




OPEN

The influence of genetic structure on phenotypic diversity in the Australian mango (*Mangifera indica*) gene pool

Melanie J. Wilkinson^{1,2}, Risa Yamashita³, Maddie E. James^{1,2}, Ian S. E. Bally⁴, Natalie L. Dillon⁴, Asjad Ali⁴, Craig M. Hardner^{3,5} & Daniel Ortiz-Barrientos^{1,2,5}

Genomic selection is a promising breeding technique for tree crops to accelerate the development of new cultivars. However, factors such as genetic structure can create spurious associations between genotype and phenotype due to the shared history between populations with different trait values. Genetic structure can therefore reduce the accuracy of the genotype to phenotype map, a fundamental requirement of genomic selection models. Here, we employed 272 single nucleotide polymorphisms from 208 *Mangifera indica* accessions to explore whether the genetic structure of the Australian mango gene pool explained variation in trunk circumference, fruit blush colour and intensity. Multiple population genetic analyses indicate the presence of four genetic clusters and show that the most genetically differentiated cluster contains accessions imported from Southeast Asia (mainly those from Thailand). We find that genetic structure was strongly associated with three traits: trunk circumference, fruit blush colour and intensity in *M. indica*. This suggests that the history of these accessions could drive spurious associations between loci and key mango phenotypes in the Australian mango gene pool. Incorporating such genetic structure in associations between genotype and phenotype can improve the accuracy of genomic selection, which can assist the future development of new cultivars.

Horticultural tree crops are vital for sustainable food production¹ and ornamental and industrial use. Tree crops can be more sustainably cultivated over time than annual field crops, thus helping to manage food supply for an increasing world population². To create new tree fruit cultivars with improved productivity and quality, we must develop breeding technologies that overcome biological limitations to their production. Tropical species, such as mango, are often large and vigorous³, leading to canopies that rapidly outgrow their orchard space. This generates shade, providing a breeding ground for disease⁴. To avoid the adverse effects of tree size, trees are traditionally planted at low density and heavily pruned each year⁴, leading to a reduction in overall production per hectare and an increased cost per unit output. Consequently, a quest to breed smaller, less vigorous trees while maintaining high yields of quality fruit is underway^{5,6}. Such efforts will produce mango that can be grown in intensive, high-density orchards that produce more fruit per hectare⁷.

Traditional tree breeding is slow, as evaluations require an assessment of phenotypic performance in mature trees over many years to account for the effects of variable spatial and temporal environments on phenotypic diversity. These evaluations, in combination with a long juvenile phase (typically 2–4 years⁴), can result in a selection process of up to or longer than 10 years from field planting⁸, making the rapid development of new cultivars unfeasible. The time for cultivar development could be reduced by predicting future phenotypic performance in young individuals using genomic selection, as demonstrated in apples⁹, sweet cherry¹⁰ and strawberry¹¹. Genomic selection uses genotype to phenotype maps from a training population to predict phenotypic variation in untested populations using marker data^{12,13}. Thus, once a genomic selection model has been created, the length and expense of phenotyping key traits may be reduced. Genomic selection for tree size and vigour

¹School of Biological Sciences, The University of Queensland, Brisbane, QLD 4072, Australia. ²Australian Research Council Centre of Excellence for Plant Success in Nature and Agriculture, The University of Queensland, Brisbane, QLD 4072, Australia. ³Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD 4072, Australia. ⁴Queensland Department of Agriculture and Fisheries, Mareeba, QLD 4880, Australia. ⁵These authors contributed equally: Craig M. Hardner and Daniel Ortiz-Barrientos. ✉email: m.wilkinson2@uq.edu.au

of progeny could therefore improve the breeding process and reduce the cost of mango breeding compared to traditional breeding approaches.

The primary assumption of genomic selection is that genetic markers are closely linked on a chromosome with the causative loci that contribute to the trait of interest¹⁴. In general, the closer the marker is to the causative loci, the more accurate the genotype to phenotype map. However, genetic structure can create statistical associations between loci that are not physically linked. This occurs because evolutionary forces such as migration, drift and mutation can make allelic combinations between unlinked loci more common than expected by chance¹⁵. Genetic structure can therefore create spurious associations between genetic markers and traits. Furthermore, genetic structure is often prevalent in modern crops, particularly those moving across the world via human migrations, which likely experienced drastic fluctuations in population size and suffered from inbreeding after crossing genetically related individuals with favourable traits¹⁶.

Differentiating uninformative loci due to genetic structure from those linked to causative loci is a common problem observed in genetic studies of human disease^{17,18} and the study of trait evolution across diverse taxa^{19,20,21,23}. Fortunately, we can improve the accuracy of the genotype to phenotype map by accounting for genetic covariation between traits and markers due to genetic structure^{24–26}, a practice that can potentially improve the quality of horticultural breeding programs that start from highly variable germplasm collections. Here, we evaluate the assumption that horticultural trait variation segregates independently from genetic structure using *Mangifera indica* in the gene pool of the Australian Mango Breeding Program.

Mango is a major horticultural tree crop worldwide, yet an understanding of the domestication history is still debated. The centre of origin of the genus *Mangifera* is Southeast Asia, but the origin of the species *M. indica* is still under question. Based on the fossil record, Mukherjee²⁷ and Blume²⁸ suggested that mango originated in the Malay Archipelago less than 2.58 million years ago. However, recent molecular taxonomy suggests it evolved within a large area of Northwest Myanmar, Bangladesh and Northeast India²⁹. From this area, human migration and trading led to the dispersion of mangoes to many regions of the world³⁰.

Several studies have evaluated the genetic structure of domesticated mango^{31,32,33,34,35,36,38}. Yet, to our knowledge, there have been no published studies on the effects of genetic structure on phenotypic variation in mango accessions. One study with 60 mango accessions from India accounted for genetic structure in a marker trait analysis³⁵, however, Lal et al.³⁵ did not assess the effect of genetic structure on their genotype to phenotype map. Without understanding the effect of genetic structure on phenotypic diversity, we do not know whether we are creating false associations between genetic markers and key mango traits. Here, we directly examined the effects of genetic structure on the creation of spurious associations between genetic markers and three traits – trunk circumference (a proxy for tree size), fruit blush colour and intensity – in the Australian mango gene pool. We assessed 272 SNP markers genotyped in 208 *M. indica* accessions imported worldwide and revealed statistical associations between genetic markers and traits arising from genetic structure. These results will help guide future studies incorporating genetic structure into their genomic selection models.

Results

Genetic structure in the Australian mango gene pool. Genetic structure was found in both a hierarchical cluster analysis (HCA) and a principal component analysis (PCA) across all 208 *M. indica* accessions (Fig. 1). Consistent with a recent origin of all accessions, the HCA created a dendrogram with only short branches in the centre (Fig. 1a), indicating few genetic differences separate the clusters. The optimal number of genetic clusters was $K = 4$, as indicated by the HCA and the elbow plot. The elbow plot from the HCA shows diminishing returns in the amount of variance explained after five clusters (Fig. S1). In the dendrogram, cluster 1 is the most genetically differentiated cluster, which only contains accessions imported from Southeast Asia. Cluster 1 is most distinct from clusters 2 and 3. In contrast, cluster 4 is more similar to cluster 1 (Fig. 1a) and contains a mixture of samples across geographical regions (e.g., South Asia, Southeast Asia, Americas, and Oceania; Table 1; Fig. 2). In the reduced principal component (PC) space (Fig. 1b), genetic clusters largely overlap, with South Asian accessions (mostly Indian accessions) primarily concentrated in the centre of the multivariate space. Genetic clusters from Southeast Asia, the Americas, and Oceania occur towards the edges of the genotypic space, with Southeast Asia distinctly separated in the PC1 axis.

In agreement with the HCA and PCA results above, we identified genetic clusters across the 208 *M. indica* accessions (Fig. 3) using the Bayesian clustering approach implemented in STRUCTURE³⁹. Most accessions contained large amounts of admixture or shared ancestral polymorphism, where portions of their genome were assigned to different genetic groups. When genetic differentiation was separated into only two groups ($K = 2$, see Methods), Southeast Asia formed one group, while all other accessions were in a second group (Fig. 3). Relaxing this constraint to $K = 3$ revealed the Americas and Oceania accessions each form a group. Populations are almost indistinguishable when K is larger than 4. Consistent with the elbow plot discussed above, the Evanno method⁴⁰ and the log probability of K values show that $K = 4$ was the optimal number of clusters (Fig. S2). Most accessions show signatures of admixture as indicated by diversity from multiple groups. Admixture signals are particularly pronounced in accessions from South Asia, mainly those from India, which do not form a distinct genetic group with any K -value.

Together, the HCA, PCA and STRUCTURE results suggest that mango accessions of the Australian mango gene pool consist of four genetic groups. Southeast Asian accessions are most differentiated relative to the rest of the world, suggesting that these accessions might have evolved differently, thus creating a heterogenous gene pool for cultivar creation in the Australian Mango Breeding Program.

Patterns of genetic diversity across the Australian mango gene pool. Genetic diversity analyses revealed high levels of heterozygosity and variable patterns of inbreeding across regions (Table 2). Levels of

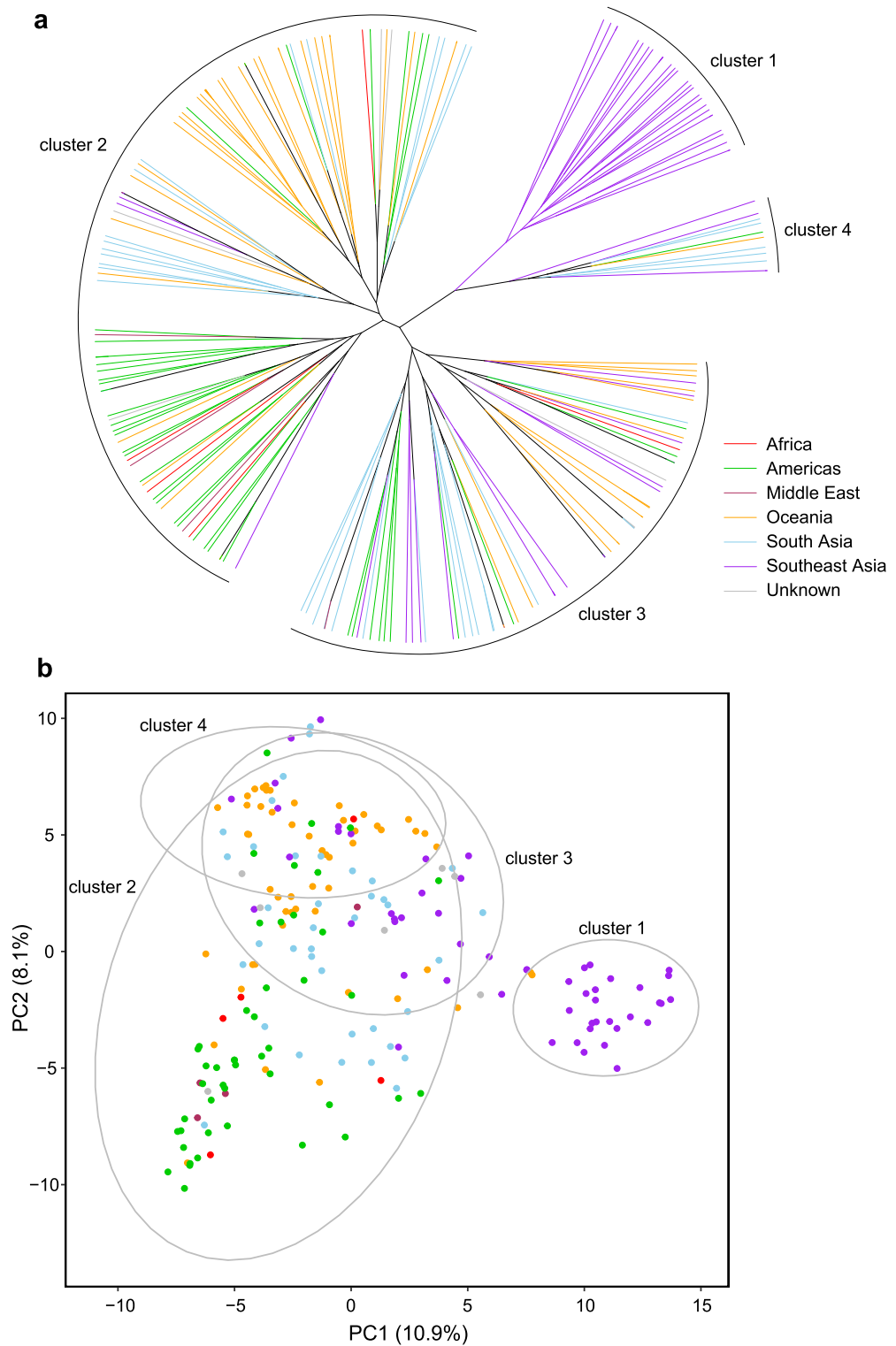


Figure 1. Genetic structure analyses for $K=4$ of the 208 accessions of *M. indica* from six geographical regions across the world. **(a)** A circular dendrogram showing the hierarchical cluster analysis using complete linkage clustering. Each branch represents an individual with the colour of the branch representing the geographical region the sample was imported into Australia from. **(b)** Principal components analysis, where the ellipses (95% probability) represent the four clusters from the hierarchical cluster analysis.

Country	Country code	cluster 1	cluster 2	cluster 3	cluster 4	Country total
Africa						
East Africa	EAF			1		1
Kenya	KEN		1			1
South Africa	ZAF		3			3
Americas						
Brazil	BRA		1	1		2
Jamaica	JAM		1	1	1	3
Saint Lucia	LCA			1		1
United States of America	USA		32	8		40
Middle East						
Israel	ISR		3	1		4
Oceania						
Australia	AUS		37	15		52
French Polynesia	PYF			1	1	2
South Asia						
India	IND		14	15	5	34
Pakistan	PAK		1			1
Sri Lanka	LKA		1	2		3
Southeast Asia						
Indonesia	IDN		2	6	2	10
Malaysia	MYS	1		2	1	4
Malesia	MLS		1	2	1	4
Myanmar	MMR	1				1
Philippines	PHL	3				3
Singapore	SGP			1		1
Thailand	THA	18	1	3		22
Vietnam	VNM	5		4		9
unknown			4	3		7
Cluster total		28	102	67	11	208

Table 1. The number of accessions of *M. indica* from each country of import and their assigned genetic clusters from the hierarchical cluster analysis for $K = 4$ calculated from 272 biallelic SNPs. Countries have been grouped into six geographical regions of import.

expected heterozygosity (H_E) and observed heterozygosity (H_O) were high across the world, with the Americas having the highest levels of observed heterozygosity ($H_O = 0.49$) and Southeast Asia having the lowest ($H_O = 0.39$). Accessions from the Americas contain an excess of heterozygote individuals (i.e., a negative inbreeding coefficient; $F_{IS} = -0.11$; 95% CI -0.13 to -0.08). On the other hand, accessions from Southeast Asia are mildly inbred (i.e., a positive inbreeding coefficient; $F_{IS} = 0.08$; 95% CI 0.06 to 0.11). Private alleles were absent in all regions, indicating either a large intermixing population or the presence of ancestral polymorphisms that have not been sorted across geography.

Genetic differentiation comparisons showed variable patterns of F_{ST} between genetic clusters and between regions of import. Comparisons between regions have low levels of F_{ST} which range from -0.016 to 0.112 (Table 3a). Southeast Asia and the Middle East, closely followed by the comparison between Southeast Asia and the Americas, showed the highest level of genetic differentiation ($F_{ST} = 0.112$ and 0.107 , respectively). In contrast, F_{ST} between clusters ranged from 0.051 to 0.286 , with cluster 1 comparisons having the highest values (Table 3b). Overall, there is low genetic divergence amongst regions of the Australian mango gene pool and high genetic divergence between genetic clusters.

Genetic structure and region of import influence phenotypic diversity. Phenotypic correlation analyses revealed associations between fruit blush colour and intensity but not between them and trunk circumference. Trunk circumference, a continuous trait, was highly variable at 9 years, ranging from 27 to 70 cm, while categorical fruit traits were less variable (see Fig. S3 for photos of each fruit blush colour and intensity category). In a single-factor linear model, fruit blush colour and intensity were strongly correlated (LR $\chi^2 = 373.168$, $df = 4$, $p < 0.0001$, $R^2 = 0.61$). However, given that 39% of mango accessions lacked fruit blush colour and therefore lacked fruit blush intensity, we removed 'no blush' and retested the association. It led to a significant yet weaker association between the fruit traits (LR $\chi^2 = 95.077$, $df = 3$, $p < 0.0001$, $R^2 = 0.28$), indicating the importance of no blush in our understanding of the genetics of blush in mango. We found no correlation between trunk circumference and fruit blush colour (Fig. S4; $F_{4,203} = 1.093$, $p = 0.3613$, $R^2 = 0.02$) and trunk circumference and fruit

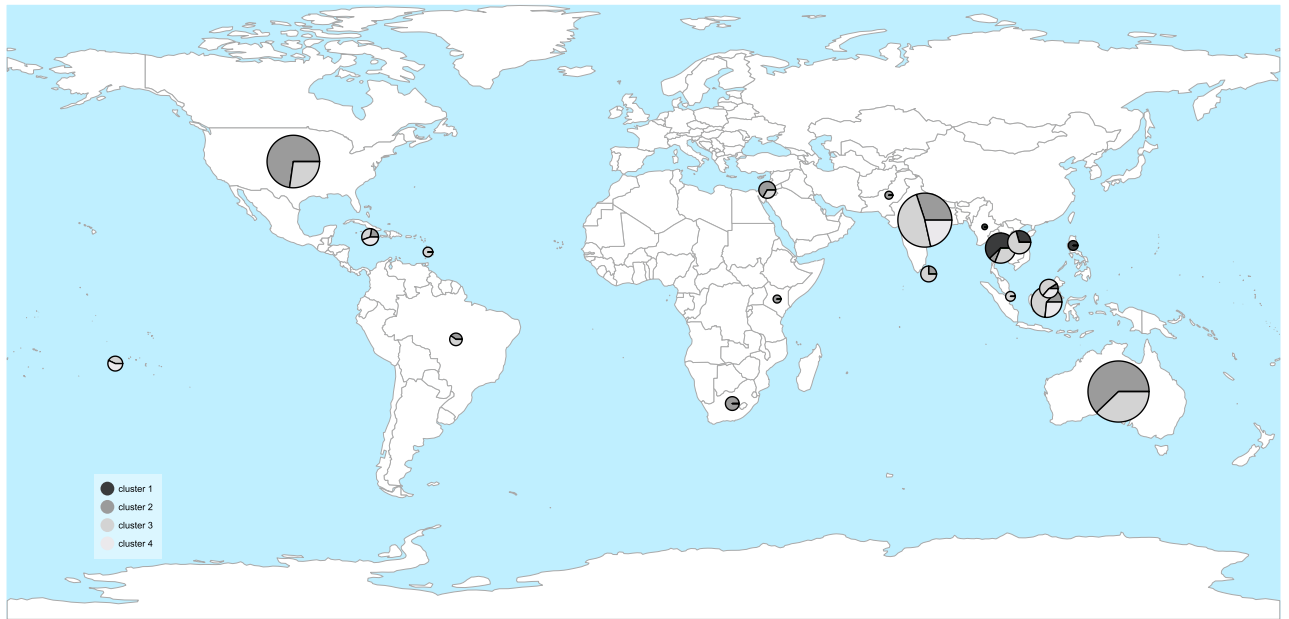


Figure 2. Genetic structure across geography of the 208 *M. indica* accessions. Cluster numbers (K=4) were determined using a hierarchical cluster analysis (Fig. 1). The size of each pie chart reflects the number of accessions imported from each country. The world map was created in “rworldmap” v1.3–6 R-package (<https://cran.r-project.org/web/packages/rworldmap/>).

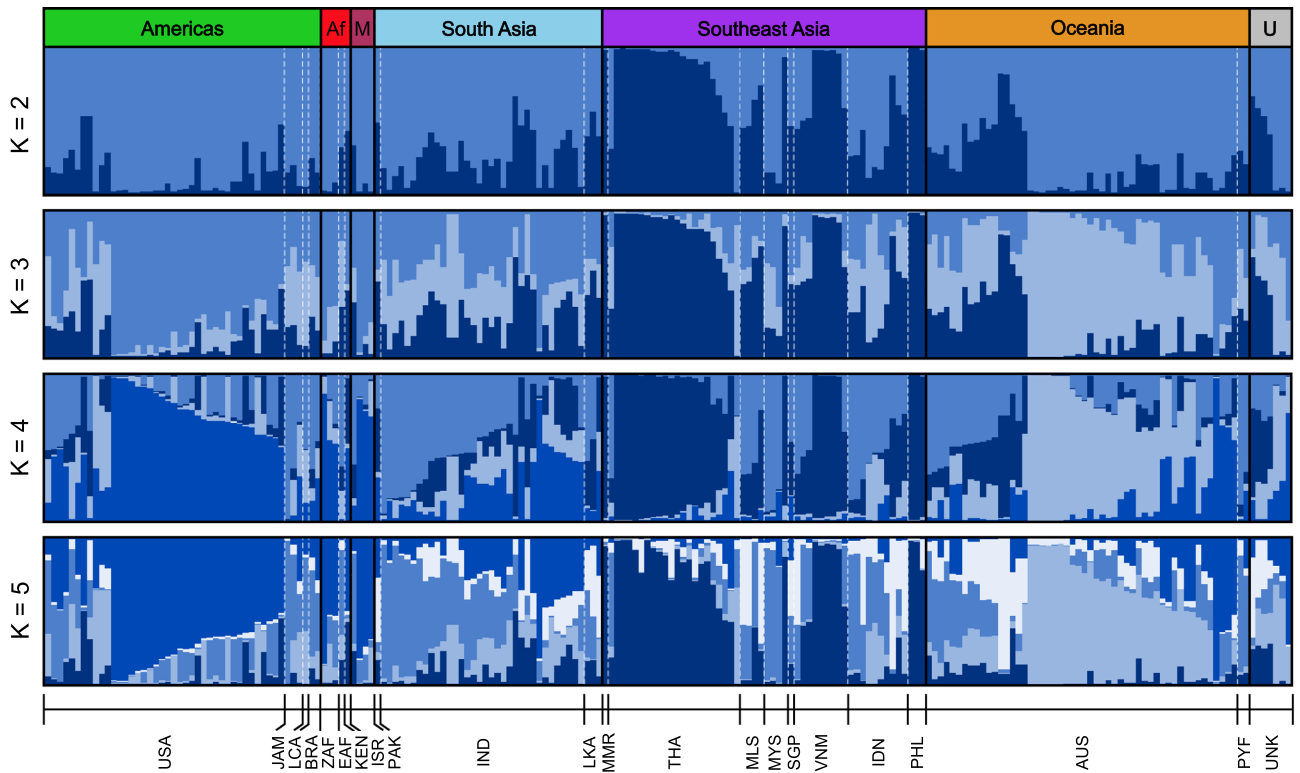


Figure 3. Genetic structure of 208 *M. indica* individuals using STRUCTURE for K=2 to K=5. Each bar represents an individual with the shades of blue representing the ancestry proportions to each cluster. Individuals are sorted by geographical region (black lines), where Af = Africa, M = Middle East and U = unknown, and country (white dotted lines). Refer to Table 1 for information on each country code.

Region	H _O	H _E	F _{IS} (95% CI's)	Pr
Africa	0.46	0.45	-0.02 (-0.08 to 0.05)	0
Americas	0.49	0.44	-0.11* (-0.13 to -0.08)	0
Middle East	0.46	0.45	-0.01 (-0.08 to 0.05)	0
Oceania	0.43	0.42	-0.02 (-0.05 to 0.00)	0
South Asia	0.45	0.45	0.00 (-0.03 to 0.02)	0
Southeast Asia	0.39	0.42	0.08* (0.06 to 0.11)	0

Table 2. Genetic diversity for 208 *M. indica* accessions across six geographic regions using 272 SNPs. H_O, observed heterozygosity; H_E, expected heterozygosity; F_{IS}, inbreeding co-efficient with 95% confidence intervals, where *CI's do not overlap with 0; Pr, the number of private alleles.

Region	Africa	Americas	Middle East	Oceania	South Asia	Southeast Asia
a						
Africa	-	-0.012 to 0.004	-0.036 to 0.005	0.022 to 0.055	0.010 to 0.035	0.062 to 0.095
Americas	-0.004	-	-0.008 to 0.012	0.051 to 0.070	0.036 to 0.053	0.092 to 0.121
Middle East	-0.016	0.002	-	0.059 to 0.099	0.009 to 0.036	0.088 to 0.134
Oceania	0.039*	0.060*	0.079*	-	0.031 to 0.046	0.070 to 0.095
South Asia	0.022*	0.044*	0.022*	0.038*	-	0.049 to 0.069
Southeast Asia	0.078*	0.107*	0.112*	0.082*	0.060*	-
Cluster	1	2	3	4		
b						
1	-	0.151 to 0.191	0.135 to 0.174	0.250 to 0.322		
2	0.171*	-	0.043 to 0.059	0.127 to 0.170		
3	0.153*	0.051*	-	0.050 to 0.074		
4	0.286*	0.148*	0.062*	-		

Table 3. Pairwise F_{ST} for *M. indica*. a) F_{ST} estimates for the geographical regions and b) clusters. F_{ST} estimates are below the diagonal, 95% confidence interval above the diagonal are based on 1000 bootstrap replicates. Clusters (K = 4) were identified using hierarchical cluster analysis (Fig. 1). *95% confidence intervals do not overlap with 0.

blush intensity (Fig. S5; F_{4,203} = 1.473, p = 0.2118, R² = 0.03), suggesting trunk circumference is likely to be genetically independent of these fruit traits.

Fruit blush traits are strongly associated with the region of import in the Australian mango gene pool. In single trait linear models, region of import showed a significant effect on fruit blush colour (Fig. 4a; LR χ^2 = 77.768, df = 12, p < 0.0001, R² = 0.14) and fruit blush intensity (Fig. 4b; LR χ^2 = 98.936, df = 3, p < 0.0001, R² = 0.18), but not trunk circumference (F_{3,188} = 1.970, p = 0.1200, R² = 0.03). Of the regions that had more than ten samples, trunk circumference ranged from a mean of 48.1 ± 1.8 (n = 38) in South Asia to a mean of 52.5 ± 1.3 (n = 46) in the Americas (Table S1). For fruit blush colour (Table S2), 67% of accessions from Southeast Asia had no blush colour, while only 11% from the Americas had no blush, with most having red blush (43%). For fruit blush intensity (Table S3), the Americas had 41% of accessions with a medium blush intensity that resembled the Haden accession. In comparison, Oceania had 39% of accessions with slight blush intensity resembling the Kensington Pride accession. Contrastingly, 94% of Southeast Asian accessions and 82% of South Asian accessions had no blush or barely visible blush intensity.

Fruit blush colour, intensity and trunk circumference were all associated with the four clusters assigned in the HCA. Cluster assignment had a significant effect on fruit blush colour (LR χ^2 = 47.074, df = 12, p < 0.0001, R² = 0.08) and the presence of blush (LR χ^2 = 28.046, df = 3, p < 0.0001, R² = 0.10), where 18% of individuals in cluster 1 had blush, whereas 70% and 69% of individuals from clusters 2 and 3 had blush, respectively. Cluster 1 is more likely to have lower blush intensity than the other clusters when the 'no blush' category is excluded (LR χ^2 = 12.274, df = 3, p = 0.0065, R² = 0.04; odds ratios between cluster 1 and clusters 2 to 4 ranged from 3.8 to 10.5). Finally, cluster had a significant effect on trunk circumference (F_{3,204} = 18.410, p < 0.0001, R² = 0.21), where cluster 1 (mean = 52.3 ± 1.5, n = 28) and cluster 2 (mean = 53.7 ± 0.8, n = 102) had the largest trunk circumference and cluster 4 had the smallest (mean = 36.8 ± 2.9, n = 11). Overall, we expect that genetic diversity and factors specific to the region of import will likely influence the genotype to phenotype map of these key mango traits.

Discussion

Genetic structure arises from evolutionary processes such as mutation, migration and genetic drift, which drive shifts in allelic frequency that could cause statistical associations between random genetic markers and traits⁴¹. Such variation arising from genetic structure is often confounded with loci contributing to trait variation in

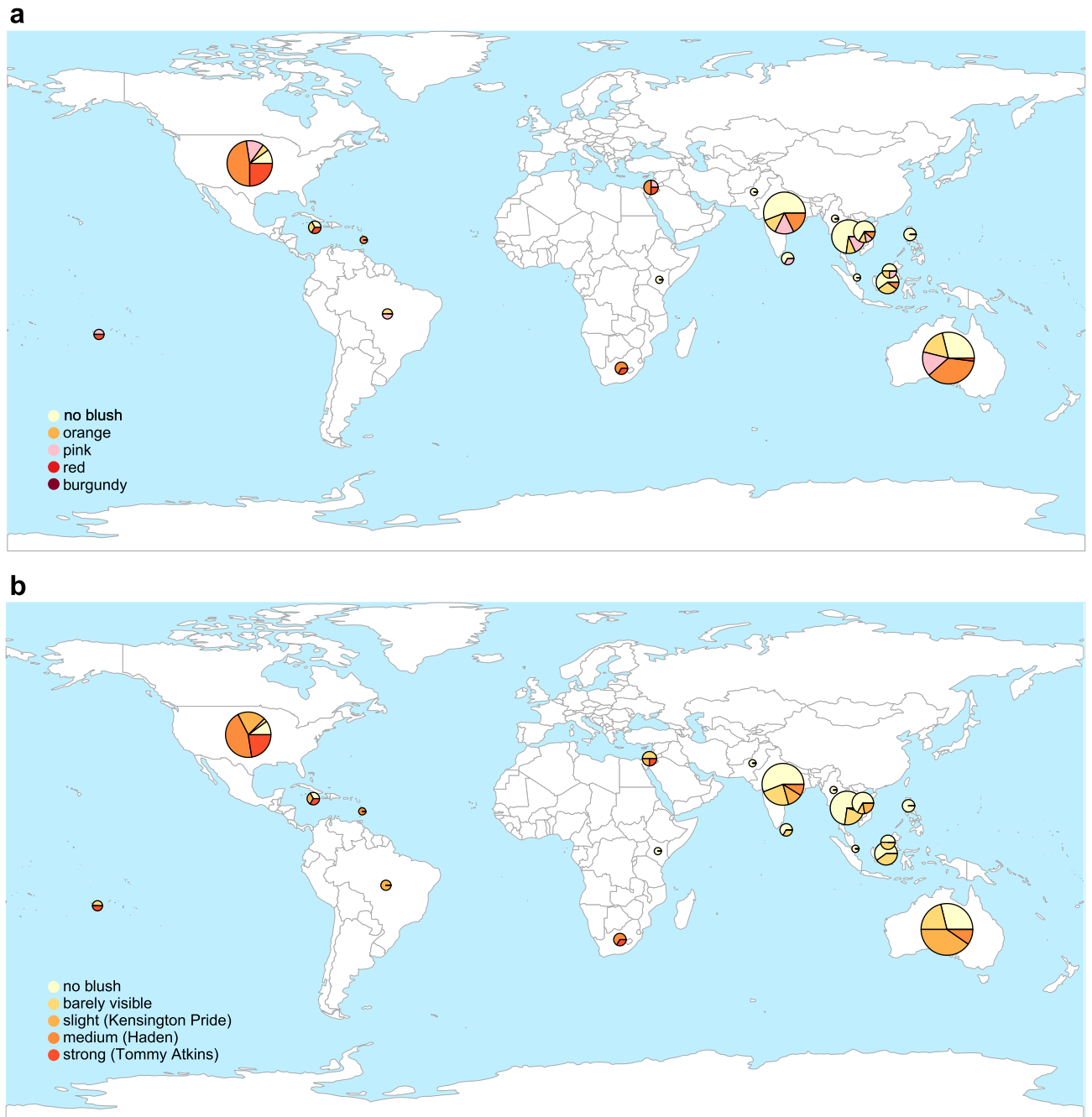


Figure 4. Fruit blush colour and intensity across geography of the 208 *M. indica* accessions. **(a)** Fruit blush colour is split into five categories. **(b)** Fruit blush intensity increases from no blush to strong blush on an ordinal scale, where the accessions in brackets best reflect the colour intensity. The size of each pie chart reflects the number of accessions imported from each country. The world map was created in “rworldmap” v1.3-6 R-package (<https://cran.r-project.org/web/packages/rworldmap/>).

association studies^{17,18,19,20,21,23}, which can misrepresent the genotype to phenotype map assumed in genomic selection models. Our study shows how genetic structure in *M. indica* can lead to statistical associations between genetic markers and three phenotypic traits measured in this study – trunk circumference, fruit blush colour and intensity. This suggests that the genetic architecture of these horticultural traits contains noise arising from the conflation of phenotypic and historical differences in the Australian mango gene pool. Such noise can create spurious associations that hinder the selection of new cultivars, so we recommend that future studies in mango breeding take this into consideration.

Genetic variability and divergence in the Australian mango gene pool can be understood in two ways. On the one hand, accessions imported from different regions are weakly differentiated. On the other hand, genetic clusters are strongly differentiated, implying the existence of clear genetic groups. Results described in Fig. 2 reveal that genetic clusters are distributed across regions, implying that their genetic structure is shared across the

world. The net effect of this nested relationship between geographic region and genetic cluster is low F_{ST} values amongst the regions yet high levels of F_{ST} amongst genetic clusters. This relationship can be used to hypothesise the causes of genetic divergence in the Australian mango gene pool.

In our study, cluster 1 (containing only Southeast Asian countries) comprises the most genetically differentiated accessions from across the world. Previous studies support this observation; Warschafsky and von Wettberg³¹ showed that accessions from Southeast Asia cluster together in a STRUCTURE plot, and Dillon et al.³² reached a similar conclusion using genetic distance analyses of 254 mango accessions. Surprisingly, we did not find private alleles (exclusive alleles) to Southeast Asia, such as those found in Warschafsky and von Wettberg³¹ (74 private alleles from a total of 364 SNPs; 20%). It is unclear what is driving this difference in the number of private alleles between the two studies. However, some of the factors that could be contributing to the variation in sampled loci include different cultivars, contrasting sequencing techniques (SNP chip vs. Restriction site associated DNA markers), different approaches for calling variants, and for filtering of the minor allele frequency^{42,43}. The genetic differentiation observed between Southeast Asia and the rest of the world might have been driven by regional cultural differences. For example, in Southeast Asia, mangoes are incorporated into savoury dishes, which might have led to the selection of immature mangoes that stay green while ripening and therefore lack blush³¹. On the other hand, red blush is favoured around the world⁴⁴, likely accentuating genetic differentiation between accessions from Southeast Asia and the rest of the world.

Artificial selection for these cultural preferences may have driven some of the genetic differentiation identified in the Australian mango gene pool. It is well accepted that selecting one trait can incidentally lead to the evolution of other traits through genetic linkage^{45,46}. The genetic architecture of selected traits will largely determine the extent of this correlated evolution. In this study, we show that fruit blush colour and intensity are highly correlated, which might imply a shared genetic architecture. Therefore, selection for either of these traits could partially drive the evolution of the other. For example, the evolution of low blush intensity, but not trunk circumference, might have arisen from selection of low levels of blush colour in Southeast Asia. Selection of polygenic traits and recruitment of pleiotropic genes can also affect levels of genetic differentiation across the genome. Selection for trunk circumference, which is a polygenic trait⁴⁷, might therefore drive changes in allelic frequencies across many loci. In contrast, fruit colour pigments and their levels, are often controlled by fewer loci in simpler biochemical pathways^{48–50}. In general, we expect genes controlling plant growth and development^{51–53} to be important drivers of genetic differentiation between accessions and merit further attention considering the influence of the genetic architecture of selected traits on population structure.

Polyembryony could have contributed to the origin of genetic differences between Southeast Asia and other accessions. Southeast Asian accessions are typically polyembryonic, where all but one (the zygotic embryo) of the multiple somatic embryos are genetically identical to the maternal parent. Polyembryony is likely to easily be maintained under moderate to strong selection as it is thought to be inherited through a single dominant gene^{54,55}. A high level of polyembryony can freeze the genetic diversity in a population, as instead of allowing hybridisation and creating unique individuals through recombination, it propagates genetically identical individuals⁵⁶. Polyembryony can therefore create genetic bottlenecks if only a fraction of the original genetic diversity is propagated, consistent with the signature of inbreeding in Southeast Asian accessions we found in this study. Furthermore, previous studies have found genetic clustering of mango accessions according to their ability to produce polyembryonic seed^{57,58}. However, embryo type is conflated with geographic region in these studies, where Southeast Asian accessions dominate the polyembryony types. Therefore, without future work teasing apart the contribution of polyembryony and geographic region than we lack an understanding of the various causes of polyembryonic selection and inbreeding on the genetic diversity of tree crops.

Genetic diversity and partitioning of genetic structure influence prediction accuracy in genomic selection models across horticultural crops^{16,24,26,59–61}. For instance, increasing genetic diversity by using a variety of races or genetic clusters in the training and validation sets produced higher prediction accuracies in rice, sorghum¹⁶ and wheat⁶¹. But genetic diversity is known to reduce prediction accuracy when estimation error is high, which occurs in small populations or when there is low marker density^{61,62}. By definition, using markers close to the causative variants will augment prediction accuracy during breeding; however, this is hard to achieve with low-density genotyping techniques such as SNP chips and Genotyping by Sequencing. With sequencing that covers the entire genome (e.g., whole genome sequencing), factors influenced by linkage disequilibrium can be better controlled, such as finding markers in tight linkage with causative loci. As such, population size, marker density, the genetic structure of the population, and the genetic architecture of the chosen traits will play a significant role in the accuracy of genomic selection models.

To ameliorate the adverse effects of genetic structure in genomic selection models, there are two major approaches used across horticultural crops^{16,24,26,59,61}. The first approach includes principal components from genetic structure analyses as covariates in the model^{63–66}. However, this method can double-count genetic structure because some elements are included in the model through the genomic relationship matrix⁶⁷. Another common approach for accounting for genetic structure in genomic selection models is ensuring an equal contribution across genetic clusters in training and validation sets. This stratified sampling approach has been shown to increase prediction accuracy in sorghum¹⁶ and maize, and could be an effective method in the Australian mango gene pool. In general, choosing the most accurate genomic selection model will largely depend on the breeding population's genetic structure and the number of samples.

Conclusion

The results of this study reveal that a horticultural species spread across the world has a genetic structure that can create statistical associations between three key traits and genetic markers. To remove the effects of spurious markers, breeders should fully characterise the genetic structure of their breeding population. This will allow

them to incorporate sample stratification to improve the performance of genomic selection models. Together with best practices of genomic selection (e.g., whole genome sequencing and large population size), these considerations can improve the genotype to phenotype map to assist in choosing individuals with accurate breeding values and help advance future parental selection. We hope our study encourages other horticultural breeding programs to follow similar methods.

Methods

Ethics statement. All plant material used in this research was sourced and collected from the Walkamin Research Station, Queensland (17.1341°S, 145.4271°E), where trees are held as a living collection. The Department of Agriculture and Fisheries granted permission as stated in the National Tree Genomics Program – Phenotype Prediction project (AS17000) for use and collection of materials from mango trees from their government station. This study complies with relevant institutional, national, and international guidelines and legislation.

Accessions. A total of 208 *M. indica* accessions were used from the gene pool collection of the Australian Mango Breeding Program at Walkamin Research Station. These accessions were imported from 21 countries across six geographical regions and were grafted onto the uniform polyembryonic rootstock, Kensington Pride. See Table 1 for the complete list of countries and sample sizes.

Genotyping. To identify some of the genotypic diversity in the Australian mango gene pool, we used the genotypes from Kuhn et al.⁶⁸. DNA isolation for these genotypes was described in Kuhn et al.⁶⁹. Briefly, young leaf samples were collected from Walkamin Research Station and the glasshouse at Mareeba Research Facility, Queensland (17.0075°S, 145.4295°E). DNA was extracted using 20 mg of fresh sample with the Qiagen Plant DNeasy kit. SNP genotyping was performed on these DNA samples using the Fluidigm EP-1 platform with 384 biallelic SNP markers. Finally, 272 SNP markers were selected for further analyses, where 236 markers belong to one of 20 linkage groups (7–20 markers per linkage group), and the location of the remaining 36 markers in the genome is unknown⁶⁸. Genotypically identical individuals across the 272 SNPs were consolidated, leaving 208 mango accessions for the analyses. On average, 98% of the 272 SNPs used in this study were successfully genotyped in every accession.

Hierarchical cluster analysis. To examine the genotypic clustering of the mango accessions due to genotypic similarity, we performed a hierarchical cluster analysis (HCA) of the 208 *M. indica* accessions. First, pairwise genetic distances between all accessions were calculated using the percentage method by the “ape” v5.3 R-package⁷⁰. The HCA was conducted by “stats” v3.6.2 R-package with complete linkage clustering. This computes all pairwise dissimilarities between the accessions in a cluster and accessions in another cluster and considers the largest value of these dissimilarities as a distance between the two clusters. To assess the optimal number of clusters, we used the elbow method⁷¹, which plots the total within-cluster sum of squares (WSS) against the number of clusters to show the ‘elbow’ where the WSS rate of decrease slows and indicates diminishing returns with more clusters⁷².

Principal components analysis. We assessed the major patterns of genetic similarity among the 208 mango accessions in multivariate space using a principal components analysis (PCA) with 272 SNPs. Missing SNP data were imputed using the regularised iterative PCA algorithm with the “missMDA” v1.17 R-package⁷³. The PCA was performed using the “stats” v3.6.2 R-package⁷⁴. Ellipses were constructed for each of the four clusters in the HCA to identify the position of every individual in a cluster in multivariate space with 95% probability.

Structure analysis. We determined levels of admixture between all 208 *M. indica* accessions with STRUCTURE v2.3.4³⁹. STRUCTURE is a Bayesian Markov chain Monte Carlo (MCMC) program that assigns individuals into genetic clusters (K) based on their genotypes by assuming Hardy Weinberg equilibrium within a cluster. It gives each accession an admixture coefficient to depict the proportion of the genome originating from a particular K cluster. We ran the admixture model and the correlated allele frequency model⁷⁵ with ten independent runs of 100,000 burn-in and 100,000 MCMC iterations for K=1 to K=7. We visually inspected summary statistics of MCMC runs to ensure convergence of model parameters. Results were summarised and plotted in the “pophelper” v2.2.7 R-package⁷⁶. The optimal K value (which represents the most likely number of sub-populations) was estimated by the Evanno method⁴⁰, which uses the second-order rate of change in the log probability of data between successive K values in the R-package StructureSelector⁷⁷. The optimal K value was also estimated using LnP(K), the mean log probability of the data. We also followed suggestions by Pritchard et al.⁷⁸ and Lawson, et al.⁷⁹ and plotted the lowest K values that capture the primary structure in the data.

Genetic diversity and genetic differentiation. To examine the level of differentiation between the clusters and geographical regions, Weir and Cockerham’s pairwise F_{ST} and 95% confidence intervals were estimated by “hierfstat” v0.4.22 R-package⁸⁰. Each accession was assigned to a cluster based on the HCA, and each country of import was grouped into six geographic regions. We calculated 95% confidence intervals for each pairwise comparison using 1000 bootstrap replicates. Significance was determined by whether the confidence interval overlapped with 0.

Measures of genetic diversity were calculated for all 208 *M. indica* accessions for each of the six geographic regions. A genind object was created in “adegenet” v2.1.2 R-package^{81,82} for input into “hierfstat” v0.4.22

R-package⁸⁰ to calculate observed heterozygosity (H_o), expected heterozygosity (H_E) and the inbreeding coefficient (F_{IS}). To determine whether F_{IS} was significantly different from 0, we calculated 95% confidence intervals for each pairwise comparison using 1,000 bootstrap replicates. The number of private alleles (Pr) was calculated with the “poppr” v2.8.6 R-package^{83,84}.

Phenotyping. To capture some of the phenotypic diversity in the Australian mango gene pool, we measured three traits in all 208 mango accessions – trunk circumference, fruit blush colour and fruit blush intensity. Trunk circumference was used as a proxy for tree size, as it has been found to be a strong indicator of tree size in other tree crops^{85–87}. Trunk circumference was measured 10 cm above the graft when the trees were 9 years old at Walkamin Research Station. After maturity (> 5 years old), fruit blush colour and intensity were assessed once a year using ten ripe fruits from each mango accession for at least 2 years. Fruits were taken from the outside of the tree, where they are exposed to full sun and have well developed blush. Fruit blush included five categories: no blush, orange, pink, red and burgundy (Fig. S3a). Fruit blush intensity was recorded as five ordinal variables increasing in colour intensity (Fig. S3b), where the accessions in brackets best reflect the colour intensity: no blush, barely visible, slight (Kensington Pride), medium (Haden) and strong (Tommy Atkins).

The effect of region of import and genetic structure on phenotypic diversity. Tests of association were undertaken to examine the relationship between traits. Chi-square likelihood ratios were used to test phenotypic association amongst the categorical traits of fruit blush and intensity. We then performed the same analysis with the ‘no blush’ category removed to test whether the association remains. A linear model was performed to test for an association between trunk circumference and fruit blush colour, and also trunk circumference and fruit blush intensity.

To understand the effect of region of import on both genotype and phenotype in the Australian mango gene pool, we tested its association with genetic structure and phenotypic diversity. We investigated the influence of geographic region on phenotypic diversity for three key mango phenotypes – trunk circumference, fruit blush colour and intensity. We performed a likelihood-ratio chi-square test for fruit blush colour (categorical) and intensity (ordinal) against the region of import and a linear model for trunk circumference. Region of import was the explanatory variable in each model and included the regions shown in Table 1, excluding unknown regions ($n=7$) and regions with low samples sizes, including the Middle East ($n=4$), and Africa ($n=5$).

We then tested for an effect of genetic structure on the three phenotypes using the optimal cluster assignment of $K=4$ from the HCA. Likelihood-ratio chi-square tests were performed for whether cluster explained (1) fruit blush colour, and (2) the presence ($n=127$) vs absence of blush ($n=81$), irrespective of the intensity of blush. We then removed the individuals with no blush from the dataset to test whether there was a significant difference in fruit blush intensity between clusters for just the individuals with fruit blush using a likelihood-ratio chi-square test with an odds ratio. Finally, we performed a mixed linear model to test the effect of cluster on trunk circumference. JMP v15.2.0 (SAS 2015) produced all statistical results reported here.

Data availability

All data generated or analysed during this study are included in this published article (and its supplementary information files).

Received: 5 September 2022; Accepted: 21 November 2022

Published online: 30 November 2022

References

- Smith, J. R. *Tree Crops, A Permanent Agriculture*. (Lulu.com, 2015).
- Molnar, T. J., Kahn, P. C., Ford, T. M., Funk, C. J. & Funk, C. R. Tree crops, a permanent agriculture: concepts from the past for a sustainable future. *Resources* **2**, 457–488. <https://doi.org/10.3390/resources2040457> (2013).
- Mizani, A. *et al.* in *XI International Symposium on Integrating Canopy, Rootstock and Environmental Physiology in Orchard Systems 1228.1228* edn 167–174 (International Society for Horticultural Science (ISHS), Leuven, Belgium).
- Bally, I. S. *Mangifera indica* (mango). *Traditional Trees of Pacific Islands. Their Culture, Environment, and Use*, 441–464 (2006).
- Bally, I. S. E., Ping, L. & Johnson, P. R. Mango Breeding. In *Breeding Plantation Tree Crops: Tropical Species* (eds Mohan Jain, S. & Priyadarshan, P. M.) 51–82 (Springer New York, New York, NY, 2009). https://doi.org/10.1007/978-0-387-71201-7_2.
- Bally, I. S. E. & Dillon, N. L. Mango (*Mangifera indica* L.) Breeding. In *Advances in Plant Breeding Strategies: Fruits: Volume 3* (eds Al-Khayri, J. M. *et al.*) 811–896 (Springer International Publishing, Cham, 2018). https://doi.org/10.1007/978-3-319-91944-7_20.
- Reddy, Y., Kurian, R. M., Ramachander, P., Singh, G. & Kohli, R. Long-term effects of rootstocks on growth and fruit yielding patterns of ‘Alphonso’ mango (*Mangifera indica* L.). *Sci. Horticult.* **97**, 95–108. [https://doi.org/10.1016/S0304-4238\(02\)00025-0](https://doi.org/10.1016/S0304-4238(02)00025-0) (2003).
- Topp, B. L., Nock, C. J., Hardner, C. M., Alam, M. & O’Connor, K. M. Macadamia (*Macadamia* spp.) Breeding. In *Advances in Plant Breeding Strategies: Nut and Beverage Crops: Volume 4* (eds Al-Khayri, J. M. *et al.*) 221–251 (Springer International Publishing, Cham, 2019). https://doi.org/10.1007/978-3-030-23112-5_7.
- Kumar, S. *et al.* Genomic selection for fruit quality traits in apple (*Malus domestica* Borkh.). *PLoS One* **7**, e36674. <https://doi.org/10.1371/journal.pone.0036674> (2012).
- Piaskowski, J. *et al.* Genomic heritability estimates in sweet cherry reveal non-additive genetic variance is relevant for industry-prioritized traits. *BMC Genet.* **19**, 23. <https://doi.org/10.1186/s12863-018-0609-8> (2018).
- Gezan, S. A., Osorio, L. F., Verma, S. & Whitaker, V. M. An experimental validation of genomic selection in octoploid strawberry. *Horticult. Res.* **4**, 16070. <https://doi.org/10.1038/hortres.2016.70> (2017).
- Desta, Z. A. & Ortiz, R. Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* **19**, 592–601. <https://doi.org/10.1016/j.tplants.2014.05.006> (2014).
- Heffner, E. L., Sorrells, M. E. & Jannink, J.-L. Genomic selection for crop improvement. *Crop Sci.* **49**, 1–12. <https://doi.org/10.2135/cropsci2008.08.0512> (2009).

14. Kijas, J. W. *et al.* Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. *Animal Genet.* **45**(5), 754–757. <https://doi.org/10.1111/age.12197> (2014).
15. Li, H. & Ralph, P. Local PCA shows how the effect of population structure differs along the genome. *Genetics* **211**, 289. <https://doi.org/10.1534/genetics.118.301747> (2019).
16. Sapkota, S. *et al.* Impact of sorghum racial structure and diversity on genomic prediction of grain yield components. *Crop Sci.* **60**, 132–148. <https://doi.org/10.1002/csc2.20060> (2020).
17. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725. <https://doi.org/10.7554/eLife.39725> (2019).
18. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702. <https://doi.org/10.7554/eLife.39702> (2019).
19. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *The Lancet* **361**, 598–604. [https://doi.org/10.1016/S0140-6736\(03\)12520-2](https://doi.org/10.1016/S0140-6736(03)12520-2) (2003).
20. Flint-Garcia, S. A. *et al.* Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* **44**, 1054–1064. <https://doi.org/10.1111/j.1365-313X.2005.02591.x> (2005).
21. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3**, e4. <https://doi.org/10.1371/journal.pgen.0030004> (2007).
22. Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* **18**, 41–50. <https://doi.org/10.1038/nrg.2016.132> (2017).
23. Quignon, P. *et al.* Canine population structure: assessment and impact of intra-breed stratification on SNP-based association studies. *PLoS One* **2**, e1324. <https://doi.org/10.1371/journal.pone.0001324> (2007).
24. Guo, Z. *et al.* The impact of population structure on genomic prediction in stratified populations. *Theor. Appl. Genet.* **127**, 749–762. <https://doi.org/10.1007/s00122-013-2255-x> (2014).
25. Teo, Y. Y. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr. Opin. Lipidol.* **19**, 133–143. <https://doi.org/10.1097/MOL.0b013e3282f5dd77> (2008).
26. Werner, C. R. *et al.* How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2020.592977> (2020).
27. Mukherjee, S. K. Origin of mango. *Indian J. Genet. Plant Breed.* **11**, 49–56 (1951).
28. Blume, C. L., Vol. 1 (1850).
29. Bompard, J. M. Taxonomy and systematics. In *The Mango: Botany, Production and Uses* (ed. Litz, R. E.) 19–41 (CABI, Wallingford, 2009). <https://doi.org/10.1079/9781845934897.0019>.
30. Mukherjee, S. & Litz, R. In *The Mango: Botany, Production and Uses* (ed R. E. Litz) (CAB international, 2009).
31. Warschefsky, E. J. & von Wettberg, E. J. B. Population genomic analysis of mango (*Mangifera indica*) suggests a complex history of domestication. *New Phytol.* **222**, 2023–2037. <https://doi.org/10.1111/nph.15731> (2019).
32. Dillon, N. L. *et al.* Genetic diversity of the Australian National Mango Genebank. *Sci. Hortic.* **150**, 213–226. <https://doi.org/10.1016/j.scienta.2012.11.003> (2013).
33. Schnell, R. *et al.* Mango genetic diversity analysis and pedigree inferences for Florida cultivars using microsatellite markers. *HortScience* **41**, 993–993. <https://doi.org/10.21273/JASHS.131.2.214> (2006).
34. Razak, S. A. *et al.* Assessment of diversity and population structure of mango (*Mangifera indica* L.) germplasm based on microsatellite (SSR) markers. *Aust. J. Crop Sci.* **13**, 315 (2019).
35. Lal, S. *et al.* Association analysis for pomological traits in mango (*Mangifera indica* L.) by genic-SSR markers. *Trees* **31**, 1391–1409. <https://doi.org/10.1007/s00468-017-1554-2> (2017).
36. Sherman, A. *et al.* Mango (*Mangifera indica* L.) germplasm diversity based on single nucleotide polymorphisms derived from the transcriptome. *BMC Plant Biol.* **15**, 277. <https://doi.org/10.1186/s12870-015-0663-6> (2015).
37. Surapaneni, M. *et al.* Population structure and genetic analysis of different utility types of mango (*Mangifera indica* L.) germplasm of Andhra Pradesh state of India using microsatellite markers. *Plant Systemat. Evolut.* **299**, 1215–1229. <https://doi.org/10.1007/s00606-013-0790-1> (2013).
38. Hirano, R., Htun Oo, T. & Watanabe, K. N. Myanmar mango landraces reveal genetic uniqueness over common cultivars from Florida, India, and Southeast Asia. *Genome* **53**, 321–330. <https://doi.org/10.1139/g10-005> (2010).
39. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967. <https://doi.org/10.1371/journal.pgen.1002967> (2012).
40. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* **14**, 2611–2620. <https://doi.org/10.1111/j.1365-294X.2005.02553.x> (2005).
41. Song, S., Dey, D. K. & Holsinger, K. E. Differentiation among populations with migration, mutation, and drift: implications for genetic inference. *Evolution* **60**, 1–12. <https://doi.org/10.1554/05-315.1> (2006).
42. Albrechtsen, A., Nielsen, F. C. & Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* **27**, 2534–2547. <https://doi.org/10.1093/molbev/msq148> (2010).
43. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942. <https://doi.org/10.1093/genetics/154.2.931> (2000).
44. Shruti, S. *et al.* Evaluation of newly developed mango (*Mangifera indica*) hybrids for their storage behaviour and peel colour. *Indian J. Agric. Sci.* **81**, 252–255 (2011).
45. Burgess, K. S., Etterson, J. R. & Galloway, L. F. Artificial selection shifts flowering phenology and other correlated traits in an autotetraploid herb. *Heredity* **99**, 641–648. <https://doi.org/10.1038/sj.hdy.6801043> (2007).
46. Roff, D. A. *Evolutionary Quantitative Genetics* (Springer Science & Business Media, Heidelberg, 2012).
47. O'Connor, K. *et al.* Genome-wide association studies for yield component traits in a macadamia breeding population. *BMC Genom.* **21**, 199. <https://doi.org/10.1186/s12864-020-6575-3> (2020).
48. Grotewold, E. The genetics and biochemistry of floral pigments. *Annu. Rev. Plant Biol.* **57**, 761–780. <https://doi.org/10.1146/annurev.arplant.57.032905.105248> (2006).
49. Koes, R., Verweij, W. & Quattrocchio, F. Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. *Trends Plant Sci.* **10**, 236–242. <https://doi.org/10.1016/j.tplants.2005.03.002> (2005).
50. Kayesh, E. *et al.* Fruit skin color and the role of anthocyanin. *Acta Physiol. Plant.* **35**, 2879–2890. <https://doi.org/10.1007/s11738-013-1332-8> (2013).
51. Smeekens, S., Ma, J., Hanson, J. & Rolland, F. Sugar signals and molecular networks controlling plant growth. *Curr. Opin. Plant Biol.* **13**, 273–278. <https://doi.org/10.1016/j.pbi.2009.12.002> (2010).
52. Teale, W. D., Paponov, I. A. & Palme, K. Auxin in action: signalling, transport and the control of plant growth and development. *Nat. Rev. Mol. Cell Biol.* **7**, 847–859. <https://doi.org/10.1038/nrm2020> (2006).
53. Mishra, B. S., Sharma, M. & Laxmi, A. Role of sugar and auxin crosstalk in plant growth and development. *Physiol. Plant.* **174**, e13546. <https://doi.org/10.1111/ppl.13546> (2022).
54. Brettell, R. I. S., Johnson, P. R., Kulkarni, V. J., Müller, W. & Bally, I. S. E. 645 edn 319–326 (International Society for Horticultural Science (ISHS), Leuven, Belgium).
55. Aron, Y., Gazit, S., Czosnek, H. & Degani, C. Polyembryony in mango (*Mangifera indica* L.) is controlled by a single dominant gene. *HortScience* **33**, 1241–1242 (1998).

56. Hollister, J. D. *et al.* Recurrent loss of sex is associated with accumulation of deleterious mutations in oenothera. *Mol. Biol. Evol.* **32**, 896–905. <https://doi.org/10.1093/molbev/msu345> (2014).
57. Abiram, K., Singh, S., Singh, R., Mohapatra, T. & Kumar, A. R. Genetic diversity studies on polyembryonic and monoembryonic mango genotypes using molecular markers. *Indian J. Horticult.* **65**, 258–262 (2008).
58. Shukla, M., Babu, R., Mathur, V. & Srivastava, D. Diverse genetic bases of Indian polyembryonic and monoembryonic mango (*Mangifera indica* L.) cultivars. *Curr. Sci.* **85**, 870–871 (2004).
59. Crossa, J. *et al.* Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* **112**, 48–60. <https://doi.org/10.1038/hdy.2013.16> (2014).
60. Isidro, J. *et al.* Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* **128**, 145–158. <https://doi.org/10.1007/s00122-014-2418-4> (2015).
61. Norman, A., Taylor, J., Edwards, J. & Kuchel, H. Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3 Genes|Genom|Genet* **8**, 2889. <https://doi.org/10.1534/g3.118.200311> (2018).
62. Lund, M. S., van den Berg, I., Ma, P., Brøndum, R. F. & Su, G. Review: how to improve genomic predictions in small dairy cattle populations. *Animal* **10**, 1042–1049. <https://doi.org/10.1017/S1751173115003031> (2016).
63. Bermingham, M. L. *et al.* Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci. Rep.* **5**, 10312. <https://doi.org/10.1038/srep10312> (2015).
64. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909. <https://doi.org/10.1038/ng1847> (2006).
65. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190. <https://doi.org/10.1371/journal.pgen.0020190> (2006).
66. Peloso, G. M., Timofeev, N. & Lunetta, K. L. Principal-component-based population structure adjustment in the North American Rheumatoid Arthritis Consortium data: impact of single-nucleotide polymorphism set and analysis method. *BMC Proc.* **3**, S108. <https://doi.org/10.1186/1753-6561-3-S7-S108> (2009).
67. Janss, L., de los Campos, G., Sheehan, N. & Sorensen, D. Inferences from genomic models in stratified populations. *Genetics* **192**, 693–704. <https://doi.org/10.1534/genetics.112.141143> (2012).
68. Kuhn, D. N. *et al.* Estimation of genetic diversity and relatedness in a mango germplasm collection using SNP markers and a simplified visual analysis method. *Sci. Hortic.* **252**, 156–168. <https://doi.org/10.1016/j.scienta.2019.03.037> (2019).
69. Kuhn, D. N. *et al.* Genetic map of mango: a tool for mango breeding. *Front. Plant Sci.* **8**, 577. <https://doi.org/10.3389/fpls.2017.00577> (2017).
70. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528. <https://doi.org/10.1093/bioinformatics/bty633> (2019).
71. Daru, B. H., Elliott, T. L., Park, D. S. & Davies, T. J. Understanding the processes underpinning patterns of phylogenetic regionalization. *Trends Ecol. Evol.* **32**, 845–860. <https://doi.org/10.1016/j.tree.2017.08.013> (2017).
72. Kassambara, A. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. Vol. 1 (Sthda, 2017).
73. Josse, J. & Husson, F. missMDA: a package for handling missing values in multivariate data analysis. *J. Stat. Softw.* **70**, 31. <https://doi.org/10.18637/jss.v070.i01> (2016).
74. R Core Team. R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* (2019).
75. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587. <https://doi.org/10.1093/genetics/164.4.1567> (2003).
76. Francis, R. M. pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32. <https://doi.org/10.1111/1755-0998.12509> (2017).
77. Li, Y. L. & Liu, J. X. StructureSelector: a web-based software to select and visualize the optimal number of clusters using multiple methods. *Mol. Ecol. Resour.* **18**, 176–177. <https://doi.org/10.1111/1755-0998.12719> (2018).
78. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959. <https://doi.org/10.1093/genetics/155.2.945> (2000).
79. Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* **9**, 3258. <https://doi.org/10.1038/s41467-018-05257-7> (2018).
80. Goudet, J. Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186. <https://doi.org/10.1111/j.1471-8286.2004.00828.x> (2005).
81. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129> (2008).
82. Jombart, T. & Ahmed, I. adegenet 1.3–1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071. <https://doi.org/10.1093/bioinformatics/btr521> (2011).
83. Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281. <https://doi.org/10.7717/peerj.281> (2014).
84. Kamvar, Z. N., Brooks, J. C. & Grünwald, N. J. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front. Genet.* <https://doi.org/10.3389/fgene.2015.00208> (2015).
85. Abd El-Wahab, R. H. *et al.* Anthropogenic effects on population structure of *Acacia tortilis* subsp. raddiana along a gradient of water availability in South Sinai Egypt. *Afr. J. Ecol.* **52**, 308–317. <https://doi.org/10.1111/aje.12121> (2014).
86. Cheng, F. S. & Roose, M. L. Origin and inheritance of dwarfing by the citrus rootstock Poncirus trifoliataflyng dragon. *J. Am. Soc. Hortic. Sci.* **120**, 286–291 (1995).
87. Sax, K. & Gowen, J. W. The place of stocks in the propagation of clonal varieties of apples. *Genetics* **8**, 458 (1923).

Acknowledgements

We thank David Kuhn formerly of Agriculture Research Services, United States Department of Agriculture (ARS-USDA) and Barbara Freeman of ARS-USDA for providing the genotypes. This research was undertaken as part of the National Tree Genomics Program – Phenotype Prediction project (AS17000) which is funded by the Hort Frontiers Advanced Production Systems fund, part of the Hort Frontiers strategic partnership initiative developed by Hort Innovation, with co-investment from The University of Queensland and Queensland Government, and contributions from the Australian Government. The data and its use is by courtesy of the State of Queensland, Australia, through the Department of Agriculture and Fisheries.

Author contributions

M.W., C.H. and D.O. designed the study. I.B., N.D. and A.A. collected and curated the data. M.W., R.Y. and M.J. performed data analysis. M.W., M.J., R.Y. and D.O. wrote the manuscript. D.O. and C.H. secured funding and were mentors. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-24800-7>.

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

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022