

Research Article

Cite this article: González-Orozco CE, Osorio-Guarín JA, Yockteng R (2023). Phylogenetic diversity of cacao (*Theobroma cacao* L.) genotypes in Colombia. *Plant Genetic Resources: Characterization and Utilization* **20**, 203–214. <https://doi.org/10.1017/S1479262123000047>

Received: 11 June 2021
Revised: 2 February 2023
Accepted: 3 February 2023
First published online: 28 February 2023



Keywords:

Amazonia; Andes; cacao germplasm; conservation; evolution; genetic diversity; South America

Author for correspondence:

Roxana Yockteng,
E-mail: ryockteng@agrosavia.co

Phylogenetic diversity of cacao (*Theobroma cacao* L.) genotypes in Colombia

Carlos E. González-Orozco¹ , Jaime A. Osorio-Guarín² 
and Roxana Yockteng^{2,3}

¹Corporación Colombiana de Investigación Agropecuaria – Agrosavia, Centro de Investigación La Libertad, Km 14 vía Puerto López, Villavicencio, Meta, Colombia; ²Corporación Colombiana de Investigación Agropecuaria – Agrosavia, Centro de Investigación Tibaitatá, Km 14 vía a Mosquera, Bogotá, Cundinamarca, Colombia and ³Muséum National d'Histoire Naturelle, UMR-CNRS 7205, Paris 75005, France

Abstract

Theobroma cacao L. (cacao) is an important tropical crop used to produce chocolate. Evolutionary relationships between cultivated and wild cacao genotypes and their genetic diversity are poorly understood. Exploring phylogenetic diversity and spatial patterns of both cultivated and crop wild relatives can improve the knowledge of the evolutionary history of a crop, giving insights into its cultivation, breeding programmes and conservation. This study identifies biodiversity priority areas in Colombia by calculating phylogenetic diversity indices using a set of 87 single nucleotide polymorphism markers. These were sourced from 279 genotypes conserved in the Corporación Colombiana de Investigación Agropecuaria (Agrosavia) germplasm collection. The Caribbean and North Andes areas exhibited the highest phylogenetic diversity and significantly high relative phylogenetic diversity. We propose that those regions where wild cacao occurs should be prioritized as conservation areas. Besides, cacao lineages that have recently diverged and are present in Arauca, Huila and Nariño areas, with significantly low relative phylogenetic diversity, should be prioritized for breeding programmes. The Amazonia genotypes were closer to the root of the phylogenetic tree, suggesting an older origin than those found in the Andes region. Our study highlights the importance of using *T. cacao* germplasm from the Amazonia region as a priority to recover relict diversity in breeding programmes and broaden the gene pool of modern cultivated cacao.

Introduction

Theobroma cacao L. is the most economically important species of the genus *Theobroma*, a member of the Malvaceae family. Cacao fruits provide the raw material for a multibillion-dollar chocolate industry and other cocoa-based products processing pharmaceutical and cosmetic industries (Wickramasuriya and Dunwell, 2018). Cocoa demand is expected to grow at 7.3% from 2019 to 2025 to reach a market value of US\$16.3 billion, due to increasing demand from emerging economies and sustained demand from developed economies (Voora *et al.*, 2022). Most of the world's cocoa production (80–90%) comes from 40 million smallholder farmers (up to five hectares), which is their primary source of revenue; therefore, this crop helps in the alleviation of poverty in cocoa-producing regions (Beg *et al.*, 2017; Benjamin *et al.*, 2018). The demand for fine or flavour chocolate has increased (Fernández-Niño *et al.*, 2021), its market has been growing at a rate of 7–11% per year since 2011 (Vignati and Gómez-García, 2020) and Colombia is one of the known producers of this kind of cocoa (Ballesteros *et al.*, 2016); however, its production is small compared to the primary producer's countries (FAO, 2022).

The increasing human population, the destruction of the Amazonian rainforests, the loss of traditional varieties and climate change have significantly altered the cacao diversity causing its vulnerability to sudden changes in weather and the appearance of new pests and diseases (CacaoNet, 2012; Cilas and Bastide, 2020). Cacao breeders look to improve yield, disease resistance and bean quality traits to support the growing global cacao industry (Bekele and Phillips-Mora, 2019; Rodríguez-Medina *et al.*, 2019). However, this breeding is limited by long-generation cycles, self-incompatibility, pollination inefficiency, challenging abiotic and biotic stress factors, including several major diseases (Bekele and Phillips-Mora, 2019), such as Frosty pod rot caused by *Moniliophthora roreri* and witches' broom caused by *Moniliophthora perniciosa* (Álvarez *et al.*, 2014; Díaz-Valderrama *et al.*, 2020).

Plant breeders rely on crop genetic resources conserved in germplasm collections to incorporate genetic diversity into commercialized cultivars and develop new materials for the sustainable cultivation of cacao (Zhang and Motilal, 2016). Around 24,000 cacao accessions, including wild and improved materials, are conserved in two international gene banks in



© The Author(s), 2023. Published by Cambridge University Press on behalf of NIAB. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Trinidad and Costa Rica and national research institutes such as CEPEC in Brazil (Lopes *et al.*, 2011; Kodoth, 2021). In Colombia, the collection conserved in the Corporación Colombiana de Investigación Agropecuaria (Agrosavia) includes around 600 wild and improved accessions (Rodríguez-Medina *et al.*, 2019). Despite the extensive collections and high genetic diversity conserved, most breeding programmes used a narrow genetic base to improve yield and resistance to pests and diseases (DuVal *et al.*, 2017; Rodríguez-Medina *et al.*, 2019; Ceccarelli *et al.*, 2022; Daymond and Bekele, 2022).

The study of crop wild relatives (CWR) is vital for agriculture and food security because they contain high levels of genetic diversity compared to cultivated crops and, as a result, can adapt to a wide range of habitats and environments, making them valuable for crop improvement (Maxted *et al.*, 2006, 2010, 2012; Heywood *et al.*, 2007; Vincent *et al.*, 2013; Zhang *et al.*, 2017; Majeed *et al.*, 2021). Therefore, plant breeders should use a broader range of genetic resources, particularly CWR, that can provide new sources of agronomic traits to develop materials that respond to present and future challenges of the crop, including climate resilience. Collecting, conserving, characterizing and using CWR genetic resources are crucial to support breeding efforts (Maxted *et al.*, 2010; Ford-Lloyd *et al.*, 2011; Dempewolf *et al.*, 2017; Ceccarelli *et al.*, 2022; Renzi *et al.*, 2022). An effective conservation of cacao plant genetic resources in developing countries will need international and regional efforts. For instance, the Global Cacao Genetic Resources Network (CacaoNet) coordinated by Bioversity International aims to optimize the conservation and use of cacao genetic resources worldwide for the benefit of breeders, researchers and farmers (CacaoNet, 2012).

Understanding the genetic diversity, population structure and genetic pedigree of cacao collections is crucial for conservation and breeding strategies. The molecular markers, simple-sequence repeats (SSR), have been extensively used to screen cacao germplasm (Borrone *et al.*, 2007; Zhang *et al.*, 2009; Aikpokpodion *et al.*, 2010; Irish *et al.*, 2010) since they have been reported as an international standard for cacao DNA fingerprinting (Lanaud *et al.*, 1999; Pugh *et al.*, 2004; Saunders *et al.*, 2004). In 2008, Motamayor *et al.* (2008) showed that the diversity of cacao based on 96 SSRs could be classified into 10 different genetic groups using collections from the Upper Amazon, Lower Amazon, Orinoco, north of South America, Central America and Guyana. In addition, Zhang *et al.* (2012), using 15 SSRs, reported a new genetic group from Bolivia with a unique genetic profile different from the groups reported by Motamayor *et al.* (2008). However, SSR markers have disadvantages, such as expensive and time-consuming processes and data sharing complications due to platform-to-platform variation (Livingstone *et al.*, 2011).

Single nucleotide polymorphism (SNP) is the most common form of DNA sequence variation between alleles and is considered an ideal codominant marker system for assessing genetic diversity in model and non-model species (Batley and Edwards, 2007; Mammadov *et al.*, 2012). SNP markers are advantageous because they are abundant, ubiquitous, amenable to high- and ultra-high-throughput automation and have a low error rate compared with SS (Mammadov *et al.*, 2012). To study the identity of cacao, Livingstone *et al.* (2015) identified 330,000 SNPs using RNA-seq data from 16 diverse cacao cultivars and generated a 6 K SNP array. Other applications of SNP markers in cacao include genetic diversity analysis (Ji *et al.*, 2013; Fang *et al.*, 2014; Cosme *et al.*, 2016; Osorio-Guarín *et al.*, 2017, 2018; Gopaulchan *et al.*, 2019, 2020; Mahabir *et al.*, 2020; Wang *et al.*, 2020), marker-trait

association studies (Romero Navarro *et al.*, 2017; McElroy *et al.*, 2018; Osorio-Guarín *et al.*, 2020; Gutiérrez *et al.*, 2021) and domestication studies (Cornejo *et al.*, 2018).

Central America was considered the first centre of cacao domestication in Mesoamerica about 1900 years ago (Miranda, 1962). However, a recent study showed that cacao was cultivated earlier (5300 years ago) in the north-western part of the Amazonia region, predominantly in southern Ecuador (Zarrillo *et al.*, 2018). In Colombia, the expansion of cacao cultivation occurred in the 17th century in the northeastern region (Guerrero-Rincón *et al.*, 1998). Besides, efforts in the country to conserve cacao diversity have been made since the 1940s to safeguard farmers' livelihoods and preserve food security with genetic material introduced from other countries and some combinations with native cacao (Rodríguez-Medina *et al.*, 2019). Colombia has high diversity of cacao CWR (González-Orozco *et al.*, 2020) and high genetic diversity of cultivated cacao based on random amplified microsatellites (RAM) markers (H_e : 0.28) regionally (Morillo *et al.*, 2014) and countrywide (H_e : 0.314) based on SNP markers (Osorio-Guarín *et al.*, 2017). High phenotypic variability using morpho-agronomic descriptors related to productivity, flower and seed traits has also been reported (Ballesteros *et al.*, 2016; López-Hernández *et al.*, 2021).

In addition to the genetic diversity and phenotypic studies, it is essential to investigate the phylogenetic relationships of current genotypes. The reconstruction of a phylogenetic tree can untangle the relationship between genotypes, and phylogenetic diversity (PD) can improve our understanding of evolutionary events that determine the current diversity within a species (Faith and Baker, 2006; Kapli *et al.*, 2020). The most traditional biodiversity metric, species richness, only considers the number of species. In contrast, PD provides a comparable, evolutionary measure of biodiversity not possible with species counts (Miller *et al.*, 2018). PD is the sum of the branch lengths of a tree that connects all studied species (Faith, 1992). This measure can be applied to any taxon regardless of its origin or rank (Fisher *et al.*, 2007; Mishler, 2021) and is widely used in plant conservation, crop science, biogeography, biodiversity and climate change (González-Orozco *et al.*, 2015; Laity *et al.*, 2015; Nagalingum *et al.*, 2015; Laffan *et al.*, 2016; Thornhill *et al.*, 2016). Using PD as a criterion in conservation planning could reduce the risk of losing entire groups or lineages (Soulebeau *et al.*, 2016).

Colombia is a potential source of unexplored cacao CWR diversity (González-Orozco *et al.*, 2021), making it a research priority for the *in situ* conservation of native genetic resources. Areas of high PD could be a potential source of genetic resources well-adapted and resilient to modern challenges due to the climate change that can be used in breeding programmes (González-Orozco *et al.*, 2021).

Osorio-Guarín *et al.* (2017) explored the phylogenetic relationships among cacao genotypes using SNP markers. However, the authors used this analysis to check if the Colombian germplasm collection represents the diversity of the species. To the best of our knowledge, this study is the first to apply PD to understand better the diversity of Colombian cacao in Agrosavia national germplasm dataset. Our study is pioneering because it combines information on its geographical distribution and evolutionary relationships to contribute to selecting *in situ* priority areas and *ex situ* management strategies of cacao germplasm in Colombia.

Materials and methods

Plant material

A total of 279 wild and cultivated accessions conserved in the Corporación Colombiana de Investigación Agropecuaria

(Agrosavia) germplasm collection were evaluated (online Supplementary Table S1). These accessions are stored *ex situ* at the Agrosavia research centre in Palmira (3°30'41"N 76°19'19"W). The wild accessions (179) were collected from habitats in the following departments: Magdalena, Guajira, Cesar, Norte de Santander, Nariño, Choco and Amazonas. The cultivated accessions (100) came from agricultural areas in the departments of Arauca, Valle del Cauca, Huila, Tolima, Cundinamarca, Antioquia and Santander (Fig. 1, online Supplementary Table S1).

Phylogenetic and genetic diversity analyses

We used an alignment of 87 SNPs (Osorio-Guarín *et al.*, 2017) and their flanking invariant region of 60 bp to avoid branch length bias in the phylogenetic analysis. This subset of SNPs belongs to an original set composed of 1560 candidate SNPs developed from cDNA sequences in cacao tissues, more specifically, flowers, cherelles, pod cortex, shoots, roots, germinated seeds and embryos from an *in vitro* culture (Argout *et al.*, 2008; Allegre *et al.*, 2012). Panel selection was based on an SNP call rate percentage higher than 90%, represented across 10 cacao chromosomes and heterozygosity results (Ji *et al.*, 2013; Fang *et al.*, 2014; Osorio-Guarín *et al.*, 2017). The protocol for SNP genotyping of cacao uses the Fluidigm 96.96 Dynamic Array™ (Fluidigm, San Francisco, CA, USA) (Osorio-Guarín *et al.*, 2017).

The nucleotide sequences of all targeting SNPs were aligned in ClustalX v1.83 (Larkin *et al.*, 2007). After concatenating the data, an evolutionary model was estimated using the option SMS of PhyML software (Lefort *et al.*, 2017). The phylogenetic tree was constructed by computing 1000 bootstrap replicates using the maximum likelihood (ML) method in the PhyML v3.0 program (Guindon *et al.*, 2010) found in a bioinformatics platform (<http://www.atgc-montpellier.fr/phyml/>). Gap sites were treated as missing data. We also conducted a Bayesian analysis using MrBayes v3.2 (Huelsenbeck and Ronquist, 2001) in CIPRES (Miller *et al.*, 2010). Using the metropolis-coupled Markov chain Monte Carlo (MCMC) algorithm, two independent runs of 50 million generations were sampled for one in every 1000 trees. The results of the MrBayes analysis were examined for convergence of parameters using Tracer v1.7 (Rambaut *et al.*, 2018), excluding the initial 10% of MCMCs. Posterior probabilities of clades were obtained from the 50% majority rule consensus of the sampled trees. The accession copoazu_75 from the species *Theobroma grandiflorum* (Wild Ex. Spreng. Schum), commonly known as copoazu, cupuassu or cacao blanco, was used as an outgroup. The ancestral distribution of Colombian cacao genotypes was reconstructed using the parsimony method with the software Mesquite v3.61 (Maddison and Maddison, 2021).

Finally, SNPs were scored as codominant markers with the software GenAlex v6.5 (Peakall and Smouse, 2006, 2012) to perform two further analyses: (1) standard measures of genetic diversity such as the effective number of alleles per locus (N_e), expected heterozygosity (H_e) and observed heterozygosity (H_o) for each department; and (2) genetic differentiation via covariance matrix with data standardization among populations based on G -Statistics (Jost's D_{EST}) (Jost, 2008) visualized through a principal coordinate analysis (PCoA).

Phylogenetic diversity analyses

The spatial analysis tool in the software Biodiverse v1.0 (<http://shawnlaffan.github.io/biodiverse/>) (Laffan *et al.*, 2010) was used

to calculate genotype richness (GR), PD, and relative phylogenetic diversity (RPD) at a spatial resolution of 1 degree, corresponding to 100 × 100 km. We used 279 geolocations of cacao genotypes from different geographical regions of Colombia (online Supplementary Table S2). GR is the number of genotypes in a grid cell referred to as observed GR (Laffan *et al.*, 2016), and PD is an observed measure of diversity, calculated as the total sum of branch lengths (in this case, each branch represents a cacao genotype) in each cell (Faith, 1992). RPD is a measure of PD using a standardization and randomization process to avoid bias caused by the number of taxa in a cell (Mishler *et al.*, 2014). RPD was calculated using the ratio of observed PD and a comparison tree with the same topology and equal branch lengths (Mishler *et al.*, 2014).

In addition, observed PD values were compared to the expected values using a randomization test of 999 iterations completed in the "rand_structured" model in Biodiverse v1.0. A two-tailed test with an α of 0.05 was applied to obtain the significance of the observed values compared to the expected values. The PD randomization test produced the significance test for the observed PD, referred to as significant phylogenetic diversity (SPD). The Colombian map, which displayed the significant values, was generated using R scripts (R Core Team and R development core team, 2008). Codes are available at https://github.com/NunzioKner/biodiverse_pipeline.

Results

Phylogenetic and genetic diversity analyses

The concatenation of the SNPs with their flanking regions produced an alignment with a total length of 21,054 bp. The Bayesian and likelihood analyses produced a similar topology using the evolution general time reversible + gamma + invariable sites (GTR + G + I), model, with a proportion of invariable sites of 0.968 and a γ shape parameter of 0.062. The reconstruction of the ancestral distribution of Colombian cacao genotypes showed that the Amazonia region was an ancestral centre of distribution (Fig. 2). A group of botanical expedition of caqueta (EBC) accessions collected in the Amazonia region were found to be the earliest diverging lineages in the tree. Some Pacific accessions collected near Tumaco (Nariño) formed a small clade (Fig. 2). Genotypes from the Andes and Caribbean regions were distributed across the phylogenetic tree. The exception was the criollo corpoica fedecacao (CRICF) accessions collected in Cesar from the Andes region, regrouping into a separate supported clade in a derived position (Fig. 2).

To avoid bias, we calculated the genetic diversity indices H_e and H_o , excluding departments with just one sample (Caquetá, Cundinamarca, Guajira and Tolima). The genetic analysis showed that H_o ranged from 0.151 to 0.467 with a mean value of 0.358, while H_e ranged from 0.242 to 0.422 with an average value of 0.367 (Table 1). The results showed that H_o is lower than H_e for Amazonas, Cesar, Chocó, Huila and Magdalena. In contrast, Antioquia, Arauca, Nariño, Norte de Santander, Santander and Valle del Cauca showed higher H_o than H_e .

The PCoA was applied to visualize and investigate the differentiation among populations. It was necessary to remove accessions from Arauca, Caquetá, Cundinamarca, Choco, Guajira, Huila, Tolima and Valle del Cauca because the D_{EST} analysis needs at least 20 individuals per population (Gerlach *et al.*, 2010). The first principal coordinate accounted for 84.73%, while the second explained 6.23% of the variation, together accounting for 90.96% of the total variation (Fig. 3). The PCoA distinguished Norte de Santander (Serranía de los Motilones), Antioquia and Magdalena



Fig. 1. Distribution of collection sites of *Theobroma cacao* genotypes across geographical regions of Colombia. Names of Colombian departments in which samples were collected are indicated. Triangles indicate grid cells used for the phylogenetic diversity (PD) analysis.

(Sierra Nevada de Santa Marta) as the most differentiated locations, while Amazonas, Santander, Cesar and Nariño clustered together on the left side of the biplot. The differentiation was significant ($P < 0.05$) and ranged between 0.001 and 0.555 using 999 permutations.

Phylogenetic diversity analyses

The 279 geolocations of cacao genotypes were mapped in 15 PD grid cells (triangles in Fig. 1). As anticipated, GR was significantly

related to PD (r^2 : 0.7955; online Supplementary Fig. S1). The highest observed PD (about 46%) and GR (68 of the 279 genotypes) were found in the Serranía del Perijá, Cesar, northern Colombia (Fig. 4(a), III*). Areas of high observed PD were also found near the northern tip of the eastern Andes range, the geographically isolated mountains of the Sierra Nevada de Santa Marta, Serranía del Perijá and Serranía de Los Motilones (Fig. 4(b)).

The second-highest concentration of PD and GR was found in the cacao-producing region in Santander, with a PD of 35.2%

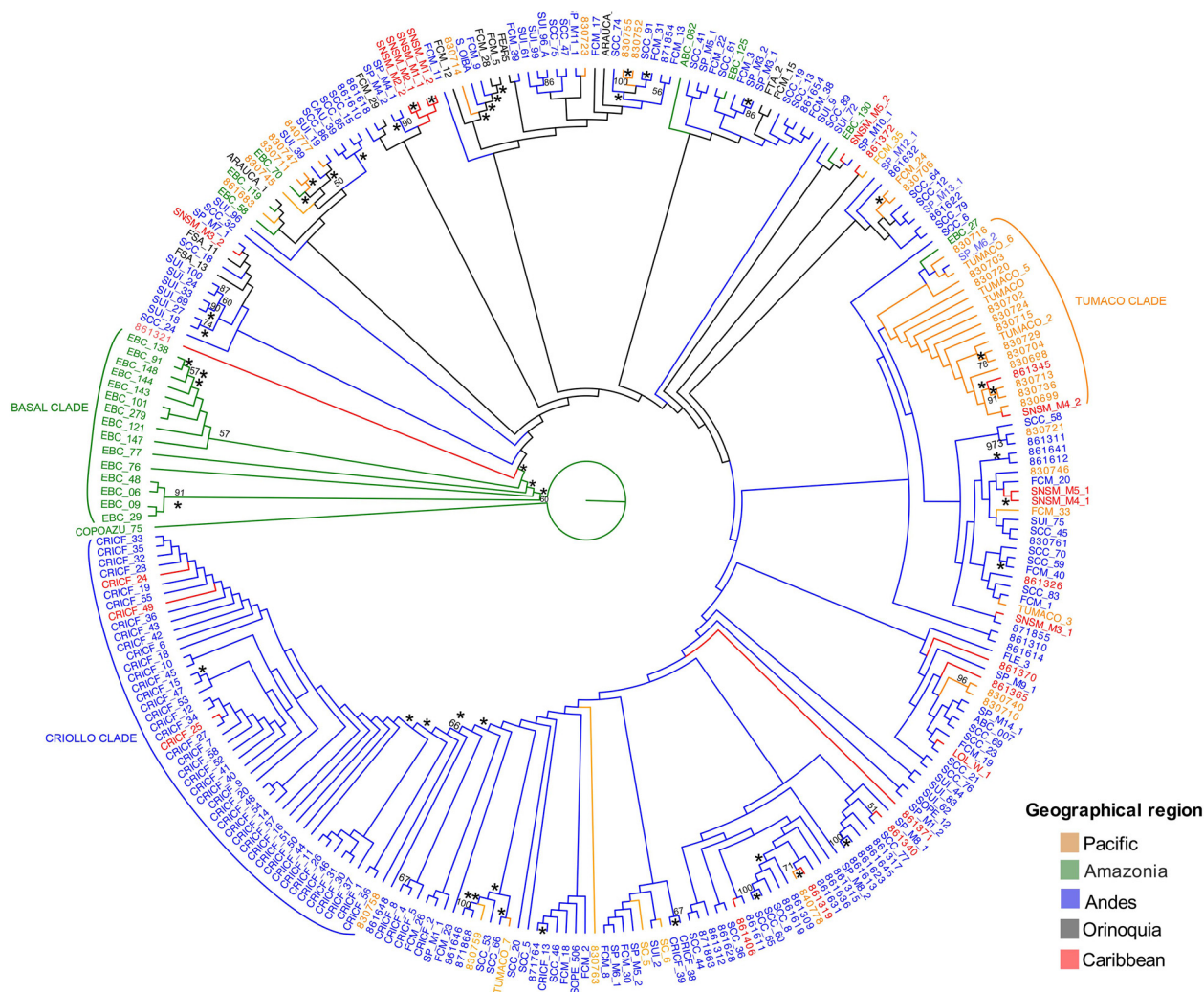


Fig. 2. Phylogeny of *Theobroma cacao* genotypes of Colombia with reconstruction of geographical distribution of genotypes. Bootstrap values higher than 50% are indicated in nodes and posterior probabilities higher than 0.9 are indicated by an asterisk (*). The basal clade corresponds to EBC accessions collected in the Amazonia region, the Tumaco clade corresponds to accessions collected from this location in the Nariño department in the Pacific region and Criollo clade corresponds to CRICF accessions collected in the Cesar department in the Andes region.

(Fig. 4(a) and (b), VI). PD was significant after a randomization test (SPD), demonstrating that the diversity found is more closely related than expected by chance (Fig. 4(c)). Using this analysis, Norte de Santander, Antioquia, Santander, Arauca, Valle del Cauca, Huila, Nariño and Amazonas show a significant SPD, indicating that these regions are diverse (Fig. 4(c)).

Areas of significantly high RPD included the Sierra Nevada de Santa Marta and the Serranía del Perijá in northeast Colombia (Fig. 4(d)). On the contrary, Arauca, Huila and Nariño had a significantly low RPD (Fig. 4(d)).

Discussion

A previous study of Colombian cacao genotypes (Osorio-Guarín *et al.*, 2017) reported high genetic diversity in the germplasm collection but did not consider the PD between materials. The present study is the first to locate centres of PD of Colombian cacao genotypes and disentangle their evolutionary relationships. Our aim was not to reconstruct historical evidence of evolutionary processes but to find a reliable way to validate the diversity of

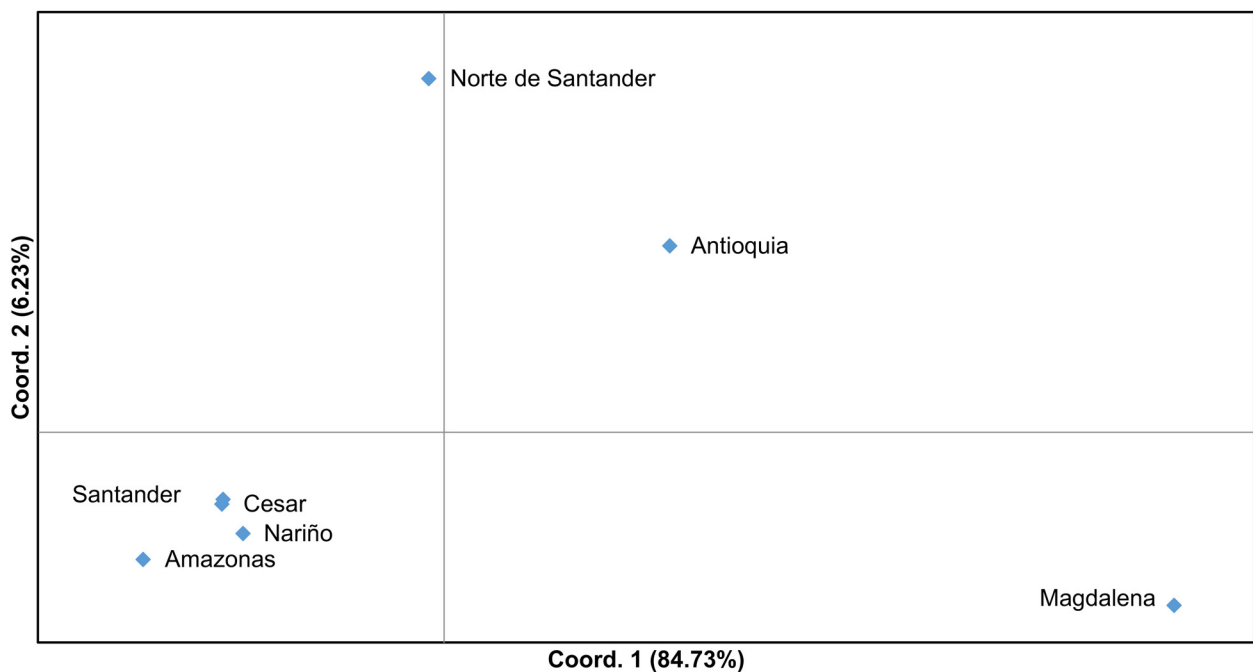
Colombian cacao based on spatial distribution patterns and form a basis for guiding further sampling and increasing the cacao gene pool available for breeding and crop improvement.

The results showed that H_o is lower than H_e for Amazonas, Cesar, Chocó, Huila and Magdalena, indicating excess homozygosity explained by inbreeding or isolation by distance in this cacao population. In contrast, Antioquia, Arauca, Nariño, Norte de Santander, Santander and Valle del Cauca showed higher H_o than H_e , resulting in an excess of heterozygosity, most likely because these locations are cacao-producing regions in which genotypes have been crossbred indiscriminately. We found that the most diverse regions based on H_e values were Magdalena, Nariño, Santander and Norte de Santander. These results can be due to the fact that Santander and Norte de Santander are Colombia’s most important producing regions where breeding programmes have been carried out, while Nariño and Magdalena are considered to have regional materials of hybrid origin (Ballesteros *et al.*, 2016; Ramos Ospino *et al.*, 2020).

The phylogenetic analysis showed low bootstrap values; however, the analysis was performed using two reconstruction

Table 1. Summary of the statistics of genetic diversity calculated with 87 SNP for 275 *Theobroma cacao* accessions

Department	Sample size	Statistic	N_e	H_o	H_e
Antioquia	29	Mean	1.625	0.402	0.357
		SE	0.034	0.022	0.016
Amazonas	21	Mean	1.524	0.226	0.306
		SE	0.037	0.016	0.018
Arauca	11	Mean	1.560	0.397	0.329
		SE	0.034	0.023	0.016
Cesar	68	Mean	1.345	0.151	0.242
		SE	0.021	0.007	0.012
Choco	7	Mean	1.712	0.320	0.390
		SE	0.034	0.020	0.015
Huila	18	Mean	1.646	0.334	0.367
		SE	0.033	0.017	0.015
Magdalena	23	Mean	1.770	0.355	0.422
		SE	0.026	0.013	0.010
Nariño	30	Mean	1.747	0.418	0.413
		SE	0.027	0.014	0.011
Norte de Santander	41	Mean	1.715	0.403	0.400
		SE	0.028	0.014	0.012
Santander	20	Mean	1.738	0.463	0.409
		SE	0.028	0.020	0.012
Valle del Cauca	7	Mean	1.723	0.467	0.398
		SE	0.032	0.025	0.014
Total		Mean	1.646	0.358	0.367
		SE	0.005	0.005	0.003

**Fig. 3.** Principal coordinate analysis (PCoA) run on GenAEx using the DEST pairwise distances.

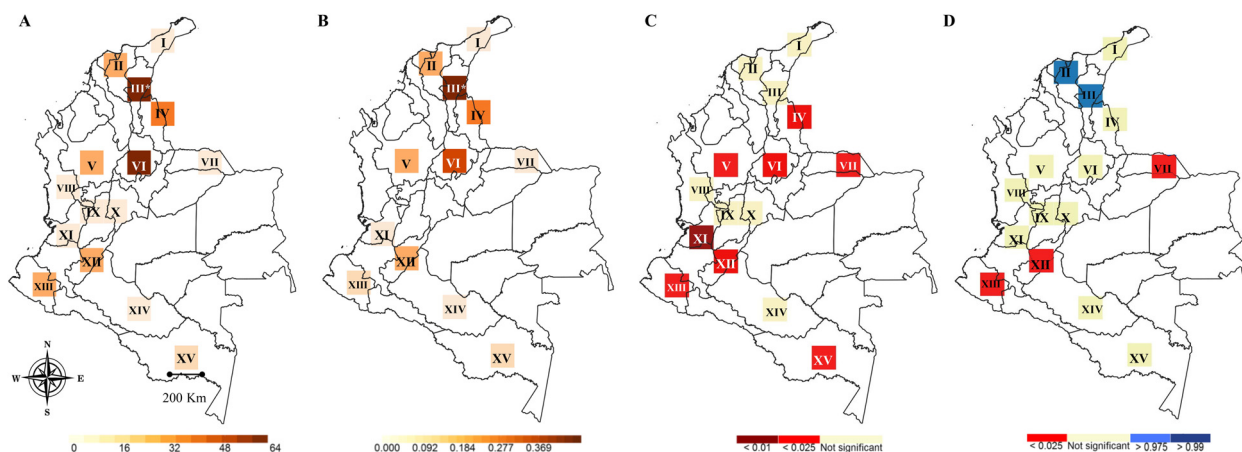


Fig. 4. Observed and randomised diversity patterns of *Theobroma cacao* genotypes in Colombia. (a) Genotypes richness (GR), (b) phylogenetic diversity (PD), (c) significant phylogenetic diversity (SPD), (d) relative phylogenetic diversity (RPD). The sites of interest are numbered as follows: I: Guajira; II: Sierra Nevada de Santa Marta in Magdalena department; III: Serranía del Perijá in Cesar department; IV: Serranía de los Motilones in Norte de Santander department; V: Antioquia; VI: Santander; VII: Arauca; VIII: Choco; IX: Tolima; X: Cundinamarca; XI: Valle del Cauca; XII: Huila; XIII: Nariño; XIV: Caquetá and XV: Amazonas. The asterisk indicates grid cells with the highest observed values of GR and PD, respectively.

methods (ML and Bayesian), resulting in similar topologies and giving robustness to our interpretations. Colombian cacao genotypes result from many historical events, including hybridization (Rodríguez-Medina *et al.*, 2019) that may cause low support of nodes, reported as a problem in phylogenetic studies (McDade, 1990). In the case of Criollo group, the low bootstrap support would be due to the high level of homozygosity (Motamayor *et al.*, 2002). We found a group of Amazonia genotypes (EBC) positioned at the root of the phylogenetic tree (Fig. 2), which is indicative of the ancestral origin and agrees with the studies that recognize the Amazonia region as the centre of origin of cultivated and wild species (Thomas *et al.*, 2012; Zarrillo *et al.*, 2018). The EBC genotypes were collected from the Amazonia region during an expedition to the low parts of the Caquetá river (Allen, 1988). Recently, a study including all the cacao genetic groups showed that samples of EBC genotypes are related to the Ecuadorian group Curaray at the base of the phylogeny (Osorio-Guarín *et al.*, 2017). In addition, a genetic structure analysis based on 9000 SNPs showed that the EBC-06, 09, 29 and EBC-48 have more than 90% of Curaray ancestry (Osorio-Guarín *et al.*, 2020).

We also recovered a clade that regroups most of the Criollo (CRICF) genotypes, a genetic group previously reported as a differentiated one by Motamayor *et al.* (2008). Half of the regional genotypes from Tumaco (Nariño) formed a clade in our phylogenetic tree, probably indicating the distinctiveness of some of the Tumaco materials. This region produces high-quality cacao, winning international prizes, including the Cocoa for Excellence in 2015 (Montoya-Restrepo *et al.*, 2015; Arango, 2017). Most genotypes from Nariño are phylogenetically related because they formed tighter groups of closely related branches, a pattern known as phylogenetic clustering (Webb *et al.*, 2002). The genetic structure analysis of these samples showed that most of their ancestry is related to a mix of the Nacional, Criollo and Amelonado genetic groups (Osorio-Guarín *et al.*, 2020).

Genotypes from the Sierra Nevada de Santa Marta (Magdalena), Serranía de Los Motilones (Norte de Santander) and Serranía del Perijá (Cesar) (excluding some of the CRICF genotypes) were distributed across the tree, suggesting that the genotypes from these sites are distantly related and do not

share a close common ancestor. Patiño Rodríguez (2002) mentioned that in the 1600s, the most crucial region for cacao cultivation was northeastern Colombia (Norte de Santander). Later in the 1950s, the south Pacific region (Valle del Cauca) was the central cacao-producing region, explaining the presence of non-phylogenetic-related cacao genotypes (Patiño Rodríguez, 2002). In these two cases, we observed long and distantly related branches, a pattern known as phylogenetic overdispersion (Webb *et al.*, 2002). Cacao materials with desired agronomic traits from different sources were probably transported to these producing regions, explaining why they are not necessarily closely related. The rest of the genotypes mostly come from the Andes region and are distributed in different shallow clades across the tree (Fig. 2), showing a mixed pattern. For example, some genotypes from Norte de Santander and Antioquia are strongly related (phylogenetic clustering), and some are broadly distributed across the tree (phylogenetic overdispersion).

Areas with a high GR often coincide with areas with high PD, as we found in our study (Fig. 4(a) and (b)). Other studies have also reported a significant correlation (Mishler *et al.*, 2014; Qian *et al.*, 2019; Manish, 2021). However, an index such as GR cannot explain the complexity of the evolutionary events causing the current diversity of taxa. For instance, a study of the flora biodiversity hotspots of the Cape Peninsula in Africa that explored the utility of these indices showed that it is more beneficial to use a decoupled PD from GR because this complex diversity has a solid phylogeographic structure as a consequence of endemic radiations (Forest *et al.*, 2007). For our study, we applied various indices to ensure diversity was understood in different forms.

Observed biodiversity patterns can be deceptive because of different sampling biases (Swenson, 2009; Schmidt-Lebuhn *et al.*, 2012; Tucker and Cadotte, 2013), such as the effect of remoteness on the sampling of field collections. For instance, genotypes from the Pacific, Orinoquia and Amazonia regions, which are remote and difficult to access, are under-represented. Updated biodiversity indices such as randomized RPD have been developed to examine the over-representation of long and deep branches and avoid this bias, which produces significantly high RPD values and is related to phylogenetic overdispersion. The over-representation of short or shallow branches produces significantly

low RPD values related to phylogenetic clustering in the tree (Mishler *et al.*, 2014; Laffan *et al.*, 2016).

We found that the Sierra Nevada de Santa Marta in Magdalena and the Serranía del Perijá in Cesar exhibited significantly high RPD, possibly due to the geographic isolation of these cacao populations (blue grid cells in Fig. 4(d)). Significantly high RPD (phylogenetic overdispersion) can be explained by the occurrence of genotypes in an area containing relicts from past climate change or by a strong environmental heterogeneity that makes different niches available (Mayfield and Levine, 2010; de Bello *et al.*, 2013). It can also be explained by competition between CWRs that do not permit their co-occurrence in the same place (Webb *et al.*, 2002).

The Sierra Nevada de Santa Marta (Magdalena), the Serranía del Perijá (Cesar), and the Serranía de los Motilones (Norte de Santander) are mountains disconnected from the Andes Mountain range. These regions contain long branches, implying the conglomeration of genotypes that possibly diversified recently with some genotypes with an older history. Isolated mountain ranges create micro-niches with singular ecological and climate conditions (dry and hot) that provide unique environments for agriculture and a concentration of endemic species (Webb and Peart, 2000; Webb *et al.*, 2002; Cooper *et al.*, 2011). Our results agree with Bryant *et al.* (2008), who found that angiosperms are more phylogenetically dispersed at higher elevations. In concordance with these findings, the Jost D_{EST} analysis showed that genotypes from the mountain ranges such as Serranía de Los Motilones (Norte de Santander) and Sierra Nevada de Santa Marta (Magdalena) are more differentiated.

The Serranía del Perijá is a hotspot recognized for its high plant endemism levels (Cuatrecasas, 1964; Rivera Díaz and Fernández Alonso, 2003). Most of the CRICF genotypes in this site belong to the Criollo genetic group (Osorio-Guarín *et al.*, 2017), the most genetically differentiated cacao group not only morphologically but also in quality and taste (Motamayor *et al.*, 2013). Despite its distinctiveness, the diversity (H_e : 0.032 and polymorphic sites: 50%) of the Criollo group was found to be low in our study, agreeing with the results of Motamayor *et al.* (2002), which could explain the predominantly short branches in the phylogenetic tree, implying that a population bottleneck probably occurred in this region. The selection and inbreeding of a few individuals caused the reduced genetic diversity and the closely related genotypes.

In contrast, Arauca, Huila and Nariño had significantly low RPDs (phylogenetic clustering), likely due to the conglomeration of cacao lineages that have recently diverged and probably result from hybridization events because these sites are cacao-producing regions. Low RPD is explained by the recent divergence of lineages in an area or by the co-occurrence of close relatives in the same community, excluding weaker competitors (Mishler *et al.*, 2014). For example, Arauca genotypes are distributed in different clades with predominantly short branch lengths, suggesting a homogenization by excessive crosses of closely related genotypes. Most of the Arauca genotypes are selected regionally by the Federación Nacional de Cacaoteros (Fedecacao) based on agronomic traits. In the study of Osorio-Guarín *et al.* (2017), some Arauca materials have approximately 50% of Iquitos ancestry.

High PD, significantly low SPD and RPD values were found in the southern departments of Valle del Cauca and Huila (Fig. 4(b)), suggesting a more recent evolutionary history. Indigenous settlements in both regions have cultivated the lands for years

(1–900 AD), which may explain this high diversity. The archaeological sites of San Agustín in Huila and Calima in Valle del Cauca are known as centres of diversity for different crops (Velandia Jagua, 1999; Piperno *et al.*, 2017). Early evidence of plant food production closely related to native wild ancestors of crops such as squash, arrowroot and cocoyam has also been found in the Calima Valley (Piperno, 2011). Recent evidence in the Amazonia region of Ecuador showed that cultivated cacao has its origins in ancient indigenous sites (Zarrillo *et al.*, 2018). A similar situation could have happened in Colombia on indigenous sites under the assumption that they were centres of food exchange, including cacao.

Most of the cacao genotypes in Colombia are more closely related than expected by chance (Fig. 4(c) and (d)), which can indicate the degradation of the original wild diversity. These genotypes with short branches are probably the result of hybridization, which could be detrimental because of genetic homogenization (Olden *et al.*, 2004; Tieman *et al.*, 2017) and, therefore, endemic losses of gene pools (Charlesworth, 2003). Two hypotheses can explain genetic homogenization: (1) intense inbreeding of the same materials in modern manipulation of crops or (2) cacao cultivation in indigenous sites, causing interbreeding of older lineages.

Furthermore, the PD would generate baseline information to improve cultivated cacao, introducing new genetic resources in a breeding programme whose agronomic performance should be previously assessed. This germplasm would then broaden the gene pool and increase population variation to solve the problems of the crop.

Conclusions

The application of PD helps analyse genetic diversity in germplasm collections, understand the evolutionary relationships among cacao genotypes in Colombia and identify centres of diversity and conservation. It is necessary to prioritize areas climatically stable with over-dispersed genotypes and stressed environments with clustered genotypes to conserve Colombian cacao diversity. Unlike cacao genotypes found in most parts of Colombia with predominantly short and closely related branches, the Amazonia region features long and distantly related branches, making it a likely location for wild cacao diversity. Since samples from the Amazonia region are underrepresented, collection in this region should be a priority to increase relict diversity for further genetic improvements of cultivated cacao. As well as regions with a significantly low RPD, such as Arauca, Huila and Nariño, with a conglomeration of cacao, recently diverged lineages. The Caribbean and northern Andes regions are the main areas where PD and significantly high RPD tend to concentrate. Particularly, the North Andes regions of Magdalena, Cesar and Norte de Santander, which have both relict and recent cacao diversity, should be prioritized as conservation areas. Collecting germplasm from selected priority areas would improve *ex-situ* holdings, provide potential new diversity for cacao improvement and increase the genetic diversity in cultivated materials.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S1479262123000047>

Acknowledgments. We thank Nunzio Knerr, Joe Miller, Mario Porcel and Gina Garzón for providing valuable comments on the manuscript. Special thanks to Allende Pesca and Orlando Guiza for sharing their insights on the history of cacao cultivation in Colombia. We thank Jhon Berdugo,

Eliana Báez and Roberto Coronado for their involvement in developing the genotype dataset. The manuscript was proofread and edited by Julia Alice Veronica de Raadt.

References

- Aikpokpodion PO, Kolesnikova-Allen M, Adetimirin VO, Guiltinan MJ, Eskes AB, Motamayor J-C and Schnell RJ (2010) Population structure and molecular characterization of Nigerian field genebank collections of cacao, *Theobroma cacao* L. *Silvae Genetica* **59**, 273–285.
- Allegre M, Argout X, Boccara M, Fouet O, Roguet Y, Bérard A, Thévenin JM, Chauveau A, Rivallan R, Clement D, Courtois B, Gramacho K, Boland-Augé A, Tahi M, Umaharan P, Brunel D and Lanaud C (2012) Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of *Theobroma cacao* L. *DNA Research* **19**, 23–35.
- Allen JB (1988) *Geographical Variation and Population Biology in Wild Theobroma cacao*. U.K: University of Edinburgh.
- Álvarez JC, Martínez SC and Coy J (2014) Estado de la moniliasis del cacao causada por *Moniliophthora roreri* en Colombia. *Acta Agronomica* **63**, 388–399.
- Arango J (2017) Evaluación del Efecto de Técnicas de Fermentación en el Sabor Y aroma de Cacao CCN-51 (*Theobroma cacao* L.) en la Zona de Tumaco-Nariño. Bogota, Colombia: Universidad Nacional de Colombia.
- Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci AM, Da Silva C, Cascardo J, Allegre M, Kuhn D, Verica J, Courtois B, Loor G, Babin R, Sounigo O, Ducamp M, Guiltinan MJ, Ruiz M, Alemanno L, Machado R, Phillips W, Schnell R, Gilmour M, Rosenquist E, Butler D, Maximova S and Lanaud C (2008) Towards the understanding of the cacao transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions. *BMC Genomics* **9**, 512.
- Ballesteros W, Lagos TC and L HF (2016) Morphological characterization of elite cacao trees (*Theobroma cacao* L.) in Tumaco, Nariño, Colombia. *Revista Colombiana de Ciencias Hortícolas* **9**, 313.
- Batley J and Edwards D (2007) SNP applications in plants. In Oraguzie NC, Rikkerink EHA, Gardiner SE and De Silva HN (ed.), *Association Mapping in Plants*. New York: Springer, pp. 95–102.
- Beg MS, Ahmad S, Jan K and Bashir K (2017) Status, supply chain and processing of cocoa – a review. *Trends in Food Science & Technology* **66**, 108–116.
- Bekele F and Phillips-Mora W (2019) Chapter 12: cacao (*Theobroma cacao* L.) breeding. In Al-Khayri JM, Jain SM and Johnson DV (eds), *Advances in Plant Breeding Strategies: Industrial and Food Crops*. Verlag, New York: Springer, pp. 1–744.
- Benjamin T, Lundy M, Abbott P, Burniske G, Croft M, Fenton M, Kelly C, Rodríguez-Camayo F and Wilcox M (2018) *An Analysis of the Supply Chain of Cacao in Colombia*. Palmira, Colombia: Purdue University and International Center for Tropical Agriculture (CIAT).
- Borrone JW, Brown JS, Kuhn DN, Motamayor JC and Schnell RJ (2007) Microsatellite markers developed from *Theobroma cacao* L. expressed sequence tags. *Molecular Ecology Notes* **7**, 236–239.
- Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ and Green JL (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences* **105**, 11505–11511.
- CacaoNet (2012) A global strategy for the conservation and use of cacao genetic resources, as the foundation for a sustainable cocoa economy. In Laliberté B (ed.), *Montpellier*, France: Bioversity International, pp. 1–28.
- Ceccarelli V, Lastra S, Loor Solórzano RG, Chacón WW, Nolasco M, Sotomayor Cantos IA, Plaza Avellán LF, López DA, Fernández Anchundia FM, Dessauw D, Orozco-Aguilar L and Thomas E (2022) Conservation and use of genetic resources of cacao (*Theobroma cacao* L.) by gene banks and nurseries in six Latin American countries. *Genetic Resources and Crop Evolution* **69**, 1283–1302.
- Charlesworth D (2003) Effects of inbreeding on the genetic diversity of populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* **358**, 1051–1070.
- Cilas C and Bastide P (2020) Challenges to cocoa production in the face of climate change and the spread of pests and diseases. *Agronomy* **10**, 1232.
- Cooper N, Freckleton RP and Jetz W (2011) Phylogenetic conservatism of environmental niches in mammals. *Proceedings of the Royal Society B: Biological Sciences* **278**, 2384–2391.
- Cornejo OE, Yee M-C, Dominguez V, Andrews M, Sockell A, Strandberg E, Livingstone D, Stack C, Romero A, Umaharan P, Royaert S, Tawari NR, Ng P, Gutierrez O, Phillips W, Mockaitis K, Bustamante CD and Motamayor JC (2018) Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Communications Biology* **1**, 167.
- Cosme S, Cuevas HE, Zhang D, Oleksyk TK and Irish BM (2016) Genetic diversity of naturalized cacao (*Theobroma cacao* L.) in Puerto Rico. *Tree Genetics & Genomes* **12**, 88.
- Cuatrecasas J (1964) Cacao and its allies: a taxonomic revision of the genus *Theobroma*. *Contributions from the US Herbarium* **35**, 379–614.
- Daymond A and Bekele F (2022) Cacao. In Priyadarshan PM and Jain SM (eds), *Cash Crops: Genetic Diversity, Erosion, Conservation and Utilization*. Cham: Springer International Publishing, pp. 23–53.
- de Bello F, Vandewalle M, Reitalu T, Lepš J, Prentice HC, Lavorel S and Sykes MT (2013) Evidence for scale- and disturbance-dependent trait assembly patterns in dry semi-natural grasslands. *Journal of Ecology* **101**, 1237–1244.
- Dempewolf H, Baute G, Anderson J, Kilian B, Smith C and Guarino L (2017) Past and future use of wild relatives in crop breeding. *Crop Science* **57**, 1070–1082.
- Díaz-Valderrama JR, Leiva-Espinoza ST and Catherine Aime M (2020) The history of cacao and its diseases in the Americas. *Phytopathology* **110**, 1604–1619.
- DuVal A, Gezan SA, Mustiga G, Stack C, Marelli JP, Chaparro J, Livingstone D, Royaert S and Motamayor JC (2017) Genetic parameters and the impact of off-types for *Theobroma cacao* L. In a breeding program in Brazil. *Frontiers in Plant Science* **8**, 1–12.
- Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61**, 1–10.
- Faith DP and Baker AM (2006) Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evolutionary Bioinformatics* **2**, 121–128.
- Fang W, Meinhardt LW, Mischke S, Bellato CM, Motilal L and Zhang D (2014) Accurate determination of genetic identity for a single cacao bean, using molecular markers with a nanofluidic system, ensures cocoa authentication. *Journal of Agricultural and Food Chemistry* **62**, 481–487.
- FAO STAT Crops and Livestock Products (2022). Available at <https://www.fao.org/faostat/en/#data/QCL> (accessed 14 September 2022).
- Fernández-Niño M, Rodríguez-Cubillos MJ, Herrera-Rocha F, Anzola JM, Cepeda-Hernández ML, Aguirre Mejía JL, Chica MJ, Olarte HH, Rodríguez-López C, Calderón D, Ramírez-Rojas A, Del Portillo P, Restrepo S and González Barrios AF (2021) Dissecting industrial fermentations of fine flavour cocoa through metagenomic analysis. *Scientific Reports* **11**, 8638.
- Fisher KM, Wall DP, Yip KL and Mishler BD (2007) Phylogeny of the Calymperaceae with a rank-free systematic treatment. *The Bryologist* **110**, 46–73.
- Ford-Lloyd B V, Schmidt M, Armstrong SJ, Barazani O, Engels J, Hadas R, Hammer K, Kell SP, Kang D, Khoshbakht K, Li Y, Long C, Lu B-R, Ma K, Nguyen VT, Qiu L, Ge S, Wei W, Zhang Z and Maxted N (2011) Crop wild relatives – undervalued, underutilized and under threat? *BioScience* **61**, 559–565.
- Forest F, Grenyer R, Rouget M, Davies TJ, Cowling RM, Faith DP, Balmford A, Manning JC, Procheş Ş, van der Bank M, Reeves G, Hedderson TAJ and Savolainen V (2007) Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* **445**, 757–760.
- Gerlach G, Jueterbock A, Kraemer P, Deppermann J and Harmand P (2010) Calculations of population differentiation based on GST and D: forget GST but not all of statistics. *Molecular Ecology* **19**, 3845–3852.
- González-Orozco CE, Mishler BD, Miller JT, Laffan SW, Knerr N, Unmack P, Georges A, Thornhill AH, Rosauer DF and Gruber B (2015) Assessing biodiversity and endemism using phylogenetic methods across multiple taxonomic groups. *Ecology and Evolution* **5**, 5177–5192.

- González-Orozco CE, Galán AAS, Ramos PE and Yockteng R (2020) Exploring the diversity and distribution of crop wild relatives of cacao (*Theobroma cacao* L.) in Colombia. *Genetic Resources and Crop Evolution* 67, 2071–2085.
- González-Orozco CE, Sosa CC, Thornhill AH and Laffan SW (2021) Phylogenetic diversity and conservation of crop wild relatives in Colombia. *Evolutionary Applications* 14, 2603–2617.
- Gopaulchan D, Motilal LA, Bekele FL, Clause S, Ariko JO, Ejang HP and Umaharan P (2019) Morphological and genetic diversity of cacao (*Theobroma cacao* L.) in Uganda. *Physiology and Molecular Biology of Plants* 25, 361–375.
- Gopaulchan D, Motilal LA, Kalloo RK, Mahabir A, Moses M, Joseph F and Umaharan P (2020) Genetic diversity and ancestry of cacao (*Theobroma cacao* L.) in Dominica revealed by single nucleotide polymorphism markers. *Genome* 63, 583–595.
- Guerrero-Rincón A, Pabón VS and Ferreira EC (1998) *Los Pueblos Del Cacao: Orígenes de Los Asentamientos Urbanos En El Oriente Colombiano*. Bucaramanga, Colombia: Universidad Industrial de Santander, Escuela de Historia.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W and Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59, 307–321.
- Gutiérrez OA, Puig AS, Phillips-Mora W, Bailey BA, Ali SS, Mockaitis K, Schnell RJ, Livingstone D, Mustiga G, Royaert S and Motamayor JC (2021) SNP markers associated with resistance to frosty pod and black pod rot diseases in an F1 population of *Theobroma cacao* L. *Tree Genetics and Genomes* 17, 28.
- Heywood V, Casas A, Ford-Lloyd B, Kell S and Maxted N (2007) Conservation and sustainable use of crop wild relatives. *Agriculture, Ecosystems and Environment* 121, 245–255.
- Huelsenbeck JP and Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755.
- Irish BM, Goenaga R, Zhang D, Schnell R, Brown JS and Motamayor JC (2010) Microsatellite fingerprinting of the USDA-ARS tropical agriculture research station cacao (*Theobroma cacao* L.) germplasm collection. *Crop Science* 50, 656–667.
- Ji K, Zhang D, Motilal LA, Boccara M, Lachenaud P and Meinhardt LW (2013) Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. *Genetic Resources and Crop Evolution* 60, 441–453.
- Just L (2008) G(ST) and its relatives do not measure differentiation. *Molecular ecology* 17, 4015–4026.
- Kapli P, Yang Z and Telford MJ (2020) Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* 21, 428–444.
- Kodoth N (2021) *Cacao. Tree Crops: Harvesting Cash from the World's Important Cash Crops*. Cham, Switzerland: Springer International Publishing, pp. 153–210.
- Laffan SW, Lubarsky E and Rosauer DF (2010) Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography* 33, 643–647.
- Laffan SW, Rosauer DF, Di Virgilio G, Miller JT, González-Orozco CE, Knerr N, Thornhill AH and Mishler BD (2016) Range-weighted metrics of species and phylogenetic turnover can better resolve biogeographic transition zones. *Methods in Ecology and Evolution* 7, 580–588.
- Laity T, Laffan SW, González-Orozco CE, Faith DP, Rosauer DF, Byrne M, Miller JT, Crayn D, Costion C, Moritz CC and Newport K (2015) Phylodiversity to inform conservation policy: an Australian example. *Science of the Total Environment* 534, 131–143.
- Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A and Lagoda PJJ (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. *Molecular Ecology* 8, 2141–2143.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948.
- Lefort V, Longueville J-E and Gascuel O (2017) SMS: smart model selection in PhyML. *Molecular Biology and Evolution* 34, 2422–2424.
- Livingstone DS, Motamayor JC, Schnell RJ, Cariaga K, Freeman B, Meerow AW, Brown JS and Kuhn DN (2011) Development of single nucleotide polymorphism markers in *Theobroma cacao* and comparison to simple sequence repeat markers for genotyping of Cameroon clones. *Molecular Breeding* 27, 93–106.
- Livingstone D, Royaert S, Stack C, Mockaitis K, May G, Farmer A, Saski C, Schnell R, Kuhn D and Motamayor JC (2015) Making a chocolate chip: development and evaluation of a 6 K SNP array for *Theobroma cacao*. *DNA Research* 22, 279–291.
- Lopes UV, Reis Monteiro W, Pires JL, Clement D, Yamada MM and Gramacho KP (2011) Cacao breeding in Bahia, Brazil: strategies and results. *Crop Breeding and Applied Biotechnology* 1, 73–81.
- López-Hernández MdP, Sandoval-Aldana AP, García-Lozano J and Criollo-Nuñez J (2021) Estudio morfoagronómico de materiales de cacao (*Theobroma cacao* L.) de diferentes zonas productoras en Colombia. *Ciencia y Agricultura* 18, 98–109.
- Maddison WP and Maddison DR (2021) Mesquite: a modular system for evolutionary analysis.
- Mahabir A, Motilal LA, Gopaulchan D, Ramkissoon S, Sankar A and Umaharan P (2020) Development of a core SNP panel for cacao (*Theobroma cacao* L.) identity analysis. *Genome* 63, 103–114.
- Majeed S, Chaudhary MT, Hulse-Kemp AM and Azhar MT (2021) Chapter 1 – introduction: crop wild relatives in plant breeding. In Azhar MT and Wani SHBT-WG for GI in CP (ed.), *Wild Germplasm for Genetic Improvement in Crop Plants*. Cambridge, United States: Academic Press, pp. 1–18.
- Mammadov J, Aggarwal R, Buyyarapu R and Kumpatla S (2012) SNP markers and their impact on plant breeding. *International Journal of Plant Genomics* 2012, 1–10.
- Manish K (2021) Species richness, phylogenetic diversity and phylogenetic structure patterns of exotic and native plants along an elevational gradient in the Himalaya. *Ecological Processes* 10, 64.
- Maxted N, Ford-Lloyd B V, Jury S, Kell S and Scholten M (2006) Towards a definition of a crop wild relative. *Biodiversity and Conservation* 15, 2673–2685.
- Maxted N, Kell S, Toledo Á, Dulloo E, Heywood V, Hodgkin T, Hunter D, Guarino L, Jarvis A and Ford-Lloyd B (2010) A global approach to crop wild relative conservation: securing the gene pool for food and agriculture. *Kew Bulletin* 65, 561–576.
- Maxted N, Kell S, Ford-Lloyd B, Dulloo E and Toledo Á (2012) Toward the systematic conservation of global crop wild relative diversity. *Crop Science* 52, 774–785.
- Mayfield MM and Levine JM (2010) Opposing effects of competitive exclusion on the phylogenetic structure of communities. *Ecology Letters* 13, 1085–1093.
- McDade L (1990) Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implications for cladistic analysis. *Evolution* 44, 1685–1700.
- McElroy MS, Navarro AJR, Mustiga G, Stack C, Gezan S, Peña G, Sarabia W, Saucicela D, Sotomayor I, Douglas GM, Migicovsky Z, Amores F, Tarqui O, Myles S and Motamayor JC (2018) Prediction of cacao (*Theobroma cacao*) resistance to *Moniliophthora* spp. diseases via genome-wide association analysis and genomic selection. *Frontiers in Plant Science* 9, 1–12.
- Miller MA, Pfeiffer W and Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop, GCE 2010*, 1–7.
- Miller JT, Jolley-Rogers G, Mishler BD and Thornhill AH (2018) Phylogenetic diversity is a better measure of biodiversity than taxon counting. *Journal of Systematics and Evolution* 56, 663–667.
- Miranda F (1962) Wild cacao in the Lacandona forest. *Cacao (Turrialba)* 7, 7.
- Mishler B (2021) *What, If Anything, Are Species?* Portland, USA: CRC Press.
- Mishler BD, Knerr N, González-Orozco CE, Thornhill AH, Laffan SW and Miller JT (2014) Phylogenetic measures of biodiversity and neo- and paleo-endemism in Australian Acacia. *Nature Communications* 5, 4473.
- Montoya-Restrepo IA, Montoya-Restrepo LA and Lowy-Ceron PD (2015) Oportunidades para la actividad cacaotera en el municipio de Tumaco, Nariño, Colombia. *Entramado* 11, 48–59.

- Morillo Y, Morillo AC, Muñoz JE, Ballesteros W and González A (2014) Caracterización molecular con microsatélites amplificados al azar (RAMs) de 93 genotipos de cacao (*Theobroma cacao* L.). *Agronomía Colombiana* 32, 315–325.
- Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A and Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity* 89, 380.
- Motamayor JC, Lachenaud P, Wallace J, Looor R, Kuhn DN, Brown S, Schnell RJ, da Silva e Mota JW, Looor R, Kuhn DN, Brown JS and Schnell RJ (2008) Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). *PLoS ONE* 3, 1–8.
- Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, Saski C, Jenkins J, Podicheti R, Zhao M, Scheffler BE, Stack JC, Feltus FA, Mustiga GM, Amores F, Phillips W, Marelli JP, May GD, Shapiro H, Ma J, Bustamante CD, Schnell RJ, Main D, Gilbert D, Parida L and Kuhn DN (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology* 14, r53.
- Nagalingum NS, Knerr N, Laffan SW, González-Orozco CE, Thornhill AH, Miller JT and Mishler BD (2015) Continental scale patterns and predictors of fern richness and phylogenetic diversity. *Frontiers in Genetics* 6, 132.
- Olden JD, Poff NL, Douglas MR, Douglas ME and Fausch KD (2004) Ecological and evolutionary consequences of biotic homogenization. *Trends in Ecology & Evolution* 19, 18–24.
- Osorio-Guarín JA, Berdugo-Cely J, Coronado RA, Zapata YP, Quintero C, Gallego-Sánchez G and Yockteng R (2017) Colombia a source of cacao genetic diversity as revealed by the population structure analysis of germplasm bank of *Theobroma cacao* L. *Frontiers in Plant Science* 8, 1994.
- Osorio-Guarín JA, Quackenbush CR and Cornejo OE (2018) Ancestry informative alleles captured with reduced representation library sequencing in *Theobroma cacao*. *PLoS ONE* 13, 1–14.
- Osorio-Guarín JA, Berdugo-Cely JA, Coronado-Silva RA, Baez E, Jaimes Y and Yockteng R (2020) Genome-wide association study reveals novel candidate genes associated with productivity and disease resistance to *Moniliophthora* spp. in cacao (*Theobroma cacao* L.). *G3: Genes, Genomes, Genetics* 10, 1713–1725.
- Patino Rodríguez VM (2002) *Historia y Dispersión de Los Frutales Nativos Del Neotrópico*. Cali, Colombia: Centro Internacional de Agricultura Tropical (CIAT).
- Peakall R and Smouse P (2006) genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6, 288–295.
- Peakall R and Smouse P (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics* 1, 6–8.
- Piperno DR (2011) The origins of plant cultivation and domestication in the new world tropics. *Current Anthropology* 52, S453–S470.
- Piperno DR, Ranere AJ, Dickau R and Aceituno F (2017) Niche construction and optimal foraging theory in Neotropical agricultural origins: a re-evaluation in consideration of the empirical evidence. *Journal of Archaeological Science* 78, 214–220.
- Pugh T, Fouet O, Risterucci AM, Brottier P, Abouladze M, Deletrez C, Courtois B, Clement D, Larmande P, N'Goran JAK and Lanaud C (2004) A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theoretical and Applied Genetics* 108, 1151–1161.
- Qian H, Deng T, Jin Y, Mao L, Zhao D and Ricklefs RE (2019) Phylogenetic dispersion and diversity in regional assemblages of seed plants in China. *Proceedings of the National Academy of Sciences* 116, 23192–23201.
- Rambaut A, Suchard MA, Xie D and Drummond A (2018) Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67, 901–904.
- Ramos Ospino A, Gómez Alvaréz M, Machado-Sierra E and Aranguren Y (2020) Caracterización fenotípica y genotípica de cultivares de cacao (*Theobroma cacao* L.) de Dibulla, La Guajira, Colombia [Phenotypic and genotypic characterization of cacao cultivars (*Theobroma cacao* L.) from Dibulla, La Guajira, Colombia. *Ciencia & Tecnología Agropecuaria* 21, 1–17.
- R Core Team and R development core team (2008) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Renzi JP, Coyne CJ, Berger J, von Wettberg E, Nelson M, Ureta S, Hernández F, Smýkal P and Brus J (2022) How could the use of crop wild relatives in breeding increase the adaptation of crops to marginal environments? *Frontiers in Plant Science* 13, 886162.
- Rivera Díaz O and Fernández Alonso JL (2003) Análisis corológico de la flora endémica de la Serranía de Perijá, Colombia. *Anales del Jardín Botánico de Madrid* 60, 347–369.
- Rodríguez-Medina C, Caicedo Arana A, Sounigo O, Argout X, Alvarado GA and Yockteng R (2019) Cacao breeding in Colombia, past, present and future. *Breeding Science* 69, 373–382.
- Romero Navarro JA, Phillips-Mora W, Arciniegas-Leal A, Mata-Quirós A, Haiminen N, Mustiga G, Livingstone III D, van Bakel H, Kuhn DN, Parida L, Kasarskis A, Motamayor JC, Romero-Navarro JA, Phillips-Mora W, Arciniegas-Leal A, Mata-Quirós A, Haiminen N, Mustiga G, Livingstone III D, van Bakel H, Kuhn DN, Parida L, Kasarskis A and Motamayor JC (2017) Application of genome wide association and genomic prediction for improvement of cacao productivity and resistance to black and frosty pod diseases. *Frontiers in Plant Science* 8, 1905.
- Saunders JA, Mischke S, Leamy EA and Hemeida AA (2004) Selection of international molecular standards for DNA fingerprinting of *Theobroma cacao*. *Theoretical and Applied Genetics* 110, 41–47.
- Schmidt-Lebuhn AN, Knerr NJ and González-Orozco CE (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. *Journal of Biogeography* 39, 2072–2080.
- Soulebeau A, Pellens R, Lowry PP, Aubriot X, Evans MEK and Haevermans T (2016) Conservation of phylogenetic diversity in Madagascar's largest endemic plant family, Sarcolaenaceae BT. In Pellens R and Grandcolas P (eds), *Biodiversity Conservation and Phylogenetic Systematics: Preserving our Evolutionary Heritage in an Extinction Crisis*. Cham: Springer International Publishing, pp. 355–374.
- Swenson NG (2009) Phylogenetic resolution and quantifying the phylogenetic diversity and dispersion of communities. *PLoS ONE* 4, e4390.
- Thomas E, van Zonneveld M, Loo J, Hodgkin T, Galluzzi G and van Etten J (2012) Present spatial diversity patterns of *Theobroma cacao* L. in the Neotropics reflect genetic differentiation in Pleistocene refugia followed by human-influenced dispersal. *PLoS ONE* 7, e47676.
- Thornhill AH, Mishler BD, Knerr NJ, González-Orozco CE, Costion CM, Crayn DM, Laffan SW and Miller JT (2016) Continental-scale spatial phylogenetics of Australian angiosperms provides insights into ecology, evolution and conservation. *Journal of Biogeography* 43, 2085–2098.
- Tieman D, Zhu G, Resende MFR, Lin T, Nguyen C, Bies D, Rambla JL, Beltran KSO, Taylor M, Zhang B, Ikeda H, Liu Z, Fisher J, Zemach I, Monforte A, Zamir D, Granell A, Kirst M, Huang S and Klee H (2017) A chemical genetic roadmap to improved tomato flavor. *Science* 355, 391–394.
- Tucker CM and Cadotte MW (2013) Unifying measures of biodiversity: understanding when richness and phylogenetic diversity should be congruent. *Diversity and Distributions* 19, 845–854.
- Velandia Jagua CA (1999) The archaeological culture of San Agustín. Towards a new interpretation. In Politis G and Alberti B (eds), *Archaeology in Latin America*. London, New York: Routledge, pp. 185–215.
- Vignati F and Gómez-García R (2020) Iniciativa Latinoamericana Del Cacao: Boletín No. 8. CAF – Banco de Desarrollo de América Latina: Caracas.
- Vincent H, Wiersema J, Kell S, Fielder H, Dobbie S, Castañeda-Álvarez NP, Guarino L, Eastwood R, León B and Maxted N (2013) A prioritized crop wild relative inventory to help underpin global food security. *Biological Conservation* 167, 265–275.
- Voora V, Larrea C, Huppé G and Nugnes F (2022) IISD's State of Sustainability Initiatives review: standards and investments in sustainable agriculture.
- Wang B, Motilal LA, Meinhardt LW, Yin J and Zhang D (2020) Molecular characterization of a cacao germplasm collection maintained in Yunnan, China using single nucleotide polymorphism (SNP) markers. *Tropical Plant Biology* 13, 359–370.

- Webb CO and Peart DR** (2000) Habitat associations of trees and seedlings in a Bornean rain forest. *Journal of Ecology* **88**, 464–478.
- Webb CO, Ackerly DD, McPeck MA and Donoghue MJ** (2002) Phylogenies and community ecology. *Annual Review of Ecology and Systematics* **33**, 475–505.
- Wickramasuriya AM and Dunwell JM** (2018) Cacao biotechnology: current status and future prospects. *Plant Biotechnology Journal* **16**, 4–17.
- Zarrillo S, Gaikwad N, Lanaud C, Powis T, Viot C, Lesur I, Fouet O, Argout X, Guichoux E, Salin F, Solorzano RL, Bouchez O, Vignes H, Severts P, Hurtado J, Yepez A, Grivetti L, Blake M and Valdez F** (2018) The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. *Nature Ecology & Evolution* **2**, 1879–1888.
- Zhang D and Motilal L** (2016) Origin, dispersal, and current global distribution of cacao genetic diversity BT – cacao diseases: a history of old enemies and new encounters. In Bailey BA and Meinhardt LW (eds), *Cacao Diseases: A History of Old Enemies and New Encounters*. Cham: Springer International Publishing, pp. 3–31.
- Zhang D, Mischke S, Johnson ES, Phillips-Mora W and Meinhardt L** (2009) Molecular characterization of an international cacao collection using micro-satellite markers. *Tree Genetics and Genomes* **5**, 1–10.
- Zhang D, Martínez WJ, Johnson ES, Somarriba E, Phillips-Mora W, Astorga C, Mischke S and Meinhardt LW** (2012) Genetic diversity and spatial structure in a new distinct *Theobroma cacao* L. population in Bolivia. *Genetic Resources and Crop Evolution* **59**, 239–252.
- Zhang H, Mittal N, Leamy LJ, Barazani O and Song BH** (2017) Back into the wild – apply untapped genetic diversity of wild relatives for crop improvement. *Evolutionary Applications* **10**, 5–24.