				0.000	40	0.146	41
P12a	CM12	[1]	*parviglumis*	20.6273	-104.4079	1426	Guachinango	Jalisco	2010				0.182	44	0.148	27
M1a	SMH 580	[2]	*mexicana*	19.9646	-101.2354	1844	Capacho	Michoacán	2010				0.222	45	0.000	22
M2a	CM11	[2]	*mexicana*	20.1394	-102.0684	1846	Churintzio	Michoacán	2010				0.049	41	0.333	30
M3a	SMH575	[1]	*mexicana*	20.0532	-101.0881	1849	San_Rafael	Michoacán	2009				0.105	38	0.293	41
M4a	SMH 572	[2]	*mexicana*	19.9582	-100.8497	1854	Andocutin	Guanajuato	2009				0.222	45	0.500	28
M5a	SMH576	[1]	*mexicana*	20.1607	-101.3731	1856	Tejocote_de_Clalera	Guanajuato	2009				0.000	40	0.000	2
M6a (H2)	SMH571	[1]	*mexicana*	19.9917	-100.8849	1861	San_Jose_de_las_Pilas	Guanajuato	2009				0.021	283	0.288	319
M7a	SMH 573	[2]	*mexicana*	19.9828	-100.9599	1878	Puerto_de_Cabras	Guanajuato	2009				0.000	38	0.452	31
M8a	SMH579	[2]	*mexicana*	20.1338	-101.4342	2002	Armadillo	Michoacán	2010				0.022	279	0.140	285
M9a	SMH-MGCH 582	[3]	*mexicana*	20.7195	-102.0898	2176	Jesús_María	Jalisco	2011				NA	0	NA	0
Gradient *b*																
P1b	CM03	[1]	*parviglumis*	17.1719	-99.5415	343	Tierra Colorada	Guerrero	2010				0.000	40	0.000	22
P2b (L1)	CM04	[1]	*parviglumis*	16.9811	-99.2855	581	Tecoanapa	Guerrero	2010				0.245	327	0.014	293
P3b	CM08	[1]	*parviglumis*	18.2377	-99.2180	1101	Paso_Morelos	Guerrero	2010				0.007	288	0.631	336
P4b	CM07	[1]	*parviglumis*	17.4596	-99.3683	1107	Mochitlán	Guerrero	2010				0.190	42	0.188	32
P5b	CM05	[1]	*parviglumis*	17.3918	-99.4776	1201	Chilpancingo	Guerrero	2010				0.043	281	0.079	277
P6b	CM06	[1]	*parviglumis*	17.4219	-99.4507	1251	Mazatlán	Guerrero	2010				0.048	249	0.137	248
P7b	SMH 581	[2]	*parviglumis*	19.3274	-100.4214	1383	Enandio	Michoacán	2011				0.000	38	0.267	30

P8b (IP1)	CM02	[1]	*parviglumis*	18.4110	-99.9084	1439	Alcholoa	Guerrero	2010		0.049	41	0.560	25
P9b	CM01	[1]	*parviglumis*	18.3498	-99.8411	1649	Teloloapan	Guerrero	2010		0.020	301	0.673	318
M1b (IM1)	CM09	[1]	*mexicana*	18.9741	-99.0703	1669	Huilotepec	Morelos	2010		0.000	38	0.286	7
M2b	Texcoco_Texcoco	[3]	*mexicana*	19.5025	-98.9146	2234	Texcoco	México	2011		NA	0	NA	0
M3b	Cocotitlan_Cocotitlan	[3]	*mexicana*	19.2262	-98.8661	2252	Cocotitlán	México	2011		NA	0	NA	0
M4b	Tenancingo_Tenancingo	[3]	*mexicana*	19.1595	-98.1854	2306	Tenancingo	Tlaxcala	2011		NA	0	NA	0
M5b	SanNicolas_SnNicolasBuenosAires	[3]	*mexicana*	19.1742	-97.5528	2375	San_Nicolas	Puebla	2011		NA	0	NA	0
M6b	Calpan_Calpan	[3]	*mexicana*	19.0839	-98.4874	2447	Calpan	Puebla	2011		NA	0	NA	0
M7b (H1)	CM10	[1]	*mexicana*	19.4075	-99.6271	2581	VillaSeca	Mexico	2010		0.104	309	0.426	343
Experimental fields														
CEBAJ	-	-	-	20.5222	-100.8122	1750	CEBAJ	Guanajuato	-	- -				
SENGUA	-	-	-	21.2986	-100.5164	2017	SENGUA	Guanajuato	-	- -				

[1] : C. Muñoz-Díez, et al., New Phytol. 199 (2013) 264-276, [2] : M.A. Fustier, et al., Mol Ecol. 26 (2017) 2738–56, [3] J.A. Aguirre-Liguori, et al., Mol Ecol. 26 (2017) 1-15

S2 Table. List of the 18 phenotypic traits measured in Fustier *et al*. (2019)

Trait	Description
Plant architecture	
PL Plant Height	Length of the main tiller from the stem base to the tip of the primary tassel (m)
HLE Height of the Lowest Ear	Length of the primary tiller from the stem base to the lowest ear insertion (m)
HHE Height of the Highest Ear	Length of the primary tiller from the stem base to the highest ear insertion (m)
Til number of Tillers	Number of tillers
number of Lateral Branches	Number of lateral branches on the primary tiller
NoE number of Nodes with Ears	Number of nodes with ears on the primary tiller
Leaf features	
LeL Leaf Length	Length of one intermediate leaf on the primary tiller (cm)
LeW Leaf Width	Width of one intermediate leaf on the primary tiller (cm)
LeC Leaf Color	color of leaves on the whole plant on a qualitative scale (1-4)
Reproduction	
FFT Female Flowering Time	Number of days from field planting to first visible silks
MFT Male Flowering Time	Number of days from field planting to anther dehiscence
TBr Tassel Branching	Number of branches in the tassel of the primary tiller
Grain features	
Gr number of Grains per ear	Number of grains per ear based on 5 ears for teosintes
GrL Grain Length	Average length of grains as measured on 10 mature grains (mm)
GrWi Grain Width	Average width of grains based as measured on 10 mature grains (mm)
GrWe Grain Weight	Average grain weight based on 50 mature grains (g)
GrC Grain Color	Average color intensity of mature grains on a qualitative scale (1-6)
Stomata features	
StD Stomata Density	Density of stomata based on 9 image measurements on a single leaf (mm^2)

S3 Table. List of reference and de novo candidate TE insertions.

Chromosome, Position and TE family and superfamily names are shown. For de novo insertions, we used the family and superfamily name of the individual copy recovered as the most similar sequence to the TE called. Frequencies in all four populations are provided as well as Fst values calculated for gradient 1 (H1 vs L1) and gradient 2 (H2 vs L2). Genomic context is described as the number of TE pieces in blocks (non-nested Supp. Figure S2-B), the distance and orientation (3' or 5' on the + or - strand) to the closest gene, the gene ID and its function (NA=Non Available).

Type of insertion	Chr	Position	Family / Individual copy name	Super Family	FreqH1	FreqH2	FreqL1	FreqL2	F_{ST} grad1	F_{ST} grad2	No. TE pieces in block	Distance to gene	Direction from gene	Gene strand	Gene ID	Gene function
Reference	1	23673474	RLG00003	RLG	0.761	0.672	0.000	0.000	0.614	0.506	5	-28480	5'	-	Zm00001d028130	DeSI-like protein
Reference	1	58284839	DTX12870	DTX	0.759	0.740	0.000	0.000	0.612	0.587	1	851	5'	+	Zm00001d029108	Protein STRICTOSIDINE SYNTHASE-like 11
Reference	1	150127044	DTH00163	DTH	0.057	0.000	1.000	0.717	0.892	0.559	1	132	3'	-	Zm00001d030638	PsbP domain-containing protein 1 chloroplastic
Reference	1	170241433	DHH00002	DHH	0.000	0.100	0.771	0.880	0.627	0.609	1	0	NA	-	Zm00001d030973	NA
Reference	1	170311220	DTH12326	DTH	0.056	0.071	0.875	0.975	0.674	0.819	1	23047	5'	+	Zm00001d030975	Indole-3-pyruvate monooxygenase YUCCA2
Reference	1	170316753	RLG02017	RLG	0.063	0.000	1.000	1.000	0.881	1.000	1	14940	5'	+	Zm00001d030975	Indole-3-pyruvate monooxygenase YUCCA2
Reference	1	182813390	RLC00148	RLC	0.000	0.076	0.741	0.856	0.589	0.611	1	0	NA	-	Zm00001d031231	Protein kinase superfamily protein
Reference	1	192362484	DTA00267	DTA	0.167	0.145	1.000	1.000	0.714	0.747	1	3372	3'	-	Zm00001d031511	NA
Reference	1	228929716	RLG00007	RLG	0.000	0.177	1.000	0.917	1.000	0.552	1	-2819	5'	-	Zm00001d032508	NA
Reference	1	235750205	RLG00001	RLG	0.030	0.000	0.863	0.743	0.702	0.591	9	14587	5'	+	Zm00001d032710	Vacuolar protein sorting-associated protein 2 homolog 3
Reference	1	237615972	RLC00019	RLC	0.000	0.022	1.000	0.756	1.000	0.567	1	2655	3'	-	Zm00001d032763	Pre-mRNA-processing factor 19 homolog 2
Reference	1	238216476	RLG00567	RLG	0.000	0.000	0.713	0.903	0.554	0.823	1	0	NA	-	Zm00001d032790	plastid transcriptionally active 3
Reference	1	266792210	RIL00004	RIL	0.775	1.000	0.000	0.000	0.633	1.000	1	-33824	3'	+	Zm00001d033549	NA
Reference	2	182516733	DTA00117	DTA	0.875	0.833	0.000	0.000	0.778	0.714	1	-487	3'	+	Zm00001d005658	trehalose-6-phosphate phosphatase2
Reference	2	200424852	DTH00320	DTH	0.201	0.262	1.000	1.000	0.665	0.585	1	-2973	5'	-	Zm00001d006171	Protein DETOXIFICATION 50
Reference	2	200570937	DTH00437	DTH	0.000	0.000	1.000	0.648	1.000	0.479	1	-41316	5'	-	Zm00001d006173	BAG family molecular chaperone regulator 1
Reference	2	204592932	DTH00102	DTH	0.796	1.000	0.000	0.351	0.661	0.480	3	301	5'	+	Zm00001d006323	Histone-lysine N-methyltransferase ATX4
Reference	2	224435524	DTA00263	DTA	0.318	0.000	1.000	0.826	0.517	0.704	1	0	NA	+	Zm00001d007193	SNARE-like superfamily protein
Reference	3	5298285	DTH00409	DTH	0.828	0.883	0.117	0.000	0.507	0.791	1	-1206	5'	-	Zm00001d039475	Protein ALWAYS EARLY 3
Reference	3	8735097	DTT10009	DTT	0.786	0.712	0.000	0.000	0.647	0.553	1	1464	3'	-	Zm00001d039601	NA
Reference	3	69774689	RLG00001	RLG	0.688	0.729	0.000	0.000	0.524	0.574	11	61971	5'	+	Zm00001d040839	Heavy metal transport/detoxification superfamily protein
Reference	3	96907620	RLG00001	RLG	0.833	0.643	0.000	0.000	0.714	0.474	5	70403	5'	+	Zm00001d041092	Mitogen-activated protein kinase kinase 3
Reference	3	101307350	RLG00001	RLG	1.000	0.833	0.000	0.000	1.000	0.714	3	-10297	3'	+	Zm00001d041165	Calcium-dependent protein kinase 24
Reference	3	115243093	DTH11798	DTH	0.820	0.790	0.000	0.111	0.695	0.466	1	-3386	3'	+	Zm00001d041384	NA
Reference	3	183979301	DTH00049	DTH	0.674	0.879	0.000	0.000	0.508	0.784	1	-12955	3'	+	Zm00001d042920	WUSCHEL-related homeobox 2
Reference	3	184000556	RIL00004	RIL	0.865	0.738	0.000	0.000	0.762	0.585	1	508	3'	-	Zm00001d042922	Probable mediator of RNA polymerase II transcription subunit 37c
Reference	3	186806849	RLG00008	RLG	0.000	0.000	0.784	0.935	0.645	0.878	1	0	NA	+	Zm00001d043022	Sorting nexin 2B
Reference	4	60709060	RLC00032	RLC	0.000	0.125	0.907	1.000	0.830	0.778	1	0	NA	-	Zm00001d050023	Probable WRKY transcription factor 32
Reference	4	67495360	DHH00002	DHH	0.033	0.154	1.000	1.000	0.936	0.733	1	-4703	5'	-	Zm00001d050123	GrpE protein homolog

Type of insertion	Chr	Position	Family / Individual copy name	Super Family	FreqH1	FreqH2	FreqL1	FreqL2	F_{ST} grad1	F_{ST} grad2	No. TE pieces in block	Distance to gene	Direction from gene	Gene strand	Gene ID	Gene function
Reference	4	71978762	DTH12393	DTH	0.068	0.000	0.881	0.647	0.663	0.478	1	-271	3'	+	Zm00001d050190	NA
Reference	4	81938941	RLC00032	RLC	1.000	0.700	0.000	0.000	1.000	0.538	13	63	5'	+	Zm00001d050334	Lycopene beta/epsilon cyclase protein
Reference	4	97922629	RLG00008	RLG	0.100	0.250	1.000	1.000	0.818	0.600	3	33484	5'	+	Zm00001d050535	UPF0451 C17orf61-like protein
Reference	4	104007717	DTA00102	DTA	0.093	0.079	0.917	0.865	0.679	0.620	1	288	3'	-	Zm00001d050610	Nuclear transport factor 2 (NTF2) family protein with RNA binding (RRM-RBD-RNP motifs) domain
Reference	4	104315974	RLC00064	RLC	0.216	0.174	1.000	1.000	0.645	0.704	1	0	NA	+	Zm00001d050612	DNA repair protein RAD50
Reference	4	163122639	DTA00192	DTA	0.079	0.071	0.775	0.933	0.495	0.743	1	5726	5'	+	Zm00001d051571	chromatin modification MEAF6-like protein
Reference	4	173867370	DHH00002	DHH	0.000	0.100	0.833	0.904	0.714	0.646	1	0	NA	+	Zm00001d051895	NA
Reference	4	176208733	DTH12389	DTH	0.000	0.000	0.757	0.753	0.609	0.604	1	0	NA	+	Zm00001d051998	AP-1 complex subunit gamma-1
Reference	4	179526591	RLC00032	RLC	0.000	0.195	0.913	0.925	0.840	0.541	1	0	NA	+	Zm00001d052098	Probable NOT transcription complex subunit VIP2
Reference	4	210581535	RLG00007	RLG	0.000	0.000	0.691	0.791	0.528	0.654	21	-3181	5'	-	Zm00001d053048	Probably inactive leucine-rich repeat receptor-like protein kinase
Reference	4	224034220	DTA00117	DTA	0.000	0.000	0.803	0.655	0.671	0.487	1	1199	3'	-	Zm00001d053284	Calcium-dependent lipid-binding (CaLB domain) family protein
Reference	5	1268756	DTH00067	DTH	0.155	0.077	0.910	0.742	0.572	0.457	1	247	5'	+	Zm00001d012861	ALBINO3-like protein 1 chloroplastic
Reference	5	9565456	RLC00014	RLC	0.189	0.000	1.000	0.639	0.682	0.470	21	2084	5'	+	Zm00001d013361	zinc ion binding
Reference	5	24617755	RLC00030	RLC	0.733	0.794	0.000	0.000	0.579	0.658	1	56313	5'	+	Zm00001d013914	Membrane steroid-binding protein 1
Reference	5	39613578	RLG00001	RLG	0.000	0.000	0.760	1.000	0.613	1.000	1	41371	5'	+	Zm00001d014290	Leucine aminopeptidase 2 chloroplastic
Reference	5	47701230	DTC00045	DTC	0.162	0.000	1.000	0.795	0.721	0.660	3	43121	3'	-	Zm00001d014449	Chromo domain-containing protein LHP1
Reference	5	47761681	RLG00001	RLG	0.121	0.250	1.000	0.971	0.784	0.547	1	-37	5'	-	Zm00001d014449	Chromo domain-containing protein LHP1
Reference	5	66229771	RLC00032	RLC	0.000	0.198	1.000	0.889	1.000	0.481	1	0	NA	-	Zm00001d014866	Callose synthase 3
Reference	5	66326728	RLG00003	RLG	0.000	0.062	0.833	0.819	0.714	0.581	1	13263	5'	+	Zm00001d014868	NA
Reference	5	69303223	DTX10177	DTX	0.875	0.762	0.100	0.000	0.601	0.616	1	15509	5'	+	Zm00001d014943	Protein transport protein SEC16B homolog
Reference	5	74677867	RLG00008	RLG	0.070	0.172	0.907	1.000	0.701	0.706	1	121778	5'	+	Zm00001d015082	transducin family protein / WD-40 repeat family protein
Reference	5	74696996	RLG00008	RLG	0.101	0.103	0.900	0.829	0.638	0.530	7	98201	5'	+	Zm00001d015082	transducin family protein / WD-40 repeat family protein
Reference	5	97162897	RLG00001	RLG	0.000	0.106	0.833	0.944	0.714	0.704	3	71063	5'	+	Zm00001d015553	NA
Reference	5	105630925	RLG00001	RLG	0.000	0.000	0.989	0.967	0.978	0.936	3	-305693	3'	+	Zm00001d015679	NA
Reference	5	107147795	RLX03639	RLX	0.000	0.000	0.700	0.808	0.538	0.678	1	-23132	5'	-	Zm00001d015683	NA
Reference	5	108009019	RLG00090	RLG	0.000	0.000	1.000	0.875	1.000	0.778	1	-152961	5'	-	Zm00001d015690	NA
Reference	5	110291455	RLG00003	RLG	0.000	0.000	0.875	0.900	0.778	0.818	1	-93963	3'	+	Zm00001d015711	NA
Reference	5	111121759	DTH00102	DTH	0.000	0.000	0.900	0.696	0.818	0.534	1	1023	5'	+	Zm00001d015715	NA
Reference	5	111787416	RLG00001	RLG	0.136	0.196	1.000	0.913	0.761	0.520	3	55586	3'	-	Zm00001d015718	NA
Reference	5	121519230	DTA00135	DTA	0.056	0.250	1.000	0.944	0.894	0.500	1	-3366	5'	-	Zm00001d015804	ADP-ribosylation factor GTPase-activating protein AGD12
Reference	5	137815554	RLC01507	RLC	0.000	0.000	1.000	0.720	1.000	0.563	3	0	NA	-	Zm00001d015992	NA

Type of insertion	Chr	Position	Family / Individual copy name	Super Family	FreqH1	FreqH2	FreqL1	FreqL2	F_{ST} grad1	F_{ST} grad2	No. TE pieces in block	Distance to gene	Direction from gene	Gene strand	Gene ID	Gene function
Reference	5	152222310	DTH00431	DTH	0.743	1.000	0.000	0.125	0.591	0.778	1	24600	3'	-	Zm00001d016242	Probable leucine-rich repeat receptor-like serine/threonine-protein kinase
Reference	5	184417751	RLG00007	RLG	0.000	0.000	1.000	0.941	1.000	0.889	31	-1589	5'	-	Zm00001d017072	Protein NBR1 homolog
Reference	5	184462450	DTH00327	DTH	0.000	0.250	1.000	1.000	1.000	0.600	1	-3150	3'	+	Zm00001d017076	NA
Reference	5	197600008	RLC00030	RLC	0.000	0.000	1.000	0.944	1.000	0.894	7	5963	5'	+	Zm00001d017497	Zinc finger protein ZAT9
Reference	5	209368203	DTH00389	DTH	0.000	0.370	1.000	1.000	1.000	0.460	1	-189	5'	-	Zm00001d017876	Phosphatidylinositol/phosphatidylcholine transfer protein SFH13
Reference	6	31105942	RLG00003	RLG	0.000	0.000	0.659	0.709	0.491	0.549	3	2445	5'	+	Zm00001d035526	Kinesin-like protein KIN-5D
Reference	6	44440786	DHH00002	DHH	0.725	0.746	0.000	0.000	0.569	0.595	1	0	NA	+	Zm00001d035741	Ras-related protein RABD1
Reference	6	46340343	DTH00154	DTH	0.071	0.000	1.000	0.921	0.867	0.854	1	363	5'	+	Zm00001d035763	Protein MEI2-like 1
Reference	6	48589646	RLG00008	RLG	0.063	0.000	1.000	0.910	0.881	0.835	1	-88702	3'	+	Zm00001d035783	Heavy metal transport/detoxification superfamily protein
Reference	6	49257317	DTA00061	DTA	0.000	0.000	0.838	0.929	0.721	0.867	1	534	3'	-	Zm00001d035789	Histidine--tRNA ligase cytoplasmic
Reference	6	51820972	DTH12305	DTH	0.056	0.321	1.000	1.000	0.894	0.514	1	-94946	5'	-	Zm00001d035814	GBF-interacting protein 1
Reference	6	53367718	RLG00008	RLG	0.095	0.000	0.845	0.965	0.565	0.932	3	99202	5'	+	Zm00001d035823	Tetratricopeptide repeat (TPR)-like superfamily protein
Reference	6	54306575	DHH00002	DHH	0.000	0.050	1.000	0.957	1.000	0.823	3	4170	5'	+	Zm00001d035838	60S ribosomal protein L39-1
Reference	6	54814498	RLG03877	RLG	0.000	0.143	1.000	1.000	1.000	0.750	1	6637	3'	-	Zm00001d035842	Callose synthase 5
Reference	6	54833197	DTH12393	DTH	0.149	0.073	1.000	1.000	0.741	0.864	1	-463	5'	-	Zm00001d035842	Callose synthase 5
Reference	6	55399902	RLG00001	RLG	0.000	0.000	0.750	0.700	0.600	0.538	9	55230	5'	+	Zm00001d035845	NA
Reference	6	79220842	RLG00001	RLG	0.000	0.000	0.663	0.833	0.496	0.714	13	-24885	3'	+	Zm00001d036242	ZCN15
Reference	6	114745805	DTH00434	DTH	0.000	0.000	0.734	0.728	0.580	0.572	1	0	NA	+	Zm00001d037170	Putative bZIP transcription factor superfamily protein
Reference	6	145318211	DTA00165	DTA	0.944	0.625	0.204	0.000	0.560	0.455	1	13167	3'	-	Zm00001d038018	Glutaredoxin family protein
Reference	6	158800590	RIT00001	RIT	0.089	0.125	1.000	1.000	0.837	0.778	1	0	NA	-	Zm00001d038504	Putative B3 domain-containing protein
Reference	6	170945940	DTA00102	DTA	0.000	0.000	1.000	0.969	1.000	0.940	1	435	5'	+	Zm00001d039138	GBF-interacting protein 1
Reference	7	39291581	RLG00008	RLG	0.659	0.735	0.000	0.000	0.491	0.581	5	38417	5'	+	Zm00001d019518	Photosystem I reaction center subunit IV A
Reference	7	73062093	RLG00008	RLG	0.020	0.357	0.768	1.000	0.586	0.474	1	1911	3'	-	Zm00001d019887	NA
Reference	7	78494978	DTH16373	DTH	1.000	1.000	0.312	0.000	0.524	1.000	1	8688	3'	-	Zm00001d019945	Glyceraldehyde-3-phosphate dehydrogenase GAPCP1 chloroplastic
Reference	7	144510536	DTH00057	DTH	0.250	0.000	1.000	0.879	0.600	0.784	1	23763	5'	+	Zm00001d021164	Transcription factor bHLH28
Reference	7	159192313	DTH12388	DTH	0.000	0.000	0.915	0.929	0.843	0.867	1	-1216	5'	-	Zm00001d021655	WAT1-related protein
Reference	7	161787602	RLC00148	RLC	0.262	0.000	1.000	0.790	0.585	0.653	3	-6763	5'	-	Zm00001d021736	Cellulose synthase-like protein D3
Reference	7	174848301	DTH00194	DTH	0.708	0.717	0.000	0.000	0.548	0.559	1	-577	3'	+	Zm00001d022307	10 kDa chaperonin
Reference	8	1094888	RLG00007	RLG	0.000	0.000	0.775	0.761	0.633	0.614	3	6520	5'	+	Zm00001d008206	Transcription factor MYB86
Reference	8	16673107	RLG00001	RLG	0.679	1.000	0.000	0.000	0.514	1.000	3	76471	5'	+	Zm00001d008671	Cycloartenol synthase
Reference	8	36114703	DHH00002	DHH	0.026	0.211	0.813	0.923	0.636	0.516	1	78363	5'	+	Zm00001d009100	Acid phosphatase/vanadium-dependent haloperoxidase-related protein
Reference	8	60877833	RLC00002	RLC	0.000	0.204	1.000	0.882	1.000	0.463	9	-89182	3'	+	Zm00001d009374	DNA replication licensing factor MCM4

Type of insertion	Chr	Position	Family / Individual copy name	Super Family	FreqH1	FreqH2	FreqL1	FreqL2	F_{ST} grad1	F_{ST} grad2	No. TE pieces in block	Distance to gene	Direction from gene	Gene strand	Gene ID	Gene function
Reference	8	63037462	RLG01208	RLG	0.000	0.000	1.000	1.000	1.000	1.000	1	53044	5'	+	Zm00001d009403	NA
Reference	8	73802203	DHH00002	DHH	0.000	0.000	0.919	0.750	0.850	0.600	1	791	3'	-	Zm00001d009626	protein phosphatase homolog9
Reference	8	114259086	DTH00119	DTH	0.112	0.000	1.000	0.914	0.799	0.842	1	73548	3'	-	Zm00001d010420	NA
Reference	8	131399594	RLC00032	RLC	0.000	0.050	1.000	0.736	1.000	0.493	1	0	NA	-	Zm00001d010868	Histidine--tRNA ligase cytoplasmic
Reference	8	141251550	DHH00002	DHH	0.855	0.784	0.000	0.000	0.747	0.645	1	-245	5'	-	Zm00001d011159	Ubiquitin-conjugating enzyme family protein
Reference	8	173026721	DTH00280	DTH	0.833	0.824	0.000	0.000	0.714	0.701	1	0	NA	+	Zm00001d012362	O-fucosyltransferase family protein
Reference	9	16233421	DTH00093	DTH	1.000	0.833	0.050	0.143	0.905	0.476	1	2962	5'	+	Zm00001d045211	Zinc finger protein ZAT11
Reference	9	41635183	RLG00330	RLG	0.826	1.000	0.071	0.300	0.576	0.538	35	20985	3'	-	Zm00001d045823	bZIP transcription factor family protein
Reference	9	49191495	DTH00102	DTH	0.000	0.000	0.900	0.647	0.818	0.478	1	0	NA	-	Zm00001d045951	NA
Reference	9	86369227	DTH11714	DTH	1.000	0.750	0.262	0.000	0.585	0.600	1	163752	5'	+	Zm00001d046385	NA
Reference	9	88330078	RLG00089	RLG	1.000	0.923	0.289	0.167	0.552	0.576	5	-96963	3'	+	Zm00001d046416	Putative cytochrome P450 superfamily protein
Reference	9	108795309	RLG00001	RLG	0.000	0.000	1.000	0.927	1.000	0.864	3	-1838	5'	-	Zm00001d046883	Diphthamide biosynthesis protein 3
Reference	9	113653857	DTH12390	DTH	0.065	0.242	0.775	1.000	0.517	0.610	1	-87	5'	-	Zm00001d046973	conserved oligomeric Golgi complex component-related / COG complex component-related
Reference	9	120464075	DTT00053	DTT	0.333	0.250	1.000	1.000	0.500	0.600	1	-1319	5'	-	Zm00001d047162	UDP-Glycosyltransferase superfamily protein
Reference	9	120886713	DTT00027	DTT	0.125	0.000	1.000	1.000	0.778	1.000	1	-347	3'	+	Zm00001d047170	hAT dimerisation domain-containing protein / transposase-related
Reference	9	152888848	DHH00002	DHH	0.000	0.000	1.000	0.686	1.000	0.522	1	-2374	3'	+	Zm00001d048233	F-box protein FBX14
Reference	9	158748466	RLX00219	RLX	0.690	1.000	0.000	0.000	0.527	1.000	1	-569	3'	+	Zm00001d048561	p-protein
Reference	10	33631214	DTH12996	DTH	0.851	1.000	0.000	0.100	0.741	0.818	1	86329	5'	+	Zm00001d023978	Cytochrome b5 isoform B
Reference	10	57743550	RLG00007	RLG	0.000	0.042	0.934	0.844	0.876	0.652	5	-85294	5'	-	Zm00001d024220	Putative HLH DNA-binding domain superfamily protein
Reference	10	62655082	DHH00002	DHH	0.338	0.196	1.000	0.894	0.495	0.491	1	28400	5'	+	Zm00001d024298	Histone-lysine N-methyltransferase ASHR2
Reference	10	73706985	DTM12508	DTM	0.000	0.000	1.000	0.750	1.000	0.600	1	1028	3'	-	Zm00001d024486	terpene synthase10
Reference	10	75847298	RLG00007	RLG	0.000	0.000	0.711	1.000	0.552	1.000	7	3713	5'	+	Zm00001d024520	RING/U-box superfamily protein
Reference	10	92440984	DTT12875	DTT	1.000	1.000	0.324	0.271	0.511	0.574	1	-108291	5'	-	Zm00001d024885	Putative WD40-like beta propeller repeat family protein
Reference	10	93361434	RLG00009	RLG	0.083	0.125	0.824	1.000	0.554	0.778	1	5549	3'	-	Zm00001d024893	NA
Reference	10	94430445	DTH10182	DTH	0.806	0.700	0.000	0.000	0.675	0.538	1	259	3'	-	Zm00001d024909	CO CO-LIKE TIMING OF CAB1 protein domain1
Reference	10	96457981	RLG00007	RLG	0.069	0.000	1.000	0.653	0.871	0.485	17	3324	5'	+	Zm00001d024948	Extra-large GTP-binding protein 3
Reference	10	137240817	DTT00026	DTT	0.288	0.000	1.000	0.633	0.553	0.463	1	-14931	3'	+	Zm00001d026055	Putative DUF604-domain containing /glycosyltransferase-related family protein
Reference	10	139293927	DTH00358	DTH	0.000	0.000	0.893	0.643	0.807	0.474	1	740	5'	+	Zm00001d026129	Cyclin-B2-4

Type of insertion	Chr	Position	Family / Individual copy name	Super Family	FreqH1	FreqH2	FreqL1	FreqL2	F_{ST} grad1	F_{ST} grad2	No. TE pieces in block
de novo	1	115491021	RST00049Zm00001d00003	RST	1.000	1.000	0.263	0.214	0.583	0.647	NA
de novo	1	117694940	RLX14861Zm00001d00001	RLX	1.000	1.000	0.021	0.030	0.958	0.941	NA
de novo	1	305014433	DHH00008Zm00001d00149	DHH	0.300	0.106	1.000	1.000	0.538	0.808	NA
de novo	2	13190735	RLG17139Zm00001d00001	RLG	1.000	0.935	0.022	0.224	0.957	0.519	NA
de novo	2	13190499	RLG17139Zm00001d00001	RLG	1.000	0.935	0.048	0.224	0.909	0.519	NA
de novo	2	13191000	RLG17139Zm00001d00001	RLG	1.000	1.000	0.022	0.111	0.957	0.800	NA
de novo	4	232251178	DHH00002Zm00001d09189	DHH	1.000	1.000	0.375	0.292	0.455	0.548	NA
de novo	7	142159381	DTT00024Zm00001d00131	DTT	1.000	0.833	0.231	0.100	0.625	0.540	NA

S4 Table. Genotyped bibliographic insertions PCR program.

TE insertion name, chromosome, location, order, superfamily, family are indicated along with the closest gene code, name. PCR conditions are provided with primer sequences, combinations used for PCR reactions (pairs or trios) with accompanying programs that contain touch-down (TD) cycles followed by standard (Std) cycles. Composition of cycles with annealing temperature (Tm, starting and ending Tm for TD program), number of cycles (cycles) and temperature (Tex) and duration (Dex) of the extension phase are provided together with the concentrations of primers, oligonucleotides and MgCl2 in the PCR mix. Predicted size of bands that we used to call genotypes (homozygous for the presence, heterozygous, homozygous for the absence) are reported.

Insertion name	Chro	Position	Size (bp)	Order	Superfamily	Family	Gene code Gene name (abbreviations)	Primer sequences (5'-3')	Primer pairs or trios	TD (Tm start-end, cycles, Tex, Dex)	Std (Tm, cycles, Tex, Dex)	Conc. Primers, dNTPs, MgCl2	Homo-pres	Hetero	Homo-abs
Tb1-ins	1	265683979	4885	LTR	RLC (Copia)	*Hopscotch*	AC233950.1_FG002 *TEOSINTE BRANCHED 1* (*Tb1*)	Tb1_F= TCGTTGATGCTTTGATGGATGG	Tb1_F/Tb1_R	TD (70°C-60°C, 15, 68°C, 210 sec)	Std (60°C, 20, 68°C, 210 sec with 5 sec incr. each cycle)		5000	5000 & 300	300
								Tb1_R= AACAGTATGATTTCATGGGACCG	Tb1_F/Tb1_IntR		Std (55°C, 35, 72°C, 90 sec)		1100	1100	null
								Tb1_IntR= CCTCCACCTCTCATGAGATCC							
Vgt1-ins	8	13521608	143	TIR	DTH (Pif/Harbinger) → MITE	*Tourist*	GRMZM2G700665 *RELATED TO AP2.7* (*Rap2.7*)	Vgt1_F= TGGAATGGATGTGAAGTGAG	Vgt1_F/Vgt_1R	TD (69°C-59°C, 10, 72°C, 90 sec)	Std (59°C, 25)	0.8 µM, 0.2 mM, 0.5 mM	331	188 & 331	188
								Vgt1_R= ATGCGTTGGCGATGAGCTT							
ZmCCT-ins	10	(2543 upstream ZmCCT ORF)	5122	TIR	DTC (CACTA)	--	GRMZM2G381691 *CO CO-LIKE TIMING OF CAB1* (*ZmCCT*)	CCT_F= GCACAAGAGAGATGGAGCATT	CCT_F/CCT_R/CCT_IntR	TD (70°C-60°C, 10, 72°C, 90 sec)	Std (60°C, 25)	0.5 µM, 0.25 mM, 2.5 mM	1376	1376 & 375	375
								CCT_R= ATTCTCAATCCAAGGTGCAG							
								CCT_IntR= ATTCTCAATCCAAGGTGCAG							

S5 Table. Genetic association values for Tb1-ins and Vgt1-ins insertions on 18 phenotypic trait measurements.

Trait, insertion name, chromosome and position are described. P-values corresponding to the structure effect and insertion effect for models incorporating a neutral genetic structure with 5 genetic groups (K5) and 11 populations (11pop) are indicated along with the P-values computed for the insertion:population interaction. Significative values are highlighted in yellow.

	Tb1 hopscotch insertion						vgt1 MITE insertion			
	K5	K5 interaction		11pop	11pop interaction		K5	K5 interaction		11pop
	p-value	Insertion pvalue	Pop: Ins p-value	p-value	Insertion pvalue	Pop: Ins p-value	P-value	Insertion pvalue	Pop: Ins p-value	p-value
Plant height	0	0	2.20E-16	0	0	0.34	0.61	0.96	2.20E-16	0.46
Height of lowest ear	0	0	3.33E-16	0	0	0.07	0.89	0.99	2.67E-15	0.21
Height of highest ear	0.11	0.07	2.20E-16	0.08	0.08	0.16	0.94	1	2.20E-16	0.62
Number of tillers	0.08	0.01	2.20E-16	0.1	0.1	0.46	0.04	0	2.20E-16	0.35
Number of lateral branches	0.54	0.47	4.60E-06	0.75	0.75	0.89	0.87	0.92	4.62E-05	0.52
Number of nodes with ears in the main tiller	0.29	0.23	3.82E-07	0.07	0.06	0	1	0.99	0	0.96
Leaf length	0.02	0.01	6.95E-09	0.07	0.07	0.03	0.61	0.66	0	0.62
Leaf width	0.1	0.08	0	0.13	0.12	0.07	0.06	0.05	0	0.02
Leaf coloration	0.33	0.3	0.06	0.4	0.4	0.54	0.64	0.64	0.18	0.7
Female flowering time	0	0	2.20E-16	0	0	2.44E-11	0.31	0.02	2.00E-16	0.01
Male flowering time	0.03	0	2.20E-16	0.02	0.02	6.06E-06	0.21	0.06	2.00E-16	0
Number of ramifications in the panicule	0.67	0.38	0.01	0.96	0.96	0.37	0.35	0.26	0.02	0.45
Number of grains per ear	0.02	0.01	3.61E-06	0.05	0.05	6.47E-06	0.78	0.74	4.36E-05	0.79
Grain length	0.13	0.18	2.20E-16	0	0	0.16	0.02	0	2.20E-16	0.34
Grain width	0.91	0.37	1.75E-14	0.8	0.8	0.36	0.43	0.28	1.84E-10	0.32
Grain weight	0.17	0.07	2.20E-16	0.04	0.04	0.05	0.8	0.69	2.20E-16	0.2
Grain coloration	0.65	0.65	0.1	0.57	0.57	0.47	0.11	0.1	0.1	0.03
Stomata density	0.16	0.35	2.18E-05	0.01	0.01	0.48	0.78	0.65	0.01	0.9

IV. GENERAL DISCUSSION AND PERSPECTIVES

When studying the wild relatives of domesticated species a series of questions related to the nature and functioning of artificial and natural selection come forcefully into picture. While the mechanism of natural selection to explain species evolution was initially constructed through observations on domesticated taxa (Darwin 1883), how exactly did artificial selection on wild populations operate at genetic and genomic levels to give rise to domesticated populations is an active field of research. The co-evolution of humans and crops through the new agroecological niches they constructed (Stitzer and Ross-Ibarra 2018) implies that traits that are adaptive in nature will not necessarily be favored by artificial selection and vice-versa (Allaby 2010). It is worthwhile to stress this point since in the maize-teosinte duo it is often the case that prolific maize research results help guide teosinte research questions and give hints to genes functions as well as their regulation under different environmental conditions (Camus-Kulandaivelu, Veyrieras et al. 2006) including various stresses (Hayano-Kanashiro, Calderón-Vásquez et al. 2009). Experimentally, maize inbred homozygous lines have been very useful in retrieving causal loci underlying phenotypic variation (Salvi, Corneti et al. 2011; Nannas and Kelly Dawe 2015). While the theoretical and technological approaches to study wild-domesticated species pairs have had a strong crop directed component, many studies have shown that the transfer of results from one system to the other is often not straightforward (Doebley 1984). In principle, since domesticated taxa are a subsample of the natural variation 'available' in their wild relatives, these relatives could hold a reservoir of naturally tested genetic combinations that could enable further adaptations of the cultivated species to new or changing environmental conditions (Wang, Yang et al. 2008; Warschefsky, Varma Penmetsa et al. 2014). In maize, it has been shown that many of the adaptations allowing certain races to thrive in highland conditions or under temperate climate were obtained through introgression from *mexicana* teosinte populations (Hufford, Gepts et al. 2011). Nevertheless it has also been pointed out that genomic background interactions as well as the architecture of genetic networks in maize might differ considerably from that of teosinte (Swanson-Wagner, Briskine et al. 2012; Wang, Chen et al. 2018) and thus preclude clean-cut predictions of the effect of teosinte alleles introgressed into maize.

IV.1 Achievements and limitations in the study of teosinte local adaptation.

Teosintes inhabit a great many environments throughout their distribution along Mexican geography (De Jesús Sánchez González, Corral et al. 2018). This sets an ideal scenario to ask whether local adaptation is at work (Hufford, Bilinski et al. 2012), the immediate question that

follows being how. Because environments are characterized by many features, a way to make sense through this richness is to concentrate on an environmental gradient, for instance, that associated to altitude. In our case, altitudinal gradients along teosintes' distribution permit characterizing teosintes genetic and phenotypic differences due to altitude *per se*, albeit in combination with clinal variation in other environmental variables specific to the localization of the orographic transects. While mainly described by humid and warm lowlands that transition to drier and colder highlands, there are several other factors that vary along these gradients such as bioavailable soil phosphorous content (Bayuelo-Jiménez and Ochoa-Cadavid 2014) and a number of unaccounted biotic factors potentially affecting and interacting with teosinte populations. The altitudinal syndrome that we describe for teosintes thus compiles influence of a great many conditions and factors.

Teosinte altitudinal syndrome is defined by earlier flowering, less tiller production, lower stomata density and larger, longer and heavier grains as populations gain elevation. We have several reasons to believe that this syndrome is the result of teosinte local adaptation. Firstly, mirrored patterns were recuperated somewhat independently along both gradients. For example, lowland extremes were located geographically at the largest distance and belonged to different genetic clusters, yet they showed similar phenotypes. Second, the suite of traits extracted as varying under spatially varying selection could in principle be selected by drivers other than those linked to altitude. Nevertheless we confirmed that such traits displayed altitudinal tendencies and their variation was better explained when including the populations altitude in addition to their identifier. Third, some of the spatially selected traits didn't follow the same trajectories as flowering time, and thus cannot be explained only as consequence of assortative mating along the gradients.

Certainly, measurements of fitness through proxies such as total plant seed production in reciprocal common gardens at the extremes of the gradients would be an ideal way to further test our local adaptation hypothesis. Unfortunately, extreme lowland populations were not able to develop and flower at higher grounds (Fustier, Martínez-Ainsworth et al. 2019). A partial way around these issues would be to contrast plant performance in common garden experiments with their *in-situ* observations. We calculated the position of our common garden locations in the environmental PCA projections produced from our sampled populations' coordinates with temperature and precipitation variables only. We verified that besides CEBAJ and SENGUA experimental fields being found at intermediate altitude, they were also in fact environmentally intermediate between *parviglumis* and *mexicana* populations. Unfortunately, we do not possess *in-situ* plant measurements at sampled populations. However, such information is in principle possible to obtain, and could help explain some of our observed patterns, such as the bell-shaped curve of

population plant height against altitude. Mid-altitude experimental fields may be environmentally closer to mid-elevation sampled populations, providing optimal conditions for such populations. The observed pattern may thus be due to some degree of genetic plasticity in teosinte height.

Finally, to strengthen our syndrome hypothesis, it would be ideal to prove that it is an integrated response of traits that results in fitness increase. In view that the number of grains was not measured at the whole plant level but rather on five ears per individual, we couldn't use it as a proxy of fitness. As a second attempt, we tried running DRIFTSEL on the suite of traits that composed our syndrome to test if they were being selected for as a group. But the complexity of the modeling hindered the MCMC convergence, urging us to reconsider a way to group these traits in smaller modules. Finally, another way to question if the trait co-variations we observed were more strongly defined by the advantage of this particular configuration rather than an inevitable outcome due to underlying genetic constraints, we could include experimental designs expressly aimed at decoupling the elements of our syndrome. For species faced with temporal or spatial climatic variability, syndromes that have been modeled by local adaptation are expected to retain a certain degree of potential restructurability (Ronce and Clobert 2012).

Numerous studies have highlighted the need to pay attention to the difficulties of carrying out genetic association studies when underlying genetic structure is not accurately estimated (Sohail, Maier et al. 2019) or indeed overlaps with the selective pressure modelling the trait of interest (Soularue and Kremer 2012). Natural teosinte populations present varying and sometimes strong genetic structure (Ross-Ibarra, Tenaillon et al. 2009). Using SSR data on 11 populations grown in common gardens, the optimum number of clusters was determined as K=5. Correcting for such genetic structure has very strong implications as to the number of SNPs associated to the different traits under observation, as compared to taking each sampled population as an independent genetic population. Although we knew each individual's population of origin and we controlled for the environment as much as technically possible (Barton, Hermisson et al. 2019), confounding underlying structure was an important issue.

Taking into account neutral structure at K=5 and a 5% FDR, we found numerous SNPs associated to phenotypes, and we concluded that the SNP list tested was indeed well pruned to represent potentially selected SNPs offering strong candidates. We further observed that the stronger two traits were correlated, the more associated SNPs they shared. In view of the extremely low LD between our small set of candidate SNP markers, our interpretation had been that such traits were probably polygenic (although polygenic scores *per se* were not estimated) and that their genetic determinants were highly pleiotropic. We thus went on to elaborate a series of tests taking

advantage of our four block, two location, two year experimental setting (inspired by Legrand, Larranaga et al. 2016) to further characterize the genetic constraints that might be at play in such patterns. This line of reasoning was greatly confronted when we tried to prove that structure correction was adequate. To do so, we randomized individual trait measurements within populations and then ran association analyzes. Such randomization breaks associations between genotypic and phenotypic variation within populations while maintaining those driven among populations. If associations persist after randomization, it indicates that they are driven by some population structure remaining within the K=5 groups. We did find a considerable amount of associations after the randomizations, urging us to reconsider our population clusters. In other words, our results suggest that hidden population structure within genetic clusters was pervasive. Hence, the choice of acceptable trade-off between over-correction (eg. 11 populations) and under-correction (eg. *K*=5) ought to be framed in accordance to the research question. On the one hand by correcting for 11 populations, we lose all associations driven by SNPs differentiated among populations. This may be particularly problematic for traits where the underlying neutral structure overlaps with the adaptive pattern (Figure 5C). On the other hand, under-correction leads to a high false positive rate. The fact that we found up to 83% of the tested SNPs associated to a phenotype suggests that many SNPs fall into that category. Association mapping approaches (especially when calculating polygenic scores where each SNP contributes subtly and in concert with others) have repeatedly been criticized for their excess false positives spawned from their inaccurate structure correction (Berg, Harpak et al. 2019; Sohail, Maier et al. 2019). With the data sets we have analyzed, we seem to have reached the limitations of the genetic association mapping approach.

IV.2 Role of inversions in local adaptation

The candidate SNPs chosen from HTS data on six populations were genotyped and analyzed on the genetic association panel composed of 11 populations as we have described in the previous section. With the objective of deducing the environmental drivers behind these SNPs selection signatures as well as their relationship to phenotypes, we further obtained their frequencies for a larger sample of 28 populations distributed along both altitudinal gradients. This allowed us to reveal a strong correlation of candidate SNPs with the environmental PC1, itself reflecting altitude. Since both teosinte subspecies inhabit hardly overlapping ecological niches (Hufford, Martínez-Meyer et al. 2012) and present genetic differentiation (Fukunaga, Hill et al. 2005), false positives could easily be called. This is why the intermediate population criteria included was important (Fustier, Brandenburg et al. 2017). While including this filter however, we found that candidate

SNPs were more often explained by isolation-by-environment than by geographical distance, which was not the case for neutral SNPs. We thus consider that frequency variation of candidate SNPs must be the expression of variation of ecological factors along the gradients.

The two gradients we used were traced traversing subspecies, which seem to hybridize only in sympatric areas (Warburton, Wilkes et al. 2011). Being that our candidate SNPs correlate to the environmental PC1 calculated from both gradients, it would be interesting to describe the clines that candidates follow as opposed to neutral SNPs. In view that both subspecies niches were still non-overlapping in the past, as recovered from environmental niche projections on last maximum glacial and last interglacial climatic layers (Hufford, Martínez-Meyer et al. 2012), shared polymorphisms are expected to mostly predate these events. An interesting possibility could be that the present boundary between these subspecies, as defined by their environmental niches, is in fact reinforced by chromosomal inversions. The following rationale details why we believe this should be tested. When comparing the six HTS populations, we found that candidate SNPs found in putative inversions *Inv1n* and *Inv4m* presented high F_{ST} values between subspecies, while *Inv9e* was highly differentiated within *mexicana*. Furthermore, candidate SNPs correlated with environmental PC1 (altitude correlated) were abundant in inversions. Teosinte inversions are known to follow clinal distribution with altitude (Fang, Pyhäjärvi et al. 2012; Pyhäjärvi, Hufford et al. 2013). In addition, we found that inversions were enriched for phenotypically associated candidate SNPs. If candidate SNPs within inversions are involved in ecological differentiation, they would have a more abrupt cline across the narrow hybrid zone as compared to candidate SNPs out of inversions, yet this prediction remains to be tested. Overall, although we highlighted a moderate amount of phenotypically associated SNPs through our population genomics driven method, we advocate that such associated SNPs are in fact strong candidates.

IV.3 Agents of genomic rearrangement

With the advent of whole genome sequencing technologies and the ever-growing list of bioinformatic tools developed to characterize TEs, researchers are more than ever confronted with their vast amount, notable diversity and the increasing number of ways TEs can impact genome evolution. Because of their various effects, ranging from drastic genomic rearrangements to tweaking of stress-dependent gene's expression, we enter an exciting era to better understand their vast source of genetic variability and the evolutionary dynamics that they have led in concert with their host genomes. In angiosperms, as inducers of genetic novelty, TEs may contribute to filling the lagging gap left by low nucleotide mutation rates and the vastness of phenotypic solution space

occupied. For example, it has been observed in *Arabidopsis thaliana* that TE variation among accessions affect gene expression and are not in linkage disequilibrium with neighboring SNPs, thus such variation would remain undetected if relying solely on SNP variation (Stuart, Eichten et al. 2016). In *A. thaliana* also, TEs show a mutation rate that doubles the genome-wide average for point mutations (Weng, Becker et al. 2019). In the attempt to understand TEs evolution through natural selection, population genomic tools designed for SNP data have often been directly transferred to TE variation analyzes. It might now be worthwhile to take a step back and question if the same assumptions (such as the rate of mutation, its randomness and the type of selection expected upon them) implicit in these models are valid for TE polymorphism (Villanueva-Cañas, Rech et al. 2017). This being said, the research questions addressed can help determine situations where such application is acceptable. Our objective was not to produce a full characterization of the TEs present in teosintes, as this would have needed a genome assembly and annotation procedure, with long read or nanopore sequencing for example (Ewing 2015; Carpentier, Manfroi et al. 2019). Instead, our aim was to provide a list of candidate TE insertions to have been singled out by spatially varying positive selection to contribute to teosinte local adaptation. To achieve this objective, we added a series of stringent filters that ensured us that the TE insertions that we analyzed were effectively true insertions found among our populations. Some of these choices, as for example taking since early stages only *de novo* insertions free of other *de novo* insertion calls in the immediate vicinity, served the dual purpose of homology of insertions among populations at the same time that they served for the purpose of selecting insertions which could, at a posterior stage, be easier to genotype through PCR. In the spirit of arriving to a trustworthy handful of candidate insertions we followed a similar yet simplified procedure from that employed by Fustier et al., in searching for candidate adaptive SNPs (Fustier, Brandenburg et al. 2017) taking as input the same raw read data. We searched F_{ST} outliers along both gradients with consistent directionality of population frequency. TE insertion polymorphism population frequencies are often studied through insertion frequency spectra to elaborate on the strength of purifying or occasionally positive selection acting upon them. In view of our methodological (use of B73 reference genome to map reads), technical (low-depth of coverage) and filter-driven (to find clean and trusty insertions) constraints, we encountered some extent of unknown bias in obtaining a curated TE data set. As a result, we could no longer rely on their frequency patterns at TE family level to reflect the kind and strength of selection behind them. We could nonetheless calculate their differentiation among extreme altitude populations followed by a careful examination of their genomic context in our teosinte data as well as the corresponding location in the B73 maize reference genome to further extract interesting and easy-to-type candidates.

We acknowledge the kind of population TE content and frequency analyzes that we propose in this work entails a number of biases that stem inevitably from the use of pooled population low-coverage short paired-end reads in combination with a reference genome belonging to a domesticated version of our focal organism. An additional difficulty inherent to our model system is the highly repetitive nature of its genome comprised by approximately 85% of TEs. Global evolutionary dynamics of TE content in teosinte cannot be ascertained due to the aforementioned biases and are thus beyond the scope of our presented work. Using our approach, we devised to pinpoint candidate TE insertions and argue that we have selected strong candidates to local adaptation. We are excited to explore these candidates through additional steps that we propose as perspectives of this work.

IV.4 Phenotypic consequences of TE insertions may be more versatile in teosintes than in maize

As transposable elements constitute undeniable players in angiosperm evolution, their roles in teosinte phenotypic variation and adaptation, as well as maize domestication through human driven selection, is of great interest. There is evidence that standing variation in teosinte has been a main purveyor of genetic variants on which domestication was enacted (Weber, Clark et al. 2007). In order to better understand why and how certain genetic variants were selected, it bears important to describe the wild gene pool from which they were taken and elaborate on how they might have been perceived as phenotypically attractive. The second part of this thesis aimed at exploring the TE facet of such questions. Some studies claim that artificial selection led to or selected for less plastic genotypes (Lorant, Pedersen et al. 2017). This is congruent with our results in that *Tb1-ins* has a much less constrained effect on teosinte phenotypes, to the degree of not even associating to branching nor tillering, whereas maize plants with the insertion seem strongly canalized to a non-branching phenotype regardless of maize variety (Studer, Zhao et al. 2011). Here, we should like to dwell on the kind of available evidences, that is, the logic of most experiments has been to introgress wild alleles into maize and register their effects. For instance, both teosinte alleles at *tb1* and in a gene at chromosome 3 are required to be introgressed in maize in order to produce a teosinte (branching) phenotype. The fact that maize plants may be rendered homozygous for such experiments doesn't mean that they constitute a neutral background in the sense that they won't necessarily resemble what we'd expect to find in teosinte. The question then remains of why and how was this locus selected by ancient farmers if *Tb1-ins* didn't have a very strong or noticeable effect on teosinte branching patterns. Perhaps the associations that we recovered of *Tb1-ins* with

traits such as female flowering time, plant height, height of the lowest ear and grain length could indicate that these could have been traits of ancient agronomic interest.

If we then turn to the *Vgt1-ins* TE insertion, we observed the expected phenotypic association with male flowering time. Maize was first domesticated in Mexican lowlands, following which posterior breeding allowed its migration and establishment to higher lands, and only afterwards did maize migrate towards higher and lower latitudes throughout the Americas (Tenaillon and Charcosset 2011; Kistler, Maezumi et al. 2018). A possible scenario is that *Vgt1-ins* conferred a noticeable advantage mostly when plants were taken to higher latitudes where day-length is significantly variable between seasons. If so, it would have been exposed to artificial selection more recently and could be less constrained by epistatic interactions than the *tb1* insertion, exerting similar effects in different genomic backgrounds. Other possible explanations to the fact that *Vgt1-ins* associates to flowering time in both maize and teosinte, but *Tb1-ins* does not, is that flowering time is determined by a complex network in both taxa and remains somewhat plastic, whereas *Tb1-ins*, a key determinant of the maize phenotype, has gained some autonomy throughout the domestication process that led to its fixation in all maize (Studer, Zhao et al. 2011). While at present we do not provide support for either explanation, our results do indicate that there ought to exist substantial differences in the regulatory networks associated to these insertions. Further comparison between these insertions aided by teosinte transcriptomic data could further refute or support these suggestions.

IV.5 Conclusions

Throughout this work, we have advanced our understanding of local adaptation across wild teosinte populations. First at the phenotypic level, our results point to the evolution of an altitudinal syndrome in spite of gene flow. This syndrome encompasses at least 10 phenotypic traits with evidence of spatially-varying selection. In addition to the traits we measured it would have been interesting to assess the variation of traits related to plant-soil interactions, since it has indeed been recently demonstrated that particularly soil phosphorous content is a key factor for local adaptation of highland teosintes that grow in volcanic soil (Aguirre-Liguori, Gaut et al. 2019).

Although we describe correlations between traits, we were unable address their drivers, whether they emerged from underlying functional constraints, or whether they resulted from an adaptive or a plastic response. Additional experiments would be necessary to answer these interesting questions in the framework of syndrome set-up. Those should include a broader sampling for phenotypic evaluation, with both a greater number of mother plants per population and

replicates per mother plants to better access their genetic values as well as a more diverse set of environments.

Our study brings one of the first illustrations of the link between genotypic and phenotypic variation at candidate SNPs previously recovered from HTS data using population genomic methods. In this purpose, we conducted common garden experiments, and recovered genomic regions seemingly involved in the determination of adaptive traits. In the near future, it would be very interesting to contrast the location of these regions to genomic location of genes displaying differential expression as well as genomic regions displaying differential methylation of histone (H3k27me3) marks between lowland and highland teosintes. Collaborators (B. Rhone, Y. Vigouroux and D. Grimanelli from IRD Montpellier) are in the process of analyzing both the transcriptomic and epigenomic landscape of four of our HTS teosinte populations (H1, H2, L1, L2).

Likewise, we would like to investigate the phenotypic effect of transposable element insertions. This work will be achieved in the coming months using the candidate insertions that we have recovered. It will bring an interesting comparison with SNPs. Noteworthy, our preliminary results on TEs with known phenotypic effects in maize suggest that mutations may affect different traits and in various manners in maize and teosintes. In other words, the adaptive nature of alleles seems to sometimes differ from one system to another, cautioning the use of wild genetic resources in maize breeding programs.

Finally, an important aspect of our work is the discovery that chromosomal inversions associate with phenotypic variation of multiple traits, indicating that they likely encompass suites of co-adapted alleles that together contribute to the establishment of an adaptive syndrome.

Altogether, our results raise interesting discussions on the challenges raised by the use (1) of population genomic tools to discover adaptive variation, (2) of natural populations in association mapping, and (3) of wild genetic resources in crop breeding.

IV.5 REFERENCES

Aguirre-Liguori, J. A., B. S. Gaut, et al. (2019). " Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (Zea mays parviglumis and Zea mays mexicana) " Molecular Ecology: 2814-2830.

Allaby, R. (2010). "Integrating the processes in the evolutionary system of domestication." Journal of Experimental Botany **61**: 935-944.

Barton, N., J. Hermisson, et al. (2019). "Why strucutre matters." eLife **8**.

Bayuelo-Jiménez, J. S. and I. Ochoa-Cadavid (2014). "Phosphorus acquisition and internal utilization efficiency among maize landraces from the central Mexican highlands." Field Crops Research **156**: 123-134.

Berg, J. J., A. Harpak, et al. (2019). "Reduced signal for polygenic adaptation of height in UK Biobank." eLife **8**: 1-47.

Camus-Kulandaivelu, L., J. B. Veyrieras, et al. (2006). "Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the Dwarf8 gene." Genetics **172**: 2449-2463.

Carpentier, M. C., E. Manfroi, et al. (2019). "Retrotranspositional landscape of Asian rice revealed by 3000 genomes." Nature Communications **10**.

Darwin, C. (1883). "The Variation of Animals and Plants under Domestication." Response: 495.

De Jesús Sánchez González, J., J. A. R. Corral, et al. (2018). "Ecogeography of teosinte." PLoS ONE.

Doebley, J. F. (1984). "Maize introgression into teosinte -- a reappraisal." Annals of the Missouri Botanical Garden **71**: 1100-1113.

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

Fang, Z., T. Pyhäjärvi, et al. (2012). "Megabase-scale inversion polymorphism in the wild ancestor of maize." Genetics **191**: 883-894.

Fukunaga, K., J. Hill, et al. (2005). "Genetic diversity and population structure of teosinte." Genetics **169**: 2241-2254.

Fustier, M.-A., N. E. Martínez-Ainsworth, et al. (2019). "Common gardens in teosintes reveal the establishment of a syndrome of adaptation to altitude." bioRxiv: 563585.

Fustier, M. A., J. T. Brandenburg, et al. (2017). "Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples." Molecular Ecology **26**: 2738-2756.

Hayano-Kanashiro, C., C. Calderón-Vásquez, et al. (2009). "Analysis of gene expression and physiological responses in three Mexican maize landraces under drought stress and recovery irrigation." PLoS ONE **4**.

Hufford, M. B., P. Bilinski, et al. (2012). "Teosinte as a model system for population and ecological genomics." Trends in Genetics **28**: 606-615.

Hufford, M. B., P. Gepts, et al. (2011). "Influence of cryptic population structure on observed mating patterns in the wild progenitor of maize (Zea mays ssp. parviglumis)." Molecular Ecology **20**: 46-55.

Hufford, M. B., E. Martínez-Meyer, et al. (2012). "Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight." PLoS ONE **7**.

Kistler, L., S. Y. Maezumi, et al. (2018). "Multiproxy evidence highlights a complex evolutionary legacy of maize in South America." Science **362**: 1309-1313.

Legrand, D., N. Larranaga, et al. (2016). "Evolution of a butterfly dispersal syndrome."

Lorant, A., S. Pedersen, et al. (2017). "The potential role of genetic assimilation during maize domestication." PLoS ONE.

Nannas, N. J. and R. Kelly Dawe (2015). "Genetic and genomic toolbox of Zea mays." Genetics **199**: 655-669.

Pyhäjärvi, T., M. B. Hufford, et al. (2013). "Complex patterns of local adaptation in teosinte." Genome Biology and Evolution **5**: 1594-1609.

Ronce, O. and J. Clobert (2012). "Dispersal syndromes." Dispersal ecology and evolution **155**: 119-138.

Ross-Ibarra, J., M. Tenaillon, et al. (2009). "Historical divergence and gene flow in the genus Zea." Genetics **181**: 1399-1413.

Salvi, S., S. Corneti, et al. (2011). "Genetic dissection of maize phenology using an intraspecific introgression library." BMC Plant Biology **11**.

Sohail, M., R. M. Maier, et al. (2019). "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies." eLife **8**.

Soularue, J.-P. and A. Kremer (2012). Assortative mating and gene flow generate clinal phenological variation in trees. BMC Evolutionary Biology.

Stitzer, M. C. and J. Ross-Ibarra (2018). "Maize domestication and gene interaction."

Stuart, T., S. R. Eichten, et al. (2016). "Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation." eLife **5**.

Studer, A., Q. Zhao, et al. (2011). "Identification of a functional transposon insertion in the maize domestication gene tb1." Nature Genetics **43**: 1160-1163.

Swanson-Wagner, R., R. Briskine, et al. (2012). "Reshaping of the maize transcriptome by domestication." Pnas **109**: 11878-11883.

Tenaillon, M. I. and A. Charcosset (2011). "A European perspective on maize history." Comptes Rendus - Biologies **334**: 221-228.

Villanueva-Cañas, J. L., G. E. Rech, et al. (2017). "Beyond SNPs: how to detect selection on transposable element insertions." Methods in Ecology and Evolution **8**(6): 728-737.

Wang, L., A. Yang, et al. (2008). "Creation of new maize germplasm using alien introgression from Zea mays ssp. mexicana." Euphytica **164**: 789-801.

Wang, X., Q. Chen, et al. (2018). "Genome-wide Analysis of Transcriptional Variability in a Large Maize-Teosinte Population." Molecular Plant **11**: 443-459.

Warburton, M. L., G. Wilkes, et al. (2011). "Gene flow among different teosinte taxa and into the domesticated maize gene pool." Genetic Resources and Crop Evolution **58**: 1243-1261.

Warschefsky, E., R. Varma Penmetsa, et al. (2014). "Back to the wilds: Tapping evolutionary adaptations for resilient crops through systematic hybridization with crop wild relatives." American Journal of Botany **101**: 1791-1800.

Weber, A., R. M. Clark, et al. (2007). "Major regulatory genes in maize contribute to standing variation in teosinte (Zea mays ssp. parviglumis)." Genetics **177**: 2349-2359.

Weng, M. L., C. Becker, et al. (2019). "Fine-grained analysis of spontaneous mutation spectrum and frequency in arabidopsis thaliana." Genetics **211**: 703-714.

ANNEX II. Superheroes and masterminds of plant domestication.

Natalia E. Martínez-Ainsworth[1] and Maud Irène Tenaillon[1]

[1]: Génétique Quantitative et Evolution – Le Moulon,
INRA - Université Paris-Sud - CNRS - AgroParisTech, Université Paris-Saclay

Corresponding author:
Maud I. Tenaillon, tenaillon@moulon.inra.fr
Ferme du Moulon, 91190 Gif sur Yvette, France
Phone 01 69 11 21 13
Fax 01 69 11 21 80

Superheroes and masterminds of plant domestication

Génétique de la domestication des plantes: une affaire de Super-héros et de “masterminds”.

Natalia E. Martínez-Ainsworth[1] and Maud Irène Tenaillon[1]

[1]: Génétique Quantitative et Evolution – Le Moulon,
INRA - Université Paris-Sud - CNRS - AgroParisTech, Université Paris-Saclay

Corresponding author:
Maud I. Tenaillon, tenaillon@moulon.inra.fr
Ferme du Moulon, 91190 Gif sur Yvette, France
Phone 01 69 33 23 34
Fax 01 69 33 23 80

Key words: Domestication syndrome, Human-mediated selection, Convergent evolution, Pace of domestication, Bottleneck, Gene flow.

Abstract: Domestication is one of the most fundamental changes in the evolution of human societies. Geographical origins of domesticated plants are inferred from archaeology, ecology and genetic data. Scenarios vary among species and include single, diffuse or multiple independent domestications. Cultivated plants present a panel of traits, the domestication syndrome that distinguish them from their wild relatives. It encompasses yield-, food usage-, and cultivation-related traits. Most genes underlying those traits are “masterminds” affecting regulation of gene networks. Phenotypic convergence of domestication traits across species or within species between independently domesticated forms rarely coincides with convergence at the gene level. We review here current data/models that propose a protracted transition model for domestication and investigate the impact of mating system, life cycle and gene flow on the pace of domestication. Finally we discuss the cost of domestication, pointing to the importance of characterizing adaptive functional variation in wild resources.

Mots clés: Syndrome de domestication, sélection humaine, évolution convergente, tempo de la domestication, goulot d’étranglement, flux de gènes.

Résumé: La domestication est l'un des changements les plus fondamentaux dans l'évolution des sociétés humaines. Les origines géographiques des plantes domestiquées sont inférées à partir de données archéologiques, écologiques et génétiques. Les scénarios de domestication varient d’une espèce à l’autre et comprennent des exemples de domestication unique, diffuse ou multiples et indépendantes. Les plantes cultivées présentent un panel de caractères, le syndrome de domestication qui les distinguent de leurs apparentés sauvages. Ce syndrome englobe des caractères liés au rendement, à l’utilisation et à la facilité de culture. La plupart des gènes qui sous-tendent ces caractères sont des «masterminds» affectant la régulation des réseaux de gènes. La convergence phénotypique des caractères de domestication qu’elle soit présente entre différentes espèces ou au sein d’une espèce entre des formes domestiquées indépendamment, coïncide rarement avec une convergence au niveau des gènes. Nous synthétisons ici les données et modèles actuels qui proposent un modèle de transition prolongée des formes sauvages vers les formes cultivées, et

s’intéressent à l'impact du système de reproduction, du cycle de vie et des flux géniques sur le tempo de la domestication. Enfin nous discutons du coût associé à la domestication, qui souligne l'importance de caractériser la variation fonctionnelle adaptative présente dans les ressources génétiques sauvages.

Introduction

Since their origin, hunting and gathering had been the primary mode of subsistence for modern humans. But around 12,000 years ago, humans switched from a hunter-gatherer lifestyle to an agricultural lifestyle. This transition in human behavioural ecology is known as "the Neolithic revolution". The Neolithic revolution has marked one of the most profound changes in human evolution. With reliable food stocks, human populations have increased, expanded, and built civilizations with environmental and cultural consequences that persist today. One of the primary drivers of this transition is the domestication of plants, a process whereby wild plants have been evolved into crop plants through human-mediated selection. Plant domestication has entailed co-dependency between humans and plants while promoting plant adaptation to a new ecological niche, the field. How complex were domestications? Where did they take place? How long did they last? These are some of the questions at the interface between archaeology, ecology and evolutionary genetics that have been until today actively debated, starting with the observations of Charles Darwin first published in 1868 in a book entitled "The Variation of Animals and Plants under Domestication".

1. What is plant domestication?

Domestication can be described as a set of consecutive stages that begins with the onset of domestication followed by an increase in frequency of a set of desirable traits (the domestication traits), and that culminates with the emergence of cultivated populations adapted to both human needs and a cultivated environment. Thereupon a first challenging task is to define a domestication syndrome, which is the subset of traits that collectively form the morphological and physiological differences between crops and their wild progenitors. Domestication traits were the very first targets of early farmers as opposed to traits selected later during crop diversification. We expect them to be fixed or nearly fixed in the cultivated forms as a result of intense human-driven positive selection.

Domesticated traits can be classified in three categories: (1) yield-related traits that affect propagule retention, shape and size – longer and more rigid stolons in cultivated potatoes, loss of seed shattering in cereals, indehiscent pods in legumes, increase in fruit size of cultivated tree species are some examples; (2) food usage-related traits such as reduction of chemical and physical defences, and reduction of propagule ornamentations that facilitate dispersal in the wild – loss of bitterness in cultivated almonds, loss/reduction of awns in rice and wheat fall in this category; (3)

cultivation-related traits that concern growth habit and loss of seed dormancy – the determinacy in bean cultivated forms and loss of seed dormancy in chickpea illustrate this last category. Domesticated plants often rely on human maintenance to ensure their reproductive success and domesticated traits are usually highly deleterious in the wild environment. For instance, propagule dissemination or seed dormancy are essential for survival in the wild but selected against in the field.

2. Single versus multiple domestications

At least 11 regions of the Old and New World can be considered as independent isolated centres for the origin of crops, several of which occur in Central and South America, Africa and South East Asia [1]. The Fertile Crescent is considered as the cradle of plant domestication with the emergence of major cereals such as wheat, barley, oats, rye, as well as lentils and chickpeas. Some of the related wild forms of these crops were cultivated before domestication. Hence Weiss et al. [2] have reported consistent evidence of granaries containing hundred thousands of wild barley and oat seeds in the Jordan Valley suggesting seed management and perhaps mass-selection predating domestication.

While attempting to determine the origins of crops using genetic data, it is not uncommon to arrive at conflicting interpretations. Recurrent gene flow among cultivated forms or between wild and cultivated gene pools for instance, may mask multiple domestication events. It is therefore important to merge multiple sources of data and assess congruence between archaeological findings and genetic analyses. Paleoclimatic reconstructions may also guide inferences on the ancient niches occupied by wild progenitors as reported for teosinte/maize landraces by Hufford et al. [3]. Along the same line Kraft et al. [4] have integrated evidence from paleobiolinguistics – the presence of words designating the cultivated species in an ancestral language being indicative of its importance – as a complementary geographical grid layer to that of genetic diversity and environmental niche projections in order to help refine the location of chili pepper (*Capsicum annuum*) domestication in Mexico.

Factors such as the distribution area of the crops wild progenitors as well as the rapidity of crops spread outside their center of origin have likely contributed to the emergence of the 3 described alternative domestication scenarios: a domestication event from a single gene pool in a restricted area, the best example so far being maize [5]; a diffuse domestication from wild gene

pool(s) distributed in a broader area, pearl millet domestication in the Sahel zone illustrates this situation [6] along with barley with the recent discovery that the genome of cultivated barley is a mosaic of several wild source populations [7]; multiple domestications in geographically distinct areas. Examples of the latter include the common bean, which was domesticated independently in Mexico and the Andes from two divergent gene pools [8] as well as Asian rice with two perhaps even three independent domestications [9].

3. Determinants of the domestication syndrome

Most domesticated genes so far were detected through the so-called top-down approach from phenotype to genotype. Crosses between wild and cultivated forms and examination of co-segregation of genetic markers and phenotypes in the offspring of these crosses (Quantitative Trait Loci mapping) have recovered a number of candidate regions. Genes in these regions were further identified by a combination of fine mapping, association mapping, and functional analyses including mutant complementation and gene expression assays. Analyses of patterns of polymorphism aiming at seeking footprints of selection in cultivated samples are also often performed to corroborate molecular evidence.

Table 1 presents the current domestication genes/loci list with their corresponding functional annotations. A prime example of a major domesticated gene is the teosinte branched 1 (tb1) gene. First identified from QTL mapping as a major determinant of the differences in inflorescence morphology and plant architecture between maize and teosinte, construction of a near-isogenic line containing the teosinte QTL in a maize background failed to complement the maize *Tb1* mutant allele [10]. The gene was cloned via transposon tagging, belonged to the TCP family of transcription regulator. Expression patterns were consistent with overexpression of the maize allele in the lateral primordia inducing a strong apical dominance with reduced lateral branches and feminization of the lateral terminal inflorescences [11]. Further comparison of the wild and cultivated alleles revealed a drastic reduction of diversity from 5’ UTR to 60-90 kb upstream the gene [12]. More recent work from the same team has revealed selection from standing variation at a *Hopscotch* transposable element situated in the *tb1* regulatory region. This element enhances *Tb1* expression in the cultivated form.

While Single Nucleotide Polymorphisms are the most frequently reported changes, additional examples present evidence of transposable elements being the causative domestication mutations.

Hence, a 4.1kb retrotransposon insertion in the *PvTFL1y* gene provokes growth determinacy in common bean [13], and a *Helitron* insertion is found in *barren stalk1* (*ba1*) maize gene, which regulates together with *Tb1* vegetative lateral meristem development and patterning of inflorescences [14]. One of the most recent and interesting discoveries of a domestication gene was found by Müller et al. 2016 [15]; it is a 3 bp deletion in the coding sequence of the *Arabidopsis EID1* homologous gene of cultivated tomatoes. The domesticated allele noticeably delays the phase of the circadian clock by three hours on average. *EID1* controls the network of genes that allows anticipating daily and seasonal changes and better synchronizing physiological processes. The adaptive advantage of the cultivated allele may be linked to the completion of tomato domestication outside its ancestral native range where it encountered longer days and evolved light-related damage avoidance [15].

Once causal mutations have been pinpointed, it is inevitable to wonder what kind of genes is most prevalent. Are domestication genes superheroes (structural genes) or masterminds (genes controlling regulatory network readjustments)? So far, most phenotypic changes associated with domestication seem to be orchestrated by mutations in regulatory genes (Table 1). Considering that transcription factors represent ~5% of the genes in the model species *Arabidopsis* [16], this observation is puzzling and may indicate, as Doebley [23] pointed out, that domestication is a process of genetic tinkering as opposed to genetic disassembling. In other words, domestication seems to have involved re-orchestration of gene networks and their expression by targeting "masterminds" rather than via the accumulation of null or loss-of-function mutations.

4. Genetic or phenotypic convergence?

Although different species were domesticated in different geographical locations at various times through history, it is possible to identify similar outcomes in their phenotypes which is termed phenotypic convergence – see Tenaillon and Manicacci [17]. It is of interest to pry into the nature of the genetics behind these traits, not only to better understand how adaptation proceeds, but also to address questions about the degree of genetic convergence in the evolutionary paths underlying convergent phenotypes. Was the same set of orthologous genes involved in the acquisition of similar traits among species? Were the same genes targeted by mutations within species when multiple domestications took place?

In most cases, convergence at the genetic level (Figure 1) has found little support amongst cultivated lineages. Hence as exemplified with barley non-brittle rachis, several different genes can confer similar phenotypes [18]. Exceptions include recurrent selection of orthologous genes encoding loss of seed shattering at the *Sh1* gene in sorghum, and at the *OsSh1* and *ZmSh1* genes in rice and maize respectively [19]. In the common bean, the genome scan by Schmutz et al. [8] found 59 shared domestication candidate genes between the Mesoamerican and Andean gene pools, representing 3% and 8% of each pool's candidates respectively. Kwak et al. [20] actually reported independent selection events on the *PvTFL1y* gene in each common bean gene pool. At a finer scale, different mutations may be observed on the same gene, resulting in similar domestication phenotypes as in the case of rice *Bh4* gene that generates white-hulled seeds [21]. Interestingly, such examples of repeated evolution on the same genes or orthologous genes across species are more often observed during crop diversification than domestication [22].

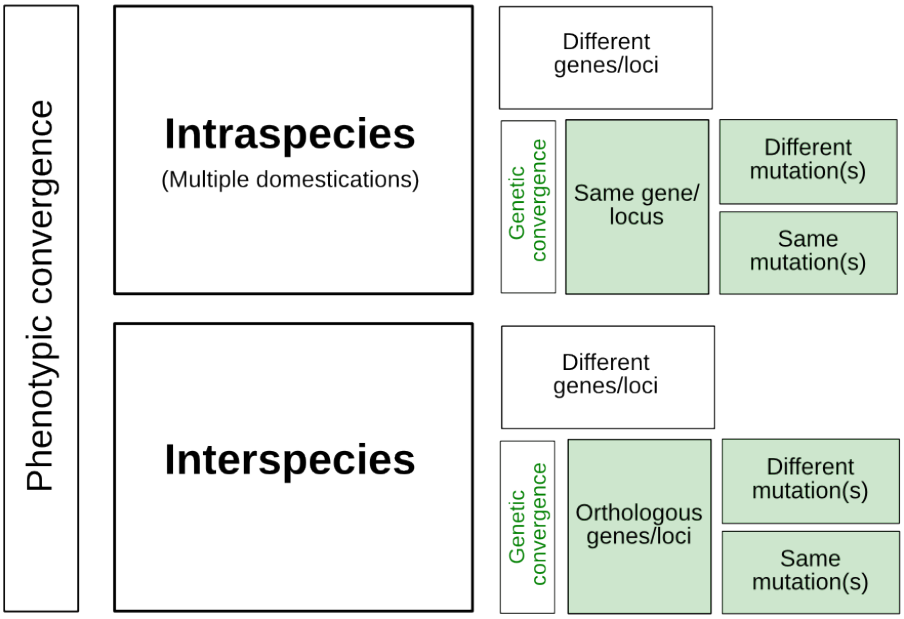


Figure 1. Levels of genetic convergence associated to phenotypic convergence in one or more species.

5. What is the pace of domestication?

First thought to be a rapid process that must have presented an immediate advantage for the early farmers, domestication process has now endorsed the status of a slow transition from wild to domesticated plants cultivation. Archaeological studies hence report the persistence within a given site of wild and cultivated forms over long time period with a slow increase of the latter. Hence,

Fuller et al. [23] have established that fixation of the non-shattering phenotype in barley, einkorn and emmer extended over a period of 2,000 to 2,500 years. In rice, there is evidence of a mix of wild (shattering) and cultivated (non-shattering) rice in Chinese sites from the lower Yangtze valley with a gradual increase of the domesticated forms from 27% (4900 BC) to 39% in 300 years [24]. But such patterns may vary from one site to another: by 6300 BC non-shattering acquisition was already complete in the middle Yangtze, suggesting an accelerated process in this area as compared to the lower Yangtze.

Many factors may influence the pace of domestication across species and sites. For instance, as discussed by Fuller [25], cultural practices related to the harvest of grains have certainly played a major role. Harvesting immature grains in cereals would delay selection for domesticated phenotypes, while storage of late-harvest mature seeds for sowing the following year would instead favour non-shattering phenotypes. Life history traits, in particular annual versus perennial life cycles have also clearly impacted domestication pace. Hence the evolution of perennial cultivated forms is affected by long juvenile periods, high level of gene flow with wild relatives, and somatic mutations transmitted by clonal propagation [26]. The rate of adaptation to the cultivated environment is also dictated by the mating system, which influences the fixation time of beneficial mutations. Glémin and Ronfort [27] have demonstrated that this rate is shorter in selfers than in outcrossers when adaptation proceeds through recessive or partially recessive mutations; a recessivity expected for domesticated traits that are most likely highly deleterious in the wild. Note that the deleterious effect of such alleles must also contribute to maintain them at very low frequency, which makes the selection from standing variation less likely. Selfing is also an efficient way to protect domesticated forms from recurrent maladaptive gene flow from sympatric wild forms. Finally, population size interferes with the aforementioned predictions by modulating the efficacy of selection. Overall, domestication proceeds faster in selfers than outcrossers and faster in large population size. The first prediction is consistent with a majority of selfers found among domesticated crops [27].

6. Consequences of domestication for the genetic diversity crop

The most notable consequence of domestication is a loss of genetic diversity. This has been observed in many species and varies from roughly 20% up to 80% loss at the nucleotide level in maize [28] and durum wheat respectively [29]. Domestication is a recent enough process to detect the footprints of what is commonly called, the domestication bottleneck, a direct consequence of

selection on a subset of wild individuals/populations. This bottleneck is likely underestimated because of the recovery of diversity since domestication through mutations, population expansion, and gene flow from wild relatives. Bottleneck scenarios have been modelled in multiple domesticated species, but the impact of gene flow has been overlooked. While there is evidence for recurrent gene flow between wild and domesticated forms, a compilation suggests that the majority of crops actually possess reproductive barriers, 38% of them being linked to either ploidy differences or reduced hybrid fitness [30]. Whether these barriers can be considered as a domestication trait is still an open question.

Both shrinks in population size and to a lesser extent impact of selective sweeps on neighbouring pre-existing variations [31] have inflated the accumulation of slightly deleterious mutations, an effect magnified in poor recombining regions [32]. There is hence a cost to domestication. It can be estimated by analysing the enrichment of nonsynonymous to synonymous derived substitutions in the cultivated form with respect to the wild form. Nabholz et al. [33] have found good evidence for such enrichment in the African rice (*Oryza glaberrima*) in comparison to its wild progenitor *Oryza barthii* and further showed that it is more pronounced in regions suffering strong drift.

Conclusion

Domestication studies continue to be a fascinating ground to delve into. By combining approaches from diverse disciplines, the origins and processes accompanying crop domestications have begun to be understood. So far, research on the genetic unravelling of domestication points to modulation in the expression of mastermind genes, which in turn exert a downstream rewiring of genetic networks. Hitherto, convergence at the gene level among crops or between crops independent domestications has rarely been observed. In fact, because the pace of domestication is influenced by many intricate factors related to life history traits, population size and trait genetic determinism in combination with cultural practices, the emerging domestication patterns are truly species-specific. They span very slow to rapid transitions embedded in single or multiple domestication events. Conversely, a consequence of domestication that has been recurrently encountered is a loss of genetic diversity that stresses the importance of assessing the functional variation of wild genetic resources to broaden the usable genetic diversity in conventional breeding programs.

Acknowledgements

We thank Georges Pelletier, Bernard Dujon and André Gallais for useful comments on the manuscript as well as all the superheroes and masterminds cited in this review for their insightful and inspiring work. The work of M.I.T is supported by the Agence Nationale de la Recherche (Project ANR 12-ADAP-0002-01) and N.E.M-A is funded by a CONACYT scholarship (579966/410748).

Bibliographical reference list

[1] G. Larson, D.R. Piperno, R.G. Allaby, M.D. Purugganan, L. Andersson, M. Arroyo-Kalin, L. Barton, C. Climer Vigueira, T. Denham, K. Dobney, A.N. Doust, P. Gepts, M.T.P. Gilbert, K.J. Gremillion, L. Lucas, L. Lukens, F.B. Marshall, K.M. Olsen, J.C. Pires, P.J. Richerson, R. Rubio de Casas, O.I. Sanjur, M.G. Thomas, D.Q. Fuller, Current perspectives and the future of domestication studies, Proc. Natl. Acad. Sci. 111 (2014) 6139–6146.

[2] E. Weiss, M.E. Kislev, A. Hartmann, Autonomous Cultivation Before Domestication, Science 312 (2006) 1608–1610.

[3] M.B. Hufford, E. Martínez-Meyer, B.S. Gaut, L.E. Eguiarte, M.I. Tenaillon, Inferences from the Historical Distribution of Wild and Domesticated Maize Provide Ecological and Evolutionary Insight, PLoS One. 7 (2012).

[4] K.H. Kraft, C.H. Brown, G.P. Nabhan, E. Luedeling, J. de J. Luna Ruiz, G. Coppens d'Eeckenbrugge, R.J. Hijmans, P. Gepts, Multiple lines of evidence for the origin of domesticated chili pepper, Capsicum annuum, in Mexico, Proc. Natl. Acad. Sci. 111 (2014) 6165–6170.

[5] Y. Matsuoka, Y. Vigouroux, M.M. Goodman, J. Sanchez G, E. Buckler, J. Doebley, A single domestication for maize shown by multilocus microsatellite genotyping., Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 6080–6084.

[6] I. Oumar, C. Mariac, J.L. Pham, Y. Vigouroux, Phylogeny and origin of pearl millet (Pennisetum glaucum [L.] R. Br) as revealed by microsatellite loci, Theor. Appl. Genet. 117 (2008) 489–497.

[7] A.M. Poets, Z. Fang, M.T. Clegg, P.L. Morrell, Barley landraces are characterized by geographically heterogeneous genomic origins, Genome Biol. 16 (2015) 173.

[8] J. Schmutz, P.E. McClean, S. Mamidi, G.A. Wu, S.B. Cannon, J. Grimwood, J. Jenkins, S. Shu, Q. Song, C. Chavarro, M. Torres-Torres, V. Geffroy, S.M. Moghaddam, D. Gao, B. Abernathy, K. Barry, M. Blair, M. a Brick, M. Chovatia, P. Gepts, D.M. Goodstein, M. Gonzales, U. Hellsten, D.L. Hyten, G. Jia, J.D. Kelly, D. Kudrna, R. Lee, M.M.S. Richard, P.N. Miklas, J.M. Osorno, J. Rodrigues, V. Thareau, C. a Urrea, M. Wang, Y. Yu, M. Zhang, R. a Wing, P.B. Cregan, D.S. Rokhsar, S. a Jackson, A reference genome for common bean and genome-wide analysis of dual domestications., Nat. Genet. 46 (2014) 707–13.

[9] P. Civáň, H. Craig, C.J. Cox, T.A. Brown, Three geographically separate domestications of Asian rice, Nat. Plants. 1 (2015) 1–5.

[10] J. Doebley, A. Stec, C. Gustus, teosinte branched1 and the origin of maize: Evidence for epistasis and the evolution of dominance, Genetics. 141 (1995) 333–346.

[11] J. Doebley, A. Stec, L. Hubbard, The evolution of apical dominance in maize, Nature. 386 (1997) 485–8.

[12] R.L. Wang, a Stec, J. Hey, L. Lukens, J. Doebley, The limits of selection during maize domestication., Nature. 398 (1999) 236–239.

[13] S.L. Repinski, M. Kwak, P. Gepts, The common bean growth habit gene PvTFL1y is a functional homolog of Arabidopsis TFL1, Theor. Appl. Genet. 124 (2012) 1539–1547.

[14] S. Gupta, A. Gallavotti, G.A. Stryker, R.J. Schmidt, S.K. Lal, A novel class of Helitron- related transposable elements in maize contain portions of multiple pseudogenes, Plant Mol. Biol. 57 (2005) 115–127.

[15] N.A. Müller, C.L. Wijnen, A. Srinivasan, M. Ryngajllo, I. Ofner, T. Lin, A. Ranjan, D. West, J.N. Maloof, N.R. Sinha, S. Huang, D. Zamir, J.M. Jiménez-Gomez, Domestication selected for deceleration of the circadian clock in cultivated tomato, Nat Genet. 48 (2016) 89–93.

[16] J.L. Riechmann, O.J. Ratcliffe, A genomic perspective on plant transcription factors, Curr. Opin. Plant Biol. 3 (2000) 423–434.

[17] M.I. Tenaillon, D. Manicacci, Maize origins: an old question under the spotlights., in: Adv. Maize (Essential Rev. Exp. Biol., The Society for Experimental Biology, 2011: pp. 89–110.

[18] T. Komatsuda, P. Maxim, N. Senthil, Y. Mano, High-density AFLP map of nonbrittle rachis 1 (btr1) and 2 (btr2) genes in barley (Hordeum vulgare L.), Theor. Appl. Genet. 109 (2004) 986–995.

[19] Z. Lin, X. Li, L.M. Shannon, C.-T. Yeh, M.L. Wang, G. Bai, Z. Peng, J. Li, H.N. Trick, T.E. Clemente, J. Doebley, P.S. Schnable, M.R. Tuinstra, T.T. Tesso, F. White, J. Yu, Parallel domestication of the Shattering1 genes in cereals, Nat. Genet. 44 (2012) 720–724.

[20] M. Kwak, O. Toro, D.G. Debouck, P. Gepts, Multiple origins of the determinate growth habit in domesticated common bean (Phaseolus vulgaris), Ann. Bot. 110 (2012) 1573–1580.

[21] B.-F. Zhu, L. Si, Z. Wang, Y. Zhou, J. Zhu, Y. Shangguan, D. Lu, D. Fan, C. Li, H. Lin, Q. Qian, T. Sang, B. Zhou, Y. Minobe, B. Han, Genetic control of a transition from black to straw-white seed hull in rice domestication., Plant Physiol. 155 (2011) 1301–1311.

[22] B.L. Gross, K.M. Olsen, Genetic perspectives on crop domestication, Trends Plant Sci. 15 (2010) 529–537.

[23] D.Q. Fuller, T. Denham, M. Arroyo-Kalin, L. Lucas, C.J. Stevens, L. Qin, R.G. Allaby, M.D. Purugganan, Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record, Proc. Natl. Acad. Sci. 111 (2014) 6147–6152.

[24] D. Q Fuller, L. Qin, Z. Yunfei, Z. Zhijun, C. Xugao, H. Leo Aoi, G.-P. Sun, The domestication process and domestication rate in rice: spikelet bases from the Lower Yangtze., Science 323 (2009) 1607–1610.

[25] D.Q. Fuller, Contrasting patterns in crop domestication and domestication rates: Recent archaeobotanical insights from the old world, Ann. Bot. 100 (2007) 903–924.

[26] B.S. Gaut, C.M. Díez, P.L. Morrell, Genomics and the Contrasting Dynamics of Annual and Perennial Domestication, Trends Genet. 31 (2015) 709–719.

[27] S. Glémin, J. Ronfort, Adaptation and maladaptation in selfing and outcrossing species: New mutations versus standing variation, Evolution (N. Y). 67 (2013) 225–240.

[28] M.B. Hufford, X. Xu, J. van Heerwaarden, T. Pyhäjärvi, J.-M. Chia, R.A. Cartwright, R.J. Elshire, J.C. Glaubitz, K.E. Guill, S.M. Kaeppler, J. Lai, P.L. Morrell, L.M. Shannon, C. Song, N.M. Springer, R.A. Swanson-Wagner, P. Tiffin, J. Wang, G. Zhang, J. Doebley, M.D. McMullen, D. Ware, E.S. Buckler, S. Yang, J. Ross-Ibarra, Comparative population genomics of maize domestication and improvement, Nat. Genet. 44 (2012) 808–811.

[29] A. Haudry, A. Cenci, C. Ravel, T. Bataillon, D. Brunel, C. Poncet, I. Hochu, S. Poirier, S. Santoni, S. Glémin, J. David, Grinding up wheat: A massive loss of nucleotide diversity since domestication, Mol. Biol. Evol. 24 (2007) 1506–1517.

[30] H. Dempewolf, K. a. Hodgins, S.E. Rummell, N.C. Ellstrand, L.H. Rieseberg, Reproductive isolation during domestication, Plant Cell. 24 (2012) 2710–2717.

[31] T.M. Beissinger, L. Wang, K. Crosby, A. Durvasula, M.B. Hufford, J. Ross-ibarra, O. Biology, I. State, Recent demography drives changes in linked selection across the maize genome, bioRxiv (2015) 031666.

[32] E. Rodgers-Melnick, P.J. Bradbury, R.J. Elshire, J.C. Glaubitz, C.B. Acharya, S.E. Mitchell, C. Li, Y. Li, E.S. Buckler, Recombination in diverse maize is stable, predictable, and associated with genetic load., Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 3823–8.

[33] B. Nabholz, G. Sarah, F. Sabot, M. Ruiz, H. Adam, S. Nidelet, A. Ghesquière, S. Santoni, J. David, S. Glémin, Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima), Mol. Ecol. 23 (2014) 2210–2227.

Figure and table captions

Table 1. Selected list of genes/loci whose function/phenotype/selective patterns offer convincing evidence of their involvement in domestication.

Figure 1. Levels of genetic convergence associated to phenotypic convergence in one or more species.

Table 1.

Crop species	Common name	Gene name (**abbreviation**)	Trait	Gene type	**References**
Brassica oleracea	Broccoli *	*BoCAULIFLOWER* (***BoCAL***)	Affects floral primordia, alterations in inflorescence morphology	Transcription factor	**1,2**
Glycine max	Soybean	*SHATTERING1-5* (***SHAT1-5***)	Increased lignification of fiber cap cells leads to shattering-resistant pods	Transcription factor	**3**
Hordeum vulgare	Barley	*INTERMEDIUM-C* (***INT-C***)	Fertility of lateral spikelets and tillering	Transcription factor	**4**
Hordeum vulgare	Barley	*Nud* (***nud***)	Caryopsis with easily separable husks	Transcription factor	**5**
Hordeum vulgare	Barley	*SIX-ROWED SPIKE* (***HvVRS1***)	Development and fertility of lateral spikelet	Transcription factor	**6**
Oryza sativa	Rice	*BLACK HULL4* (***Bh4***)	Changes color of seed hull from black to white	Amino acid transporter protein	**7**
Oryza sativa	Rice	*GRAIN WIDTH5* (***GW5***)	Increase of grain size	Polyubiquitin-interacting protein	**8**
Oryza sativa	Rice	*OsLIGULELESS1* (***OsLG1***)	Alteration in laminar joint and ligule development forming closed panicles	Transcription factor	**9**
Oryza sativa	Rice	*OsPROSTRATE GROWTH1* (***PROG1***)	Tiller angle leads to erect growth (plant architecture)	Transcription factor	**10, 11**
Oryza sativa	Rice	*Red pericarp* (***Rc***)	Pericarp color	Transcription factor	**10, 12**
Oryza sativa	Rice	*SHATTERING4-1* (***sh4-1***)	Reduced seed shattering	Transcription factor	**13**
Phaseolus vulgaris	Common bean	*PvTERMINAL FLOWER1* (***PvTFL1y***)	Determinate shoots with a terminal inflorescence	Transcription cofactor	**14**
Solanum lycopersicum	Tomato	*LOCULE NUMBER* (***LC***)	Increase in the number of locules	Transcription factor	**15**
Solanum lycopersicum	Tomato	*FASCIATED* (***fas***)	Increase in the number of carpels and locules	Transcription factor	**16**
Solanum lycopersicum	Tomato	*fruit weight 2.2* (***fw2.2***)	Alteration in fruit size	Cell number regulator protein	**17**
Sorghum bicolor	Sorghum	*SbSHATTERING1* (***SbSH1***)	Non-shattering of seeds	Transcription factor	**18**
Triticum aestivum	Common wheat	*wheat AP2-like* (***WAP2***) (***Q***)	Allows free-threshing and spelt spike formation	Transcription factor	**19**
Zea mays	Maize	*BARREN STALK1* (***ba1***)	Prevents axillary meristem formation	Transcription factor	**20**
Zea mays	Maize	*brittle2* (***bt2***)	Increase in yield and different amylopectin properties	Enzyme	**21**
Zea mays	Maize	*grassy tillers1* (***gt1***)	Suppression of elongation of lateral ear branches	Transcription factor	**22**
Zea mays	Maize	*PROLAMIN-BOX BINDING FACTOR* (***PBF1***)	Unclear	Transcription factor	**23, 24**
Zea mays	Maize	*Ramosa1* (***ra1***)	Branching architecture	Transcription factor	**25**
Zea mays	Maize	*starch branching enzyme IIB* (***ae1***)	Amylopectin structure leading to starch pasting properties	Amylose extender	**21**
Zea mays	Maize	*teosinte branched 1* (***tb1***)	Apical dominance, short ear tipped branches	Transcription factor	**26 ,27**
Zea mays	Maize	*teosinte glume architecture 1* (***tga1***)	Softer glume leads to kernel exposition	Transcription factor	**28**
Zea mays	Maize	*Zea agamous-like1* (***Zagl1***)	Increase in female ear length	Transcription factor	**29, 30**
Zea mays	Maize	*ZmSHATTERING1* (***ZmSh1***)	Reduced seed shattering	Transcription factor	**18**

*Also includes the varieties: Brussels sprouts, Cabbage, Cauliflower, Kale and Kohlrabi.

1. Purugganan, M. D. & Fuller, D. Q. (2009). *Nature* 457, 843–848, **2.** Smith, L. & King, G. (2000). *Mol. Breed.* 603–613, **3.** Dong, Y. *et al.* (2014). *Nat. Commun.* 5, 3352, **4.** Ramsay, L. *et al.* (2011). *Nat Genet* 43, 169–72, **5.** Taketa, S. *et al.* (2004). *Theor. Appl. Genet.* 108, 1236–1242, **6.** Komatsuda, T. *et al.* (2007). *Proc. Natl. Acad. Sci. U. S. A.* 104, 1424–1429, **7.** Zhu, B.-F. *et al.* (2011). *Plant Physiol.* 155, 1301–1311, **8.** Shomura, A. *et al.* (2008). *Nat. Genet.* 40, 1023–1028, **9.** Ishii, T. *et al.* (2013). *Nat. Genet.* 45, 462–465, **10.** Huang, X. *et al.* (2012). *Nature* 490, 497–501, **11.** Jin, J. *et al.* (2008). *Nat. Genet.* 40, 1365–9, **12.** Sweeney, M. T., Thomson, M. J., Pfeil, B. E. & McCouch, S. (2006). *Plant Cell* 18, 283–294, **13.** Wu, X., Skirpan, A. & McSteen, P. (2009). *Plant Physiol.* 149, 205–19, **14.** Kwak, M., Toro, O., Debouck, D. G. & Gepts, P. (2012). *Ann. Bot.* 110, 1573–1580, **15.** Muños, S. *et al.* (2011). *Plant Physiol.* 156, 2244–2254, **16.** Cong, B., Barrero, L. S. & Tanksley, S. D. (2008). *Nat. Genet.* 40, 800–804, **17.** Cong, B. & Tanksley, S. D. (2006). *Plant Mol. Biol.* 62, 867–880, **18.** Lin, Z. *et al.* (2012). *Nat. Genet.* 44, 720–724, **19.** Simons, K. J. *et al.* (2006). *Genetics* 172, 547–555, **20.** Gallavotti, A. *et al.* (2004). *Nature* 432, 630–635, **21.** Whitt, S. R., Wilson, L. M., Tenaillon, M. I., Gaut, B. S. & Buckler, E. S. (2002).

Proc. Natl. Acad. Sci. U. S. A. 99, 12959–62, **22.** Whipple, C. J. *et al.* (2011). *Proc. Natl. Acad. Sci.* 108, E506–E512, **23.** Jaenicke-Després, V. *et al.* (2003). *Science* 302, 1206–1208, **24.** Lang, Z. *et*

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

Titre : Caractérisation des déterminants génomiques et des réponses phénotypiques de l'adaptation à l'altitude chez les téosintes (*Zea mays* ssp. *parviglumis* and ssp. *mexicana*)

Mots clés : Variation adaptative; Structuration génétique; Sélection spatialisée; Éléments transposables; Génétique d'association; Syndrome altitudinal

Résumé : Les deux sous-espèces annuelles de téosinte qui sont les plus proches parents sauvages du maïs sont d'excellents systèmes pour étudier l'adaptation locale car leur distribution couvre un large éventail de conditions environnementales. *Zea mays* ssp. *parviglumis* est distribuée dans un habitat chaud et mésique en dessous de 1800 m d'altitude, tandis que *Zea mays* ssp. *mexicana* prospère dans des conditions sèches et fraîches à des altitudes plus élevées. Nous avons combiné des approches d'écologie inverse et de génétique association afin d'identifier les déterminants de l'adaptation locale chez ces téosintes. A partir de données de séquençage haut débit (HTS) de six populations comprenant des populations de basses et hautes altitudes, une étude précédente a identifié un sous-ensemble de 171 polymorphismes nucléotidiques (SNP candidats) présentant des signaux de sélection. Nous avons utilisé ces SNP candidats pour tester l'association entre la variation génotypique et phénotypique de 18 caractères. Notre panel d'association était constitué de 1664 plantes provenant de graines de 11 populations échantillonnées le long de deux gradients d'altitude. Il a été évalué deux années consécutives dans deux jardins communs. Nous avons contrôlé sa structure neutre en utilisant 18 marqueurs microsatellites. La variation phénotypique a révélé l'existence d'un syndrome altitudinal composé de dix caractères. Nous avons ainsi observé une augmentation de la précocité de floraison, une diminution de la production de talles et de la densité en stomates des feuilles ainsi qu'une augmentation de la taille, de la longueur et du poids des grains avec l'élévation croissante du site de collecte des populations. Ce syndrome a évolué malgré des flux de gènes détectables entre populations. Nous avons montré que le pourcentage de SNP candidats associés aux différents caractères dépend de la prise en compte de la structure neutre soit en cinq groupes génétiques (73,7%), soit en onze populations (13,5%), indiquant une stratification complexe. Nous avons testé les corrélations entre les variables environnementales et les fréquences alléliques des SNP candidats sur 28 populations. Nous avons trouvé un enrichissement à la fois pour les SNP présentant des associations phénotypiques et les SNP présentant des corrélations environnementales dans trois larges inversions chromosomiques, confirmant leur rôle dans l'adaptation locale. Pour explorer la contribution de la variation structurale à l'évolution adaptative, nous nous sommes concentrés sur le contenu en éléments transposables (ET) des six populations séquencées (HTS). Ces éléments constituent environ 85% du génome du maïs et contribuent à sa variabilité fonctionnelle. Nous avons effectué la première description populationnelle des ET chez les téosintes pour deux catégories d'insertions, celles présentes et celles absentes du génome de référence du maïs. Nous avons ensuite recherché des polymorphismes liés aux ET présentant des fréquences alléliques contrastées entre populations de basse et de haute altitude. Nous avons identifié un sous-ensemble d'insertions candidates. Enfin, nous avons génotypé, dans un panel d'association, des insertions d'ET connues pour avoir contribué à l'évolution phénotypique du maïs. Contrairement à ce qui a été observé chez le maïs, certaines de ces insertions n'ont montré aucun effet phénotypique chez les téosintes, ce qui suggère que leur effet dépend du fond génétique. Notre étude apporte de nouvelles connaissances sur l'adaptation altitudinale chez les plantes. Elle ouvre la discussion sur les défis soulevés par l'utilisation (1) d'outils de génomique des populations pour identifier la variation adaptative, (2) de populations naturelles en génétique d'association, et (3) de ressources génétiques sauvages pour l'amélioration des espèces cultivées.

Title : Characterizing the genomic determinants and phenotypic responses to altitudinal adaptation in teosintes (*Zea mays* ssp. *parviglumis* and ssp. *mexicana*)

Keywords : Adaptive variation; Genetic structure; Spatially-varying selection; Transposable elements; Association mapping; Altitudinal syndrome

Abstract : Annual teosintes, the closest wild relatives of maize, are ideal systems to study local adaptation because their distribution spans a wide range of environmental conditions. *Zea mays* ssp. *parviglumis* is distributed in warm and mesic conditions below 1800 m, while *Zea mays* ssp. *mexicana* thrives in dry and cool conditions at higher altitudes. We combined reverse ecology and association mapping to mine the determinants of local adaptation in annual teosintes. Based on high throughput sequencing (HTS) data from six populations encompassing lowland and highland populations growing along two elevation gradients, a previous study has identified candidate regions displaying signals of selection. Within those regions a subset of 171 candidate single nucleotide polymorphisms (SNPs) was selected to test their association to phenotypic variation at 18 traits. Our association panel encompassed 1664 plants from seeds collected from eleven populations sampled along the elevation gradients. We benefit from phenotypic characterization of all the plants in two common gardens located at mid-altitude for two years. In addition, we controlled for neutral structure of the association panel using 18 microsatellite markers. Phenotypic variation revealed the components of an altitudinal "syndrome" constituted of ten traits evolving under spatially-varying selection. Plants flowered earlier, produced less tillers, displayed lower stomata density and carried larger, longer and heavier grains with increasing elevation of population collection site. This syndrome evolved in spite of detectable gene flow among populations. The percentage of candidate SNPs associated with traits largely depended on whether we corrected for five genetic groups (73.7%) or eleven populations (13.5%), thereby indicating a complex stratification in our association panel. We analyzed correlations between environmental variables and allele frequencies of candidate SNPs on a larger set of 28 populations. We found enrichment for SNPs displaying phenotypic associations and environmental correlations in three Mb-scale chromosomal inversions, confirming the role of these inversions in local adaptation. To further explore the contribution of structural variation to adaptive evolution, we focused on transposable element (TE) content of the HTS populations. TEs constitute ~85% of the maize genome and contribute to its functional variability via gene inactivation and modulation of gene expression. We performed the first population-level description of TEs in teosintes for two categories of insertions, those present and those absent from the maize reference genome. We next searched for TE polymorphisms with contrasted allele frequencies between lowland and highland populations. We pinpointed a subset of adaptive candidate insertions. Finally, we genotyped in our association panel TE insertions known to have contributed to maize phenotypic evolution. In contrast to what was found in maize, some of these insertions displayed no measurable phenotypic effects in teosintes, suggesting that their effect depends on the genetic background. Altogether our study brings new insights into plant altitudinal adaptation. It opens discussions on the challenges raised by the use (1) of population genomic tools to discover adaptive variation, (2) of natural populations in association mapping, and (3) of wild genetic resources in crop breeding.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France