nature plants

ARTICLES

https://doi.org/10.1038/s41477-022-01144-8

Check for updates

OPEN

State of ex situ conservation of landrace groups of 25 major crops

Julian Ramirez-Villegas[1,2,3,25]✉, Colin K. Khoury[1,4,25]✉, Harold A. Achicanoy[1,25], Maria Victoria Diaz[1,25], Andres C. Mendez[1,25], Chrystian C. Sosa[1,5,6,25], Zakaria Kehel[7], Luigi Guarino[8], Michael Abberton[9], Jorrel Aunario[10], Bashir Al Awar[11], Juan Carlos Alarcon[12], Ahmed Amri[7], Noelle L. Anglin[13,14], Vania Azevedo[13,15], Khadija Aziz[7], Grace Lee Capilit[10], Oswaldo Chavez[13], Dmytro Chebotarov[10], Denise E. Costich[12], Daniel G. Debouck[1], David Ellis[13], Hamidou Falalou[16], Albert Fiu[17], Michel Edmond Ghanem[18], Peter Giovannini[8], Alphonse J. Goungoulou[19], Badara Gueye[9], Amal Ibn El Hobyb[7], Ramni Jamnadass[20], Chris S. Jones[21], Bienvenu Kpeki[19], Jae-Sung Lee[10], Kenneth L. McNally[10], Alice Muchugi[21], Marie-Noelle Ndjiondjop[19], Olaniyi Oyatomi[9], Thomas S. Payne[12], Senthil Ramachandran[15], Genoveva Rossel[13], Nicolas Roux[22], Max Ruas[22], Carolina Sansaloni[12], Julie Sardos[22], Tri Deri Setiyono[10,23], Marimagne Tchamba[9], Ines van den Houwe[24], J. Alejandro Velazquez[12], Ramaiah Venuprasad[10], Peter Wenzl[1], Mariana Yazbek[11] and Cristian Zavala[12]

Crop landraces have unique local agroecological and societal functions and offer important genetic resources for plant breeding. Recognition of the value of landrace diversity and concern about its erosion on farms have led to sustained efforts to establish ex situ collections worldwide. The degree to which these efforts have succeeded in conserving landraces has not been comprehensively assessed. Here we modelled the potential distributions of eco-geographically distinguishable groups of landraces of 25 cereal, pulse and starchy root/tuber/fruit crops within their geographic regions of diversity. We then analysed the extent to which these landrace groups are represented in genebank collections, using geographic and ecological coverage metrics as a proxy for genetic diversity. We find that ex situ conservation of landrace groups is currently moderately comprehensive on average, with substantial variation among crops; a mean of 63% ± 12.6% of distributions is currently represented in genebanks. Breadfruit, bananas and plantains, lentils, common beans, chickpeas, barley and bread wheat landrace groups are among the most fully represented, whereas the largest conservation gaps persist for pearl millet, yams, finger millet, groundnut, potatoes and peas. Geographic regions prioritized for further collection of landrace groups for ex situ conservation include South Asia, the Mediterranean and West Asia, Mesoamerica, sub-Saharan Africa, the Andean mountains of South America and Central to East Asia. With further progress to fill these gaps, a high degree of representation of landrace group diversity in genebanks is feasible globally, thus fulfilling international targets for their ex situ conservation.

Crop landraces, also known as farmers' traditional, heritage, folk or heirloom varieties, are cultivated plant populations developed and managed by Indigenous or traditional agrarian cultures through cultivation, selection and diffusion[1]. Having recognizable characteristics and geographic origins, landraces continue to be cultivated by these communities in many regions for their unique agroecological and societal functions and services[1,2]. These typically genetically heterogeneous populations are commonly planted in a mosaic of different crop species and varieties, in combinations sustaining local agricultural resilience and adaptive capacity, human nutrition and cultural needs[1,2]. Farmer-based exchange[3] and gene flow among landrace populations, occasionally

[1]International Center for Tropical Agriculture (CIAT), Cali, Colombia. [2]CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), Cali, Colombia. [3]Wageningen University & Research (WUR), Plant Production Systems Group, Wageningen, The Netherlands. [4]San Diego Botanic Garden, Encinitas, CA, USA. [5]Pontificia Universidad Javeriana Cali, Cali, Colombia. [6]Universidad del Quindío, Armenia, Colombia. [7]International Center for Agricultural Research in the Dry Areas (ICARDA), Rabat, Morocco. [8]Global Crop Diversity Trust, Bonn, Germany. [9]International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. [10]International Rice Research Institute (IRRI), Los Baños, Philippines. [11]International Center for Agricultural Research in the Dry Areas (ICARDA), Beirut, Lebanon. [12]International Maize and Wheat Improvement Center (CIMMYT), Texcoco, México. [13]International Potato Center (CIP), Lima, Peru. [14]United States Department of Agriculture (USDA), Agricultural Research Service, Aberdeen, ID, USA. [15]International Crops Research Institute for the Semi-arid Tropics (ICRISAT), Hyderabad, India. [16]International Crops Research Institute for the Semi-arid Tropics (ICRISAT), Niamey, Niger. [17]Centre for Pacific Crops and Trees (CePaCT), Narere, Fiji. [18]Mohammed VI Polytechnic University (UM6P), Benguerir, Morocco. [19]Africa Rice Center (AfricaRice), Bouaké, Côte d'Ivoire. [20]World Agroforestry (ICRAF), Nairobi, Kenya. [21]International Livestock Research Institute (ILRI), Addis Ababa, Ethiopia. [22]Bioversity International, Montpellier, France. [23]Louisiana State University, Baton Rouge, LA, USA. [24]Bioversity International, Leuven, Belgium. [25]These authors contributed equally: Julian Ramirez-Villegas, Colin K. Khoury, Harold Achicanoy, Maria Victoria Diaz, Andres Mendez, Chrystian C. Sosa. ✉e-mail: j.r.villegas@cgiar.org; c.khoury@cgiar.org

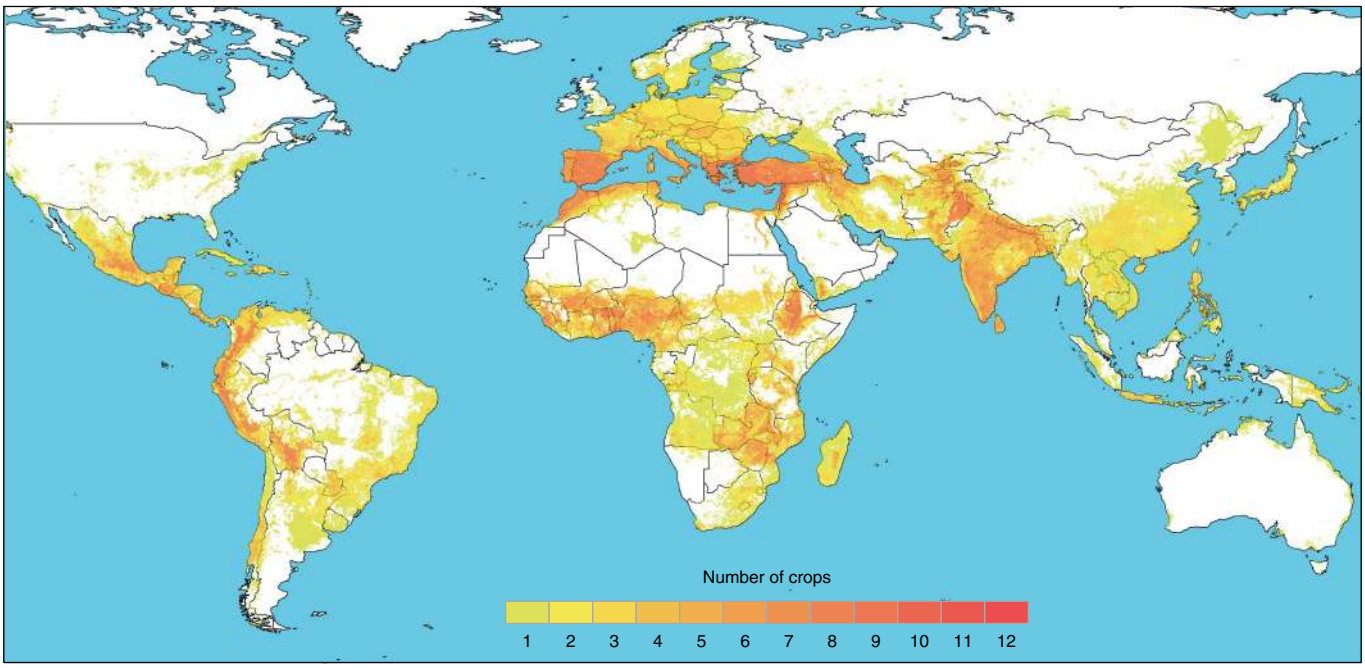


Fig. 1 | Richness map of the predicted distributions of landrace groups of 25 cereal, pulse and starchy root/tuber/fruit crops within their geographic regions of diversity. Darker colours indicate greater numbers of crop landrace groups potentially overlapping in the same 2.5-arc-minute cells, quantified in terms of number of crops. See Extended Data Fig. 1 for richness across all 71 landrace groups within the 25 crops.

also involving modern cultivars[4] or wild progenitors[5], encourage the development of new variation, while longstanding cultivation and selection lead to adaptation to local environmental and societal conditions[6].

Landrace diversity is an essential genetic resource for modern crop breeding[7] and is key to understanding agricultural origins and domestication processes[8]. Landraces are typically accessed via ex situ repositories, called genebanks, for these research purposes. Efforts to collect landraces for genebank conservation have often prioritized sampling from geographic regions and cultures wherein crops were domesticated and/or have been cultivated for a very long time, in recognition of the extraordinary genetic variation in landraces found in these environments[1,7,9]. These activities have gained urgency since the 1960s as economic, agricultural, demographic, environmental and climatic changes increasingly impact in situ populations[1,7]. The result of these collection efforts has been the assemblage of approximately three million landrace samples in international, regional, national and subnational genebanks[10].

Despite these extensive efforts, landrace diversity is not commonly considered to be comprehensively represented ex situ, and major international agreements, including the Convention on Biological Diversity (CBD) Aichi Target 13 (ref. [11]) and the Sustainable Development Goals (SDGs) Target 2.5 (ref. [12]), urgently prioritize the resolution of this conservation gap. To reach these targets, information about the current distributions of landraces and their degree of representation in genebanks is needed. To respond to this need, in this Article we employ a conservation gap analysis methodology[13] to predict the distributions and quantify the current ex situ conservation status of 71 eco-geographically distinguishable groups of landraces within 25 cereal, pulse and starchy root/tuber/fruit crops whose genetic resources are researched and conserved by CGIAR international agricultural research centres or by the Centre for Pacific Crops and Trees (CePaCT) of the Pacific Community (SPC). We identify gaps in existing ex situ collections to inform further collecting efforts.

Results

Geographic distributions of crop landrace groups. On the basis of correlations among 93,269 landrace occurrences of 25 crops (61.9% of occurrences having pre-assigned landrace group assignments and the rest inferred) and 50 environmental and socioeconomic predictor variables, landraces as a whole were predicted to be distributed on all inhabited continents, including throughout most of the world's tropical and subtropical lands (Fig. 1 and Extended Data Fig. 1). Regions with particularly high levels of richness across crops were projected in East and Southern Africa, South and Central Asia, the Mediterranean and West Asia, West Africa and the Andean mountains of South America and Mesoamerica, with landraces of up to 12 of the 25 crops potentially cultivated within single 2.5-arc-minute grid cells in Bangladesh, Ethiopia, India, Nepal and Pakistan. These geographic concentrations of landrace group diversity align well with the historically recognized centres of origin and primary regions of diversity of the world's major crops[14,15]. Notably less landrace diversity across crops was predicted to be cultivated in most temperate regions, in some very arid zones such as the Saharan Desert and in a few highly mesic areas such as the Amazon Basin.

The predicted distributions of the five major races of sorghum are provided in Fig. 2a as an example of landrace-group-level results (the Supplementary Information presents the occurrences and predicted distributions of landrace groups for all assessed crops). Sorghum landrace group ranges were modelled throughout the crop's main regions of diversity in Africa, South Asia, the Mediterranean and West Asia. Its races inhabit distinct eco-geographic ranges but also overlap in specific areas, particularly in Southern and West Africa and in South Asia. Regarding different types of assessed crops, cereal and pulse landrace group diversity was predicted to be particularly rich in South and Central Asia; West, East and Southern Africa; the Mediterranean and West Asia; Europe; the Andean mountains; and Mesoamerica. Meanwhile, starchy root/tuber/fruit crop landrace group richness was concentrated in Mesoamerica, Southeast Asia and the Pacific, South America, West Africa and South Asia (Extended Data Figs. 2–4).

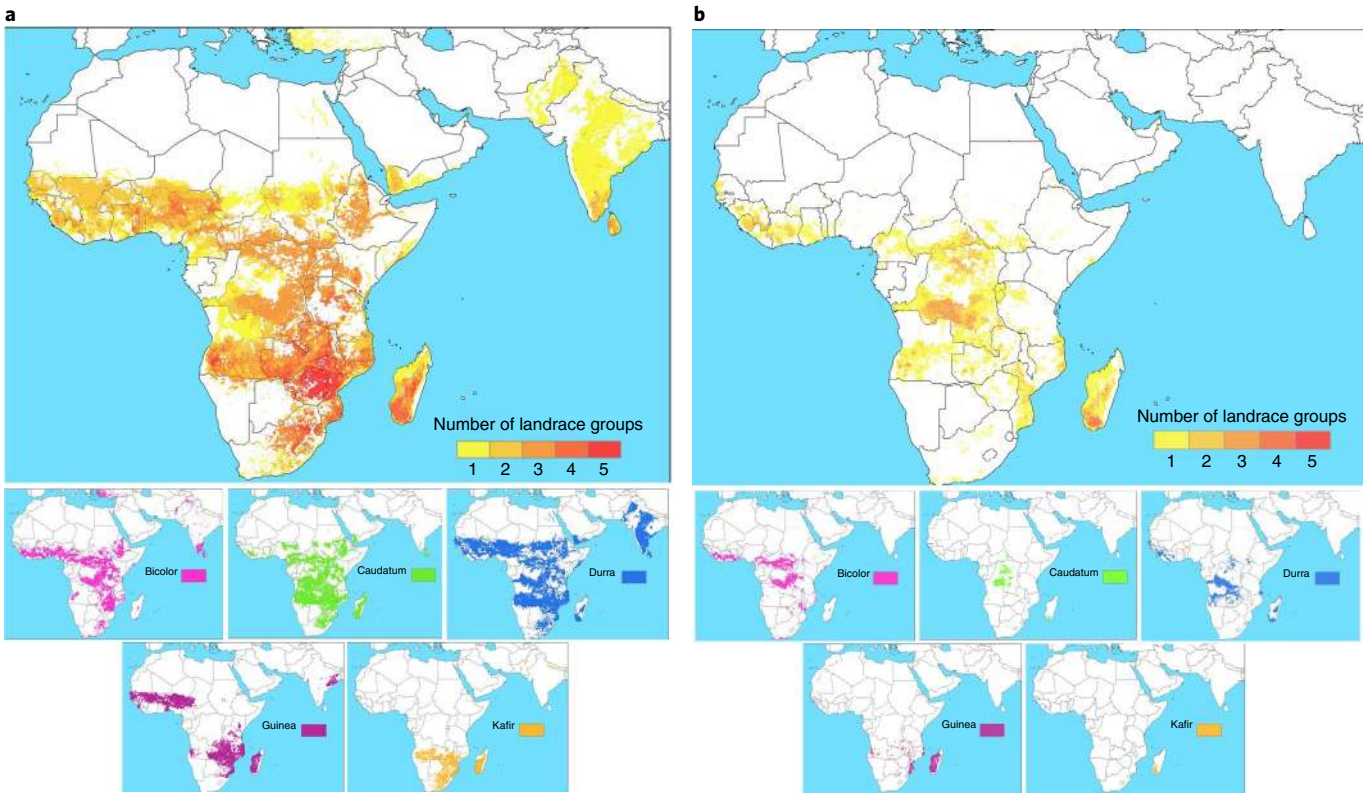


Fig. 2 | Richness maps of sorghum landrace group distributions and ex situ conservation gaps. a,**b**, Predicted distributions (**a**) and ex situ conservation gaps (**b**) for five landrace groups of sorghum in Africa, South Asia, the Mediterranean, and West Asia—namely, the races bicolor, caudatum, durra, guinea and kafir. Small maps, individual distributions of each landrace group; large maps, richness at the crop level.

Ex situ conservation status and gaps for crop landrace groups. On average, ex situ conservation of crop landrace groups—measured in terms of the extent of current cultivated geographic range and ecological variation in the range that has previously been collected from and is now conserved in genebanks—was estimated to be moderately comprehensive at present, with substantial variation among crops; an average of 63% ± 12.6% of distributions was represented ex situ (Fig. 3 and Supplementary Table 1). Measured as the mean of the estimated minimum and maximum extent of representation in genebanks, geographic and ecological variation in landrace groups of the following crops was among the most comprehensively represented: breadfruit at 81.6% conserved, bananas and plantains at 81.5%, lentils at 78.3%, common beans at 77.4%, chickpeas at 75.8%, barley at 75.5% and bread wheat at 71.3%. Conversely, the largest conservation gaps persist for pearl millet at 32.7%, yams at 43.0%, finger millet at 45.4%, groundnut at 46.5%, potatoes at 50.3% and peas at 52.4%. The maximum potential representation metrics indicate that breadfruit, lentil, banana and plantain, grasspea and chickpea landrace group variation may already be very well conserved, since all have maximum current ex situ conservation scores above 90%, while the minimum coverage metrics warn that some crops may still face extensive conservation gaps, such as pearl millet at 15.2%, groundnut at 22.6%, finger millet at 25.3%, peas at 28.1% and yams at 29.0%.

Regarding types of assessed crops, the average degree of ex situ conservation of cereal, pulse and starchy root/tuber/fruit landrace groups did not differ significantly ($P = 0.69$), measuring 59.9%, 64.6% and 64.9%, respectively. At 45.0%, 45.6% and 50.4%, their mean minimum potential representation values were also similar, as were their maximum potential representation values of 74.8%, 83.6% and 79.3%. For the final crop-type category, this finding is remarkable because these plants are represented by lower overall numbers of genebank samples (an average of 1,052.7 accessions versus 2,827.4 of pulses and 5,796.4 of cereals); these typically clonally propagated crops often require higher ex situ conservation expenditures per sample and present more substantial challenges from pests and diseases[16]. Nonetheless, cereal, pulse and starchy root/tuber/fruit crop types all included members with some of the least and most comprehensive conservation scores (Fig. 3). Moreover, these scores were not correlated with importance of the crop to global food supplies, production and trade ($r = 0.064$)

At the landrace group level, considerable differences in current conservation status were identified among groups within many crops (Supplementary Table 2). For example, geographic and ecological variation in barley landraces with covered (with hull) grains was estimated to be 89.1% conserved, while diversity in landraces with naked or hull-less grains was only 31.3% represented ex situ. Asian rice, finger millet, potato, sorghum and yam landrace groups also varied rather widely regarding current conservation in genebanks, while cassava, chickpea, common bean, cowpea, groundnut, lentil, maize, pea, pearl millet, African rice, sweetpotato and bread wheat landrace groups had more similar within-crop ex situ representation estimates.

Taking sorghum landrace groups as an example, high-confidence gaps in current ex situ conservation in terms of geographic and ecological variation were identified for all five major races in sub-Saharan Africa, with overlapping gaps concentrated in Central, West and Southern Africa, including in Madagascar (Fig. 2b). The Supplementary Information provides conservation gap maps for all assessed crops.

Across the landrace groups of all 25 crops, geographic areas identified as hotspots requiring further collecting for ex situ conservation were concentrated in South Asia; the Mediterranean and West Asia; Mesoamerica; West, East and Southern Africa; the Andean

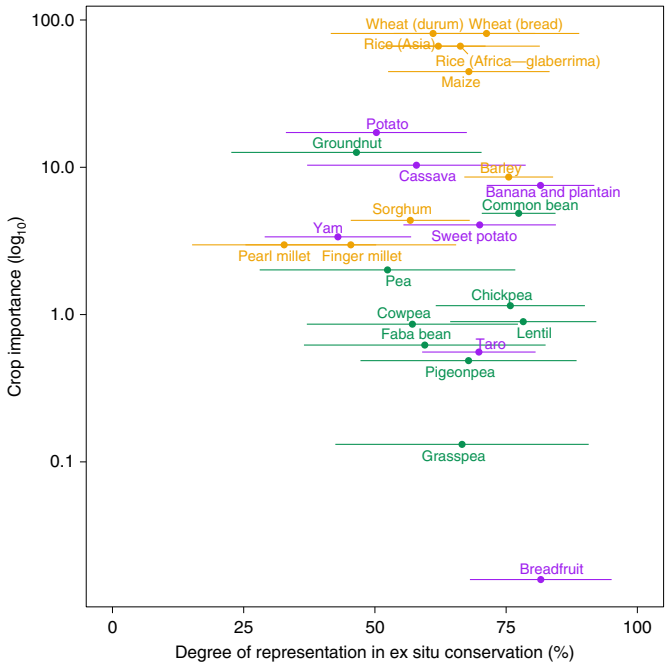


Fig. 3 | The current representation of crop landrace groups in ex situ conservation. Conservation metrics provide a scale from the lower to the upper estimates of current ex situ conservation status per crop with the averages denoted by circles. The crop importance metric indicates the current significance of the crop, averaged across global food supply, production and trade metrics (Supplementary Information). Gold, cereals; green, pulses; purple, starchy roots/tubers/fruits.

mountains; and Central and East Asia (Fig. 4 and Extended Data Fig. 5; online results at https://ciat.shinyapps.io/LGA_dashboard/). Currently, uncollected landrace groups of up to nine crops are potentially cultivated within single 2.5-arc-minute grid cells in India and Morocco and of up to eight crops in Algeria, Greece, Iran, Mexico, Pakistan, Sierra Leone and Turkey. Regarding types of assessed crops, cereal and pulse landrace group diversity was predicted to be particularly in need of further collecting in the Mediterranean and West Asia; South Asia; West, East and Southern Africa; Europe; the Andean mountains; and Mesoamerica. Conversely, starchy root/tuber/fruit crop landrace group ex situ conservation gaps were concentrated in East and Southeast Asia, South Asia, West Africa, South America and Mesoamerica (Extended Data Figs. 6–8).

Discussion

Our analysis of the ex situ conservation status of landrace groups within 25 staple crops suggests that their representation in genebanks is most often substantial, a finding that highlights the impact of extensive international, national and subnational efforts worldwide over more than a half-century, both individually and via collaborative networks and initiatives[1,7,17,18]. Conservation of landraces of these crops—or at least their eco-geographically distinguishable groups—appears to be considerably further advanced than equivalent protection for crop wild relatives (Extended Data Fig. 9)[19,20].

However, the findings also reveal that ex situ conservation gaps in terms of uncollected geographic and environmental variation across the distributions of landrace groups of these crops persist. Our quantitative and spatial results can aid in priority setting across these crops, their landrace groups and geographic regions, contributing to conservation targeting, planning and action. Further prioritization may be applied based on known or perceived threats related to economic, agricultural, technological, demographic, climatic and political change[1]. Recent decades of progress in clarifying and, in some cases, expediting the terms and conditions of genetic resources sampling and exchange[21,22] bolster anticipation that such gaps can be filled through international collaboration. With further concerted efforts to collect crop landraces of these and other crops, a high degree of representation of their diversity in genebanks appears to be feasible, and, thus, the fulfilment of the international targets of the CBD[11] and SDGs[12] regarding their ex situ conservation also seems achievable. Conducted periodically over time, the gap analysis offers a more holistic approach to assess the state of landrace conservation than simply reporting changes in counts of accessions held in genebanks[23] and, thus, may also represent a useful complement to the current indicators for these targets.

The landrace group classification and modelling processes described here demonstrate the potential to associate genetic, morphological, physiological, chemical, nomenclatural and other characteristics of cultivated plant populations with environmental and socioeconomic predictors within the regions of origin and diversity of crop taxa. These processes can be performed across a spectrum of infraspecific groups and geographic scales, depending on available knowledge and occurrence and characterization information. While our processes are based on openly available data and tools that undergo continual updating, they involve several limitations.

First, our methods are vulnerable to deficiencies in the quality, completeness and availability of occurrence and infraspecific grouping information. Many cultivated plants are insufficiently sampled, potentially due to a historical emphasis on wild rather than farming landscapes within biodiversity initiatives and persisting disconnects between biodiversity conservation and agricultural research communities[24]. Robust landrace classifications based on genetic structure, geography and other attributes also require further resolution for many crops.

The major biodiversity and conservation repository databases that we utilize here do not yet represent all pertinent national and subnational institutions worldwide; those institutions that do participate may not report all holdings and locality and characterization information is incomplete for many existing records[13,20]. Some additional information is probably present in other, smaller online databases or in offline or undigitized datasets. These gaps increase the uncertainty in our results, possibly leading to underestimations of the true degree of ex situ landrace conservation. On the other hand, the accessibility and long-term security of many such low-visibility collections are often equally uncertain[13,19]. Several processes would strengthen the conservation and potential usefulness of these genetic resources and the accuracy of analyses such as ours: the generation of characterization information and knowledge about infraspecific groups, improvements in the quality and completeness of existing occurrence information and better availability of landrace samples and their associated data, including safety duplication to better ensure long-term persistence.

Second, because our modelling method is based on statistical relationships between occurrences and environmental and socioeconomic predictor variables, it is also sensitive to the quality and comprehensiveness of these predictor datasets. Factors lacking predictor information or acting at finer scales than currently available data reflect will not be well incorporated into modelling processes. These may include environmental factors—both abiotic, such as soil characteristics or supplemental irrigation in small plots, and biotic, such as pathogen pressures or pollinator distributions—and socioeconomic drivers such as farm sizes, agronomic practices and seed system dynamics. Further, the models are unlikely to account for relatively recent disappearances of landraces unless such losses are associated with available predictors. The increasing generation of land-use-change information[25] may partially resolve this challenge. In all cases, further development of high-resolution predictor datasets with global scope will improve modelling. Deeper

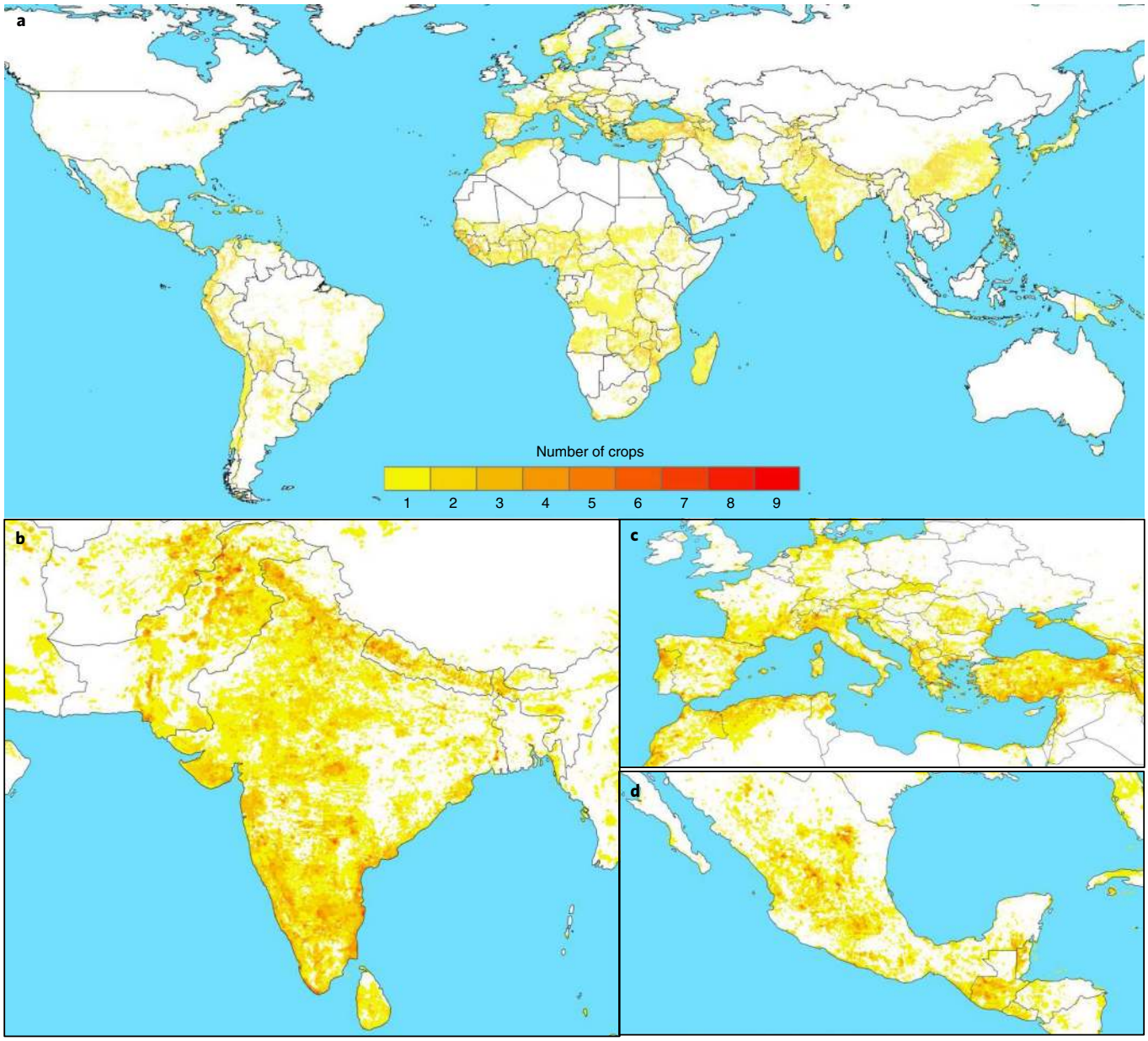


Fig. 4 | Geographic hotspots for further collection for the ex situ conservation of crop landrace groups. a, Global map of 'gap richness' across the predicted distributions of landrace groups of 25 cereal, pulse and starchy root/tuber/fruit crops within their geographic regions of diversity, indicating where landraces are expected to occur and have not yet been collected and conserved in genebanks. Darker colours indicate greater numbers of uncollected crop landrace groups potentially overlapping in the same 2.5-arc-minute cells, quantified in terms of numbers of crops. **b**–**d**, Examples of regions with particularly high gap richness in South Asia (**b**), the Mediterranean and West Asia (**c**) and Mesoamerica (**d**). See Extended Data Fig. 5 for gap richness across the 71 landrace groups within the 25 crops.

understanding of the wide range of factors affecting farmer choices regarding landrace cultivation, including apparent stochasticity[26], may also be important to improved modelling, while the limits to predicting distributions of populations whose ranges are driven by human preferences as much as environmental factors must be acknowledged.

Third, while geographic and ecological variation within predicted native ranges of plants has been shown to be an effective surrogate for direct measures of genetic diversity[27,28], the modelling and conservation metrics used here may not fully reflect the distributions of and gaps in genetic variation within crop landraces. Further, our standardized method may not take into account the differences between crop species in genetic diversity within and among their populations, which may be influenced by reproductive biology, such as by outcrossing versus inbreeding species and by the mode of pollination; by mode of propagation, such as by seed versus clonally; and by other ecological and cultural factors[3]. Moreover, our conservation gap analysis methodology is based on the assumption that the existence of an ex situ accession from a site indicates that the targeted landrace group has been adequately sampled there. In reality, landrace distinctions at finer resolution than their modelled groupings may be ignored and, thus, not fully conserved. Previous field collecting may also not have comprehensively sampled populations at the resolution needed for all conservation, plant breeding or other research aims. This drawback may be particularly applicable to landraces that are typically genetically

heterogeneous and, thus, may require large sample sizes to represent their diversity and, in particular, rare alleles. Finally, because in situ crop diversity constantly changes, developing novel variation from gene flow, recombination and mutation[6], valuable new forms may have arisen in previously collected areas. Further sampling for ex situ conservation may, therefore, be warranted within or near previously collected sites.

The combination of these vulnerabilities reinforces the importance of field reconnaissance and of partnering with Indigenous and traditional agrarian communities and associated organizations to inform further collecting activities. In this sense, our results are best considered as support tools, useful for guiding rather than prescribing taxonomic and geographic priorities[13]. Additional essential steps include ensuring adherence to international, national and local sampling and exchange policies[21,22]; assessing field work risks, particularly in regions affected by war and civil strife[29]; and determining the most appropriate timing to maximize the harvest of viable seeds and other propagules. The capacity of pertinent genebanks to receive, adequately maintain and distribute landrace diversity must be preconfirmed[1,7], and the logistics of getting collected material into relevant genebanks in a timely fashion must be established.

Further development and mobilization of landrace modelling and conservation gap analysis would ideally assess a wider range of crops, including fruits and vegetables, nuts and other groups of importance to human nutrition and agricultural livelihoods[30]. It is probable that many other crops, especially those that have not received primary focus in international or national genetic resources conservation and crop improvement efforts over the past half-century, are less well represented in ex situ conservation repositories[1,10]; thankfully, erosion of the in situ genetic diversity of these crops may be less severe thus far than in major staples[1].

Geographic expansion of the analyses beyond historical regions of diversity[9,14,15] may also aid in identifying novel variation, although further assessment of the correlation between landrace groups and spatial predictors in such regions will be necessary. To more fully address the scope of international conservation targets for landraces, these analyses must also assess the state of their in situ (on-farm) conservation; this task presents substantial challenges because emphasis in this context falls on the conditions and processes that foster landrace diversity rather than on the persistence of particular populations[1,2,31]. Given further development and expansion of the methods and scope, and the combination of the results with parallel analyses of crop wild relatives[19,20] and other socioeconomically and culturally valuable plants[32], a significantly improved understanding of distributions, protection status and conservation gaps across the major forms of crop diversity prioritized by the CBD and the SDGs should be achievable.

Methods

Crops and their landrace study areas. Food crops whose genetic resources are researched and conserved by CGIAR international agricultural research centres or by the CePaCT of the SPC were included in this study. Crop landrace distributions were modelled and conservation analyses conducted within recognized primary and, for some crops, secondary regions of diversity, where these crops were domesticated and/or have been cultivated for very long periods, and where they are, thus, expected to feature high genetic diversity and adaptation to local environmental and cultural factors (Supplementary Tables 1 and 2)[9,13]. These regions were identified through literature review (Supplementary Information) and confirmed by crop experts.

Occurrence data. Our crop landrace group distribution modelling and conservation gap analysis rely on occurrence data, including coordinates of locations where landraces were previously collected for ex situ conservation and reference sightings. For ex situ conservation records, occurrences marked as landraces were retrieved from two major online databases: the Genesys Plant Genetic Resources portal[33] and the World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture (WIEWS) of the Food and Agriculture Organization of the United Nations[34]. Occurrences were also obtained directly from individual international genebank information systems: AfricaRice, the International Transit Centre and Musa Germplasm Information System of Bioversity International[35], CePaCT, International Center for Tropical Agriculture (CIAT), International Maize and Wheat Improvement Center (CIMMYT), International Potato Center (CIP), International Center for Agricultural Research in the Dry Areas (ICARDA), International Crops Research Institute for the Semi-arid Tropics (ICRISAT), International Institute of Tropical Agriculture (IITA) and International Rice Research Institute (IRRI), as well as from the United States Department of Agriculture (USDA) Genetic Resources Information Network (GRIN)–Global[36] and the Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO)[37]. Occurrences were compiled from the Global Biodiversity Information Facility (GBIF), with 'living specimen' records classified as ex situ conservation records and the remaining serving as reference sightings for use in distribution modelling. Reference occurrences were also drawn from published literature (Supplementary Information). Duplicated observations within or between data sources were eliminated, with a preference to utilize the most original data. Coordinates were corrected or removed when latitude and longitude were equal to zero or inverted, located in water bodies or in the wrong country or had poor resolution (<2 decimal places). Occurrences were clipped to study areas per crop. The complete occurrence dataset is available in Supplementary Dataset 2.

Spatial predictors. We compiled and calculated spatially explicit gridded information for 50 potential environmental and cultural predictors of landrace distributions, including climatic, topographic, evolutionary history and socioeconomic variables (Supplementary Table 3)[13]. For climate data, we gathered or derived 39 variables from WorldClim version 2 (ref. [38]) and Environmental Rasters for Ecological Modeling (ENVIREM)[39]. We included elevation from the Shuttle Radar Topography Mission (SRTM) dataset of the CGIAR–Consortium on Geospatial Information portal[40,41]. Two crop evolutionary history proxies were included: distance to human settlements before the year AD 1500 (ref. [42]) and distance to known wild progenitor populations[13]. The eight socioeconomic variables included population density[43], distance to navigable rivers[44], percentage of the area under irrigation[45], population accessibility[46,47], geographic distributions of ethnic or cultural groups[48] and crop harvested area, production quantity and yield[49]. All predictor data were scaled to 2.5-arc-minute resolution with World Geodetic System (WGS) 84 as a datum. Extended descriptions of the sources and their justification for inclusion are provided in Ramirez-Villegas et al.[13]. The complete spatial predictor dataset is available in Supplementary Dataset 3.

Crop landrace group classifications. Crop landraces are cultivated plant populations managed by Indigenous or traditional farmers through cultivation, selection and diffusion[1]. They are typically genetically heterogeneous, although some types, such as clonally propagated populations, may be relatively homogeneous. They have recognizable characteristics, identities and geographic origins are in an ongoing process of adaptation to their local environments and societal conditions[1,2,31]. For most crops, landraces number in the thousands, with major global staple cereals such as rice and wheat potentially represented by hundreds of thousands of landraces[50,51], although precise numbers and consensus regarding differentiations among landraces within crops have not been established. Given the diversity of landraces and the complexity of environmental and cultural drivers differentiating them, our method seeks a compromise between, on the one hand, acknowledgement of this diversity and, on the other, the feasibility and performance of distribution modelling and conservation gap analysis.

For each crop, we, therefore, conducted an extensive literature review to identify recognized infraspecific groups with distinct morphological, physiological, chemical, genetic, nomenclatural or other characteristics that could be tested for environmental and cultural associations (Supplementary Table 1 and Supplementary Information). The nature of these groups varied by crop and included genepools, races, genetic clusters and geographic or environmental groupings. Crops often had more than one proposed grouping or classification.

We then built and tested classification models to determine how well the proposed groups could be predicted and distinguished based on spatial predictors, drawing from the occurrence database and training datasets compiled from the literature review. A random forest[52], a support vector machine[53], the *K*-nearest neighbour (KNN) algorithm[54] and artificial neural networks[55] were used to determine classification performance. The response variable was the group to which a given occurrence was assigned, whereas the explanatory variables were the spatial predictors. Models were combined into an ensemble using the mode—that is, the most frequent predicted value among the models—and tested using 15-fold cross-validation with 80% training and 20% testing. We accepted a given classification if each of its components was predicted with an average cross-validated accuracy of at least 80%. In the case of multiple classification proposals per crop, we selected the one with the best overall performance. Finally, we used the trained models to predict the corresponding group for occurrences missing such information. All landrace groups for all crops are provided in Supplementary Table 2, with the best-performing groups identified.

Crop landrace group distribution modelling. To predict the probability of geographic occurrence for each landrace group within each crop, we generated MaxEnt models[56,57] using the 'maxnet' R package[58]. Group-specific spatial

predictors were selected using a combination of the variance inflation factor (VIF) and a principal component analysis (PCA) to control for excessive model complexity and variable collinearity[59]. We removed variables that did not contribute significantly to the variance in the PCA, defined as contributing less than 15% to the first component, and we further discarded variables with a VIF > 10 (ref. [60]). The predictors and whether they were selected for the modelling of each landrace group are presented in Supplementary Table 4.

We generated a random sample of pseudo-absences as background points in areas that (1) were within the same ecological land units[61] as the occurrence points, (2) were deemed potentially suitable according to a support vector machine classifier that uses all occurrences and predictor variables and (3) were farther than 5 km from any occurrence[62]. The number of pseudo-absences generated per crop group was ten times its number of unique occurrences.

MaxEnt models were fitted through five-fold ($K = 5$) cross-validation with 80% training and 20% testing. For each fold, we calculated the area under the receiving operating characteristic curve (AUC), sensitivity, specificity and Cohen's kappa as measures of model performance. To create a single prediction that represents the probability of occurrence for the landrace group, we computed the median across K models. Geographic areas in the form of pixels with probability values above the maximum sum of sensitivity and specificity were treated as the final area of predicted presence[13].

Ex situ conservation status and gaps. Three separate but complementary metrics were developed to compare the geographic and environmental diversity in current ex situ conservation collections to the total geographic and environmental variation across the crop landrace group distribution model and, thus, to identify and quantify ex situ conservation gaps[13].

A connectivity gap score (S_{CON}) was calculated for each 2.5-arc-minute pixel within the distribution model by drawing a triangle[63,64] around each pixel using the three closest genebank accession occurrence locations as vertices and then deriving normalized values for the pixel based on distance to the triangle centroid and vertices[13]. The S_{CON} of a pixel is high—closer to 1 on a scale of 0–1—when its corresponding triangle is large, when the pixel is close to the centroid of the triangle or when the distance to the vertices is large. A high S_{CON} represents a greater probability of the pixel location being a gap in existing ex situ collections.

An accessibility gap score (S_{ACC}) was calculated for each 2.5-arc-minute pixel in the distribution model by computing travel time from each pixel to its nearest genebank accession occurrence location based both on distance and the speed of travel, defined by a friction surface[13,45]. Travel time scores were normalized by dividing pixel values by the longest travel time within the distribution model, with the final score ranging from 0 to 1. A high S_{ACC} value for a pixel reflects long travel times from existing genebank collection occurrences and, thus, represents a higher probability of the pixel location being a gap in existing ex situ collections.

An environmental gap score (S_{ENV}) was calculated for each 2.5-arc-minute pixel in the distribution model by conducting a hierarchical clustering analysis using Ward's method with all the predictor variables from the distribution modelling. The Mahalanobis distance between each pixel and the environmentally closest genebank accession occurrence location was then computed[13]. Environmental distance scores were normalized between 0 and 1. A high S_{ENV} value for a pixel reflects a large distance to areas with similar environments where landraces have previously been collected for genebank conservation and, thus, represents a higher probability of the pixel location being a gap in existing ex situ collections.

Spatial ex situ conservation gaps were determined from the conservation gap scores using a cross-validation procedure to derive a threshold for each score. We created synthetic gaps by removing existing genebank occurrences in five randomly chosen circular areas with a 100 km radius within the distribution model. We then tested whether these artificial gaps could be predicted by our gap analysis, identifying the threshold value of each score that would maximize the prediction of these synthetic gaps. Performance for each of the five gap areas was assessed using AUC, sensitivity and specificity. The average cross-area threshold value was calculated for each score to discern pixels with a high likelihood of finding ex situ conservation gaps and that, thus, were higher priority for further field sampling. These were pixels with combined gap scores above the threshold, assigned a value of 1, as opposed to the relatively well-conserved areas below the threshold, which were assigned a value of 0.

The three binary conservation gap scores were then mapped in combination, resulting in pixels across the distribution model with gap values ranging from 0 to 3. Pixels with a value of 0 display no connectivity, accessibility or environmental gaps and are considered well represented ex situ. Pixels with a value of 1 indicate a conservation gap in connectivity, accessibility or the environment; we consider these 'low-confidence' gaps. Pixels with a value of 2 indicate gaps in two metrics or 'medium-confidence' gaps, and values of 3 indicate gaps across all metrics or 'high-confidence' gaps. High-confidence gap areas are displayed on crop-conservation-gap maps (Fig. 2b and Supplementary Information) and conservation hotspot maps across crops (Fig. 4 and Extended Data Figs. 5–8).

The representation of crop landrace groups in current ex situ conservation collections was calculated based on the final 1–3 value conservation-gap maps. The complement of the proportion of the modelled distribution considered as a potential conservation gap by any single gap score represents the minimum estimate of current representation; the complement of the proportion considered by all three scores as a gap, which is to say high-confidence gap areas, represents the maximum estimate (Supplementary Tables 1 and 2).

While distribution modelling and conservation gap analyses were conducted at the crop landrace group level and results are presented in full in the Supplementary Information, for ease of comparison of results across crops, and to avoid bias towards crops with many landrace groups, we also calculated summary results at the crop level. Crops that had been assessed with geographic differentiations, including maize in Africa and Latin America and yams in the New World and the Old World, were also combined. For spatial results, the pixels in crop landrace group models were summed—that is, constituent landrace group models were combined. The minimum and maximum current conservation representation estimations at the crop level were then calculated based on combined spatial models.

GBIF occurrence downloads. The following occurrence downloads from the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/, 2017–2021) were used: 10.15468/dl.rrntfr, 10.15468/dl.2f2v4h, 10.15468/dl.2ywlb7, 10.15468/dl.lnfelh, 10.15468/dl.ryrmfj, 10.15468/dl.8adf61, 10.15468/dl.nff5ys, 10.15468/dl.erxs6e, 10.15468/dl.vbfgho, 10.15468/dl.mjjk3x, 10.15468/dl.uppz1n, 10.15468/dl.938bgm, 10.15468/dl.hr87hm, 10.15468/dl.k1va80, 10.15468/dl.coqpu2, 10.15468/dl.lkoo9u, 10.15468/dl.e998mp, 10.15468/dl.vfbmm7, 10.15468/dl.tnp478, 10.15468/dl.6zxsea, 10.15468/dl.0lray8, 10.15468/dl.5sjgsw, 10.15468/dl.wkju6h, 10.15468/dl.7xzfvc, 10.15468/dl.autlf5, 10.15468/dl.fe2amw, 10.15468/dl.2zblvz, 10.15468/dl.ddplkj, 10.15468/dl.jbzejg, 10.15468/dl.ej5bha, 10.15468/dl.905pxd, 10.15468/dl.pim1vs, 10.15468/dl.vdridc, 10.15468/dl.b43gyv, 10.15468/dl.nnw3z7, 10.15468/dl.bnt9jc, 10.15468/dl.f5x2cg, 10.15468/dl.ub7zbg, 10.15468/dl.sggf2v, 10.15468/dl.ath5ve, 10.15468/dl.23k3ug, 10.15468/dl.cym376, 10.15468/dl.53bwzk, 10.15468/dl.fsad7h and 10.15468/dl.fm6p7z.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Occurrence data, including spatial predictor variable results (at 2.5-arc-minute resolution) for each occurrence (Supplementary Dataset 2) and the global spatial predictor dataset (2.5-arc-minute resolution, all 50 variables) (Supplementary Dataset 3) are available at https://doi.org/10.7910/DVN/J8WAPH. Source data are provided with this paper.

Code availability

Code for the crop landrace group classification testing, distribution modelling and conservation gap analysis is available at https://github.com/CIAT-DAPA/gap_analysis_landraces.

Received: 13 October 2021; Accepted: 28 March 2022;
Published online: 9 May 2022

References

1. Khoury, C. K. et al. Crop genetic erosion: understanding and responding to loss of crop diversity. *New Phytol.* **233**, 84–118 (2021).
2. Jarvis, D. I. et al. A global perspective of the richness and evenness of traditional crop-variety diversity maintained by farming communities. *Proc. Natl Acad. Sci. USA* **105**, 5326–5331 (2008).
3. Allinne, C. et al. Role of seed flow on the pattern and dynamics of pearl millet (*Pennisetum glaucum* [L.] R. Br.) genetic diversity assessed by AFLP markers: a study in south-western Niger. *Genetica* **133**, 167–178 (2007).
4. Rojas-Barrera, I. C. et al. Contemporary evolution of maize landraces and their wild relatives influenced by gene flow with modern maize varieties. *Proc. Natl Acad. Sci. USA* **116**, 21302–21311 (2019).
5. Jarvis, D. I. & Hodgkin, T. Wild relatives and crop cultivars: detecting natural introgression and farmer selection of new genetic combinations in agroecosystems. *Mol. Ecol.* **8**, S159–S173 (1999).
6. Mercer, K. L. & Perales, H. R. Evolutionary response of landraces to climate change in centers of crop diversity. *Evol. Appl.* **3**, 480–493 (2010).
7. Gepts, P. Plant genetic resources conservation and utilization: the accomplishments and future of a societal insurance policy. *Crop Sci.* **46**, 2278–2292 (2006).
8. Meyer, R. S., DuVal, A. E. & Jensen, H. R. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.* **196**, 29–48 (2012).
9. Khoury, C. K. et al. Origins of food crops connect countries worldwide. *Proc. R. Soc. B* **283**, 20160792 (2016).
10. *Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture* (Food and Agriculture Organization of the United Nations, 2010); https://www.fao.org/3/i1500e/i1500e.pdf
11. Convention on Biological Diversity. *Strategic Plan for Biodiversity 2011-2020, including Aichi Biodiversity Targets* (Convention on Biological Diversity, 2010); https://www.cbd.int/doc/strategic-plan/2011-2020/Aichi-Targets-EN.pdf

12. Sustainable development goals. *United Nations* https://sdgs.un.org/2030agenda (2015).
13. Ramirez-Villegas, J. et al. A gap analysis modelling framework to prioritize collecting for ex situ conservation of crop landraces. *Divers. Distrib.* **26**, 730–742 (2020).
14. Vavilov, N. I. Tzentry proiskhozhdeniya kulturnykh rastenii (The centres of origin of cultivated plants). *Works Appl. Bot. Plant Breed.* **16**, 1–248 (1926).
15. Ladizinsky, G. *Plant Evolution under Domestication* (Kluwer Academic, 1998).
16. Halewood, M. et al. Germplasm acquisition and distribution by CGIAR genebanks. *Plants* **9**, 1296 (2020).
17. Plucknett, D. L., Smith N. J. H., Williams, J. T. & Murthi-Anishetty, N. *Gene Banks and the World's Food* (Princeton Univ. Press, 1987).
18. Thormann, I., Engels, J. M. M. & Halewood, M. Are the old International Board for Plant Genetic Resources (IBPGR) base collections available through the Plant Treaty's multilateral system of access and benefit sharing? A review. *Genet. Resour. Crop Evol.* **66**, 291–310 (2019).
19. Castañeda-Álvarez, N. P. et al. Global conservation priorities for crop wild relatives. *Nat. Plants* **2**, 16022 (2016).
20. Khoury, C. K. et al. Crop wild relatives of the United States require urgent conservation action. *Proc. Natl Acad. Sci. USA* **117**, 33351–33357 (2020).
21. *The International Treaty on Plant Genetic Resources for Food and Agriculture* (Food and Agriculture Organization of the United Nations, 2002).
22. *Nagoya Protocol on Access and Benefit-sharing* (Convention on Biological Diversity, 2011).
23. SDG Indicators, Metadata Repository, Goal 2. End hunger, achieve food security and improved nutrition and promote sustainable agriculture. *United Nations* https://unstats.un.org/sdgs/metadata/?Text=&Goal=2&Target (2021).
24. Scherr, S. J. & McNeely, J. A. Biodiversity conservation and agricultural sustainability: towards a new paradigm of 'ecoagriculture' landscapes. *Phil. Trans. R. Soc. B* **363**, 477–494 (2008).
25. Winkler, K., Fuchs, R., Rounsevell, M. & Herold, M. Global land use changes are four times greater than previously estimated. *Nat. Commun.* **12**, 2501 (2021).
26. Zeven, A. C. The traditional inexplicable replacement of seed and seed ware of landraces and cultivars: a review. *Euphytica* **110**, 181–191 (1999).
27. Hanson, J. O., Rhodes, J. R., Riginos, C. & Fuller, R. A. Environmental and geographic variables are effective surrogates for genetic variation in conservation planning. *Proc. Natl Acad. Sci. USA* **114**, 12755–12760 (2017).
28. Hoban, S., Kallow, S. & Trivedi, C. Implementing a new approach to effective conservation of genetic diversity, with ash (*Fraxinus excelsior*) in the UK as a case study. *Biol. Conserv.* **225**, 10–21 (2018).
29. Sperling, L. The effect of the civil war on Rwandas bean seed systems and unusual bean diversity. *Biodivers. Conserv.* **10**, 989–1010 (2001).
30. Willett, W. et al. Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. *Lancet* **393**, 447–492 (2019).
31. Brush, S. B. In situ conservation of landraces in centers of crop diversity. *Crop Sci.* **35**, 346 (1995).
32. Khoury, C. K. et al. Comprehensiveness of conservation of useful wild plants: an operational indicator for biodiversity and sustainable development targets. *Ecol. Indic.* **98**, 420–429 (2019).
33. Genesys-PGR: a gateway to genetic resources. *Global Crop Diversity Trust* https://www.genesys-pgr.org/ (2021).
34. United Nations Food and Agriculture Organization World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture. *Food and Agricultural Organization of the United Nations* http://www.fao.org/wiews/en/ (2021).
35. Ruas, M. et al. MGIS: managing banana (*Musa spp.*) genetic resources information and high-throughput genotyping data *Database* https://doi.org/10.1093/database/bax046 (2017).
36. National Plant Germplasm System, GRIN-Global Accessions. *US Department of Agriculture Agricultural Research Service* https://npgsweb.ars-grin.gov/gringlobal/search (2021).
37. Portal de Geoinformacion. *Comisión Nacional para el Conocimiento y Uso de la Biodiversidad* http://www.conabio.gob.mx/informacion/gis/ (2021).
38. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
39. Title, P. O. & Bemmels, J. B. ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography* **41**, 291–307 (2018).
40. Reuter, H. I., Nelson, A. & Jarvis, A. An evaluation of void-filling interpolation methods for SRTM data. *Int. J. Geogr. Inf. Sci.* **21**, 983–1008 (2007).
41. Jarvis, A., Reuter, H. I., Nelson, A., & Guevara, E. *Hole-filled Seamless SRTM Data V4* (International Center for Tropical Agriculture, 2008); https://cgiarcsi.community/data/srtm-90m-digital-elevation-database-v4-1/
42. Reba, M., Reitsma, F. & Seto, K. C. Spatializing 6,000 years of global urbanization from 3700 BC to AD 2000. *Sci. Data* **3**, 160034 (2016).
43. *Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11* (Center for International Earth Science Information Network, 2018); https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11
44. Rivers + lake centerlines. *Natural Earth* https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-rivers-lake-centerlines/ (2019).
45. Siebert, S., Henrich, V., Frenken, K., & Burke, J. *Update of the Global Map of Irrigation Areas to Version 5: Project Report* (Food and Agriculture Organization of the United Nations, 2013).
46. Weiss, D. J. et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336 (2018).
47. Travel time to major cities. *Publications Office of the European Commission Joint Research Centre—Institute for Environment and Sustainability* https://op.europa.eu/en/publication-detail/-/publication/20a3a771-15b3-45ac-9606-7575b9df740a/language-en (2008).
48. Weidmann, N. B., Rød, J. K. & Cederman, L.-E. Representing ethnic groups in space: a new dataset. *J. Peace Res.* **47**, 491–499 (2010).
49. You, L. et al. *Spatial Production Allocation Model (SPAM) 2005 v3.2* (MapSPAM, 2019). https://www.mapspam.info/
50. Harlan, J. *Crops and Man* (American Society of Agronomy, 1975).
51. Jones, H. et al. Approaches and constraints of using existing landrace and extant plant material to understand agricultural spread in prehistory. *Plant Genet. Resour. Charact. Util.* **6**, 98–112 (2008).
52. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **26**, 217–222 (2005).
53. Meyer, D., Leisch, F. & Hornik, K. The support vector machine under test. *Neurocomputing* **55**, 169–186 (2003).
54. Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* Vol. 2888 (eds Meersman, R. et al.) 986–996 (Springer, 2003).
55. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).
56. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**, 231–259 (2006).
57. Elith, J. et al. A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57 (2011).
58. Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. & Blair, M. E. Opening the black box: an open-source release of Maxent. *Ecography* **40**, 887–893 (2017).
59. Warren, D. L. & Seifert, S. N. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecol. Appl.* **21**, 335–342 (2011).
60. Braunisch, V. et al. Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography* **36**, 971–983 (2013).
61. Sayre, R. et al. *A New Map of Global Ecological Land Units—An Ecophysiographic Stratification Approach* (Association of American Geographers, 2014).
62. Senay, S. D., Worner, S. P. & Ikeda, T. Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PLoS ONE* **8**, e71218 (2013).
63. Lee, D. T. & Schachter, B. J. Two algorithms for constructing a Delaunay triangulation. *Int. J. Comput. Inf. Sci.* **9**, 219–242 (1980).
64. Turner, R. deldir: Delaunay triangulation and Dirichlet (Voronoi) tessellation. R package version 0.1-16. https://cran.r-project.org/web/packages/deldir/deldir.pdf (2021).

Acknowledgements

Support was provided by the CGIAR Genebank Platform (J.R.-V., C.K.K., H.A., M.V.D., A.M., C.C.S., Z.K., L.G., M.A., J.A., B.A.A., J.C.A., A.A., N.L.A., V.A., K.A., G.L.C., O.C., D.C., D.E.C., D.G.D., D.E., H.F., A.F., M.E.G., P.G., A.J.G., B.G., A.I.E.H., R.J., C.S.J., B.K., J.-S.L., K.L.M., A.M., M.-N.N., O.O., T.S.P., S.R., G.R., N.R., M.R., C.S., J.S., T.D.S., M.T., I.v.d.H., J.A.V., R.V., P.W., M.Y. and C.Z.) and by grant number 2019-67012-29733/project accession number 1019405 from the USDA National Institute of Food and Agriculture (C.K.K.).

Author contributions

J.R.-V., C.K.K., H.A., M.V.D., C.C.S., Z.K. and L.G. conceived and designed the study. J.R.-V., C.K.K., H.A., M.V.D., A.M., C.C.S., Z.K., L.G., M.A., J.A., B.A.A., J.C.A., A.A., N.L.A., V.A., K.A., G.L.C., O.C., D.C., D.E.C., D.G.D., D.E., H.F., A.F., M.E.G., P.G., A.J.G., B.G., A.I.E.H., R.J., C.S.J., B.K., J.-S.L., K.L.M., A.M., M.-N.N., O.O., T.S.P., S.R., G.R., N.R., M.R., C.S., J.S., T.D.S., M.T., I.v.d.H., J.A.V., R.V., P.W., M.Y. and C.Z. contributed data and conceptual inputs. J.R.-V., C.K.K., H.A., M.V.D., A.M., C.C.S., Z.K., K.A., O.C., A.I.E.H. and S.R. compiled and processed data and wrote and ran the code. J.R.-V., C.K.K., H.A., M.V.D., A.M., C.C.S., Z.K. and L.G. interpreted the results. J.R.-V., C.K.K., M.V.D. and A.M. wrote the paper. J.R.-V., C.K.K., H.A., M.V.D., A.M., C.C.S., Z.K., L.G., M.A., J.A., B.A.A., J.C.A., A.A., N.L.A., V.A., K.A., G.L.C., O.C., D.C., D.E.C., D.G.D., D.E., H.F., A.F., M.E.G., P.G., A.J.G., B.G., A.I.E.H., R.J., C.S.J., B.K., J.-S.L., K.L.M., A.M., M.-N.N., O.O., T.S.P., S.R., G.R., N.R., M.R., C.S., J.S., T.D.S., M.T., I.v.d.H., J.A.V., R.V., P.W., M.Y. and C.Z. edited the paper.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41477-022-01144-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-022-01144-8.

Correspondence and requests for materials should be addressed to Julian Ramirez-Villegas or Colin K. Khoury.

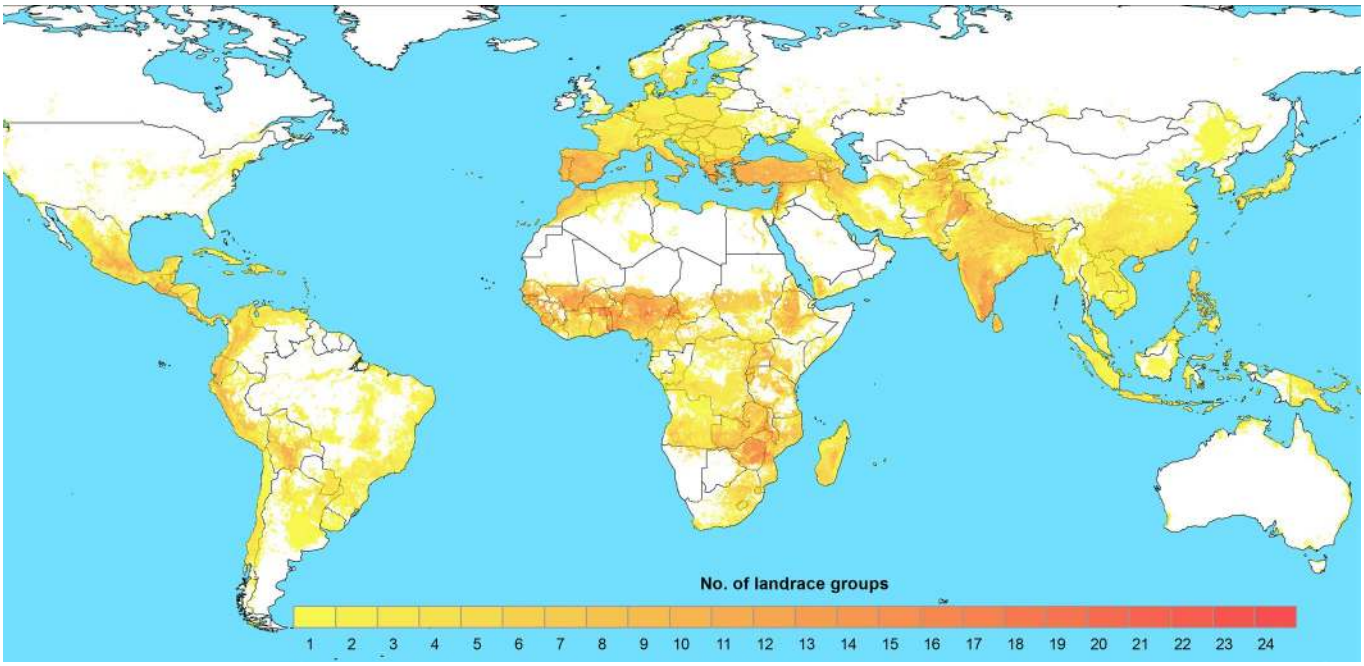
Peer review information *Nature Plants* thanks James Borell, Gayle Volk and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

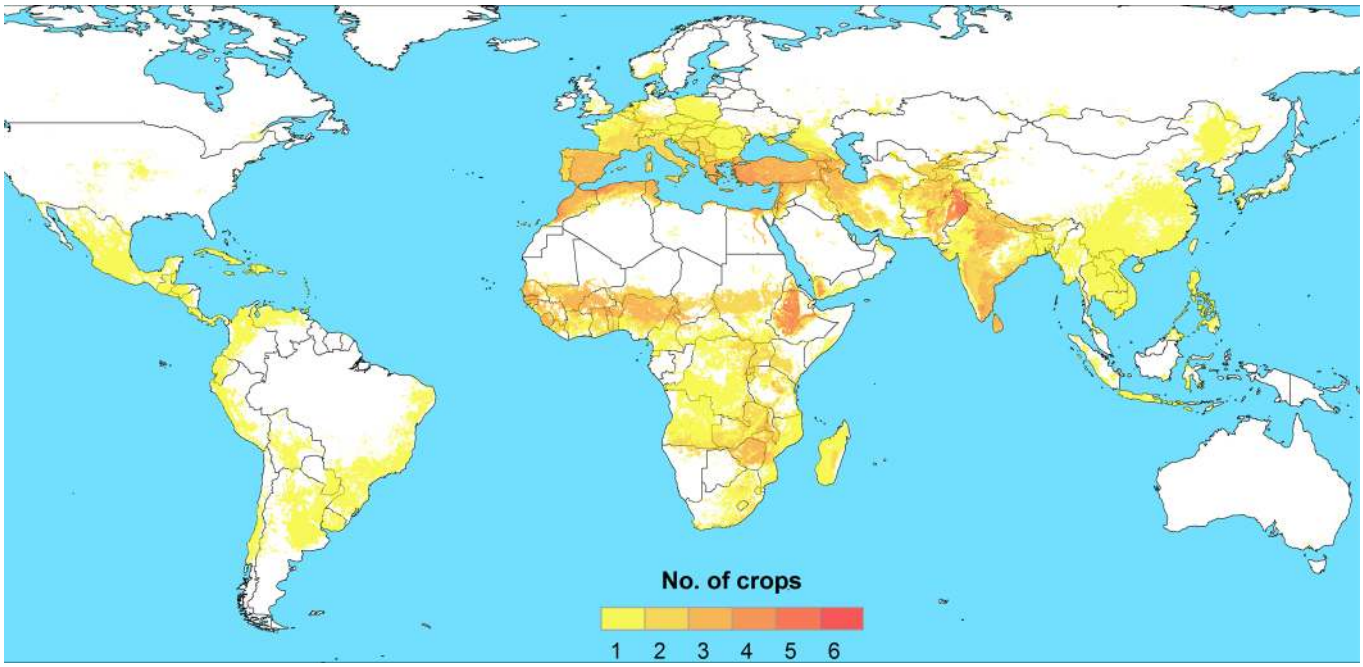
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit http://creativecommons.org/licenses/by/4.0/.

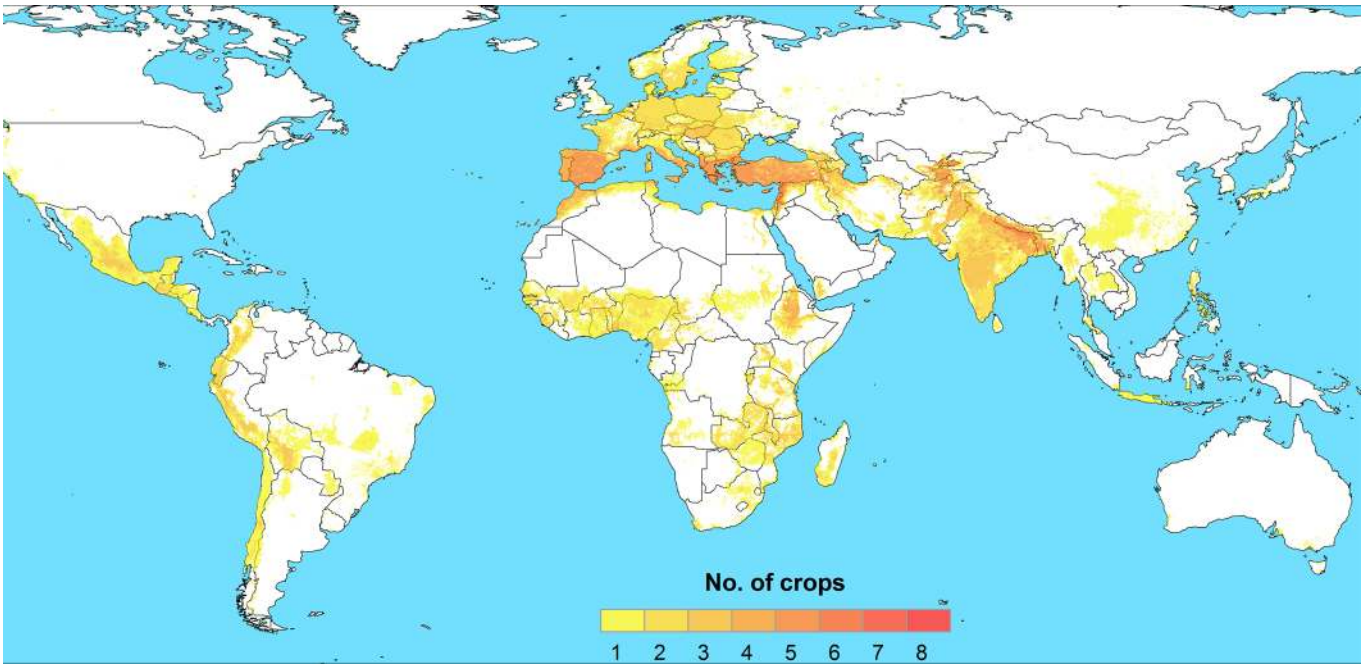
© The Author(s) 2022



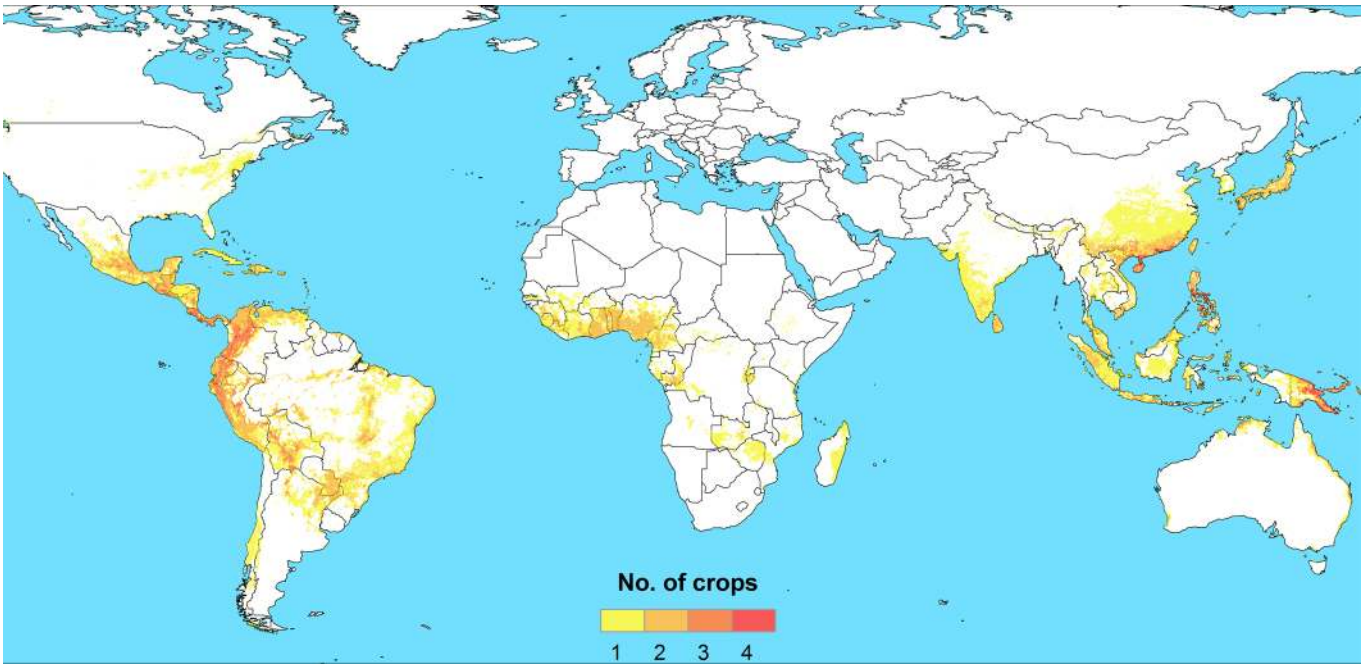
Extended Data Fig. 1 | Richness map of the predicted distributions of 71 landrace groups of 25 cereal, pulse, and starchy root/tuber/fruit crops within their geographic regions of diversity. Richness map of the predicted distributions of 71 landrace groups of 25 cereal, pulse, and starchy root/tuber/fruit crops within their geographic regions of diversity. Darker colors indicate greater numbers of crop landrace groups potentially overlapping in the same 2.5 arc-minute cells.



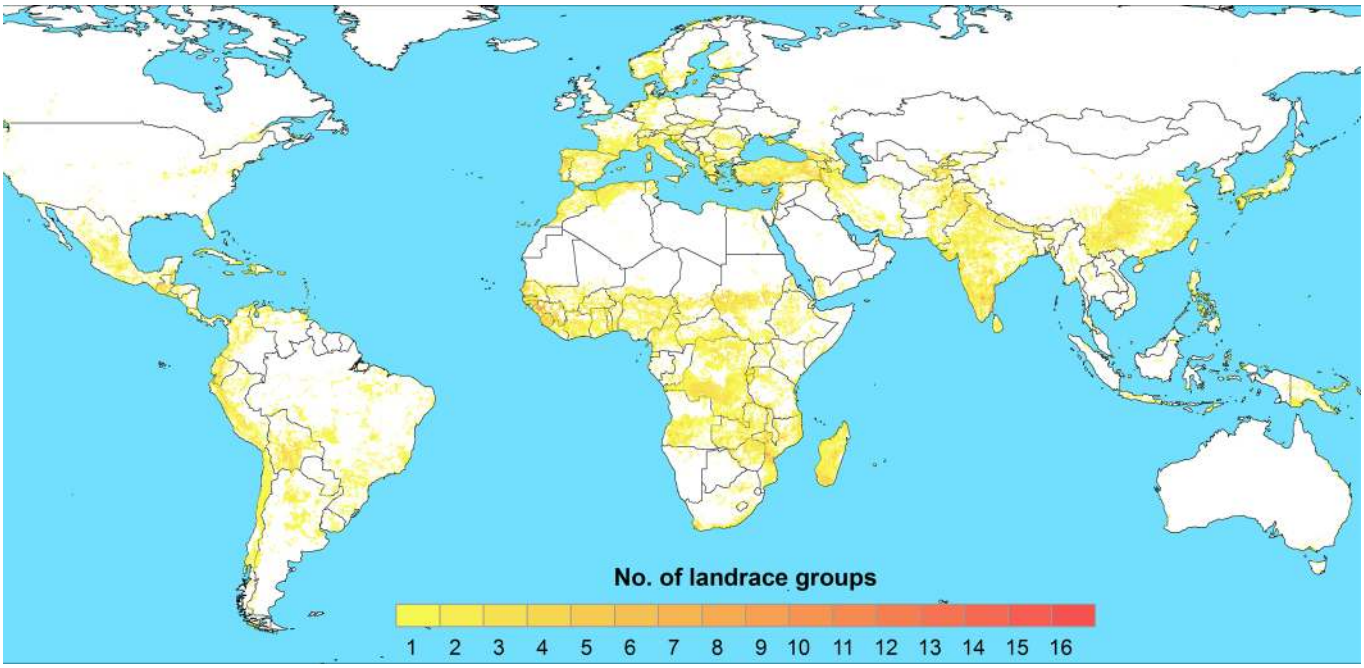
Extended Data Fig. 2 | Richness map of the predicted distributions of landrace groups of 9 cereal crops within their geographic regions of diversity. Richness map of the predicted distributions of landrace groups of 9 cereal crops within their geographic regions of diversity. Darker colors indicate greater numbers of crop landraces potentially overlapping in the same 2.5 arc-minute cells, quantified in terms of numbers of crops.



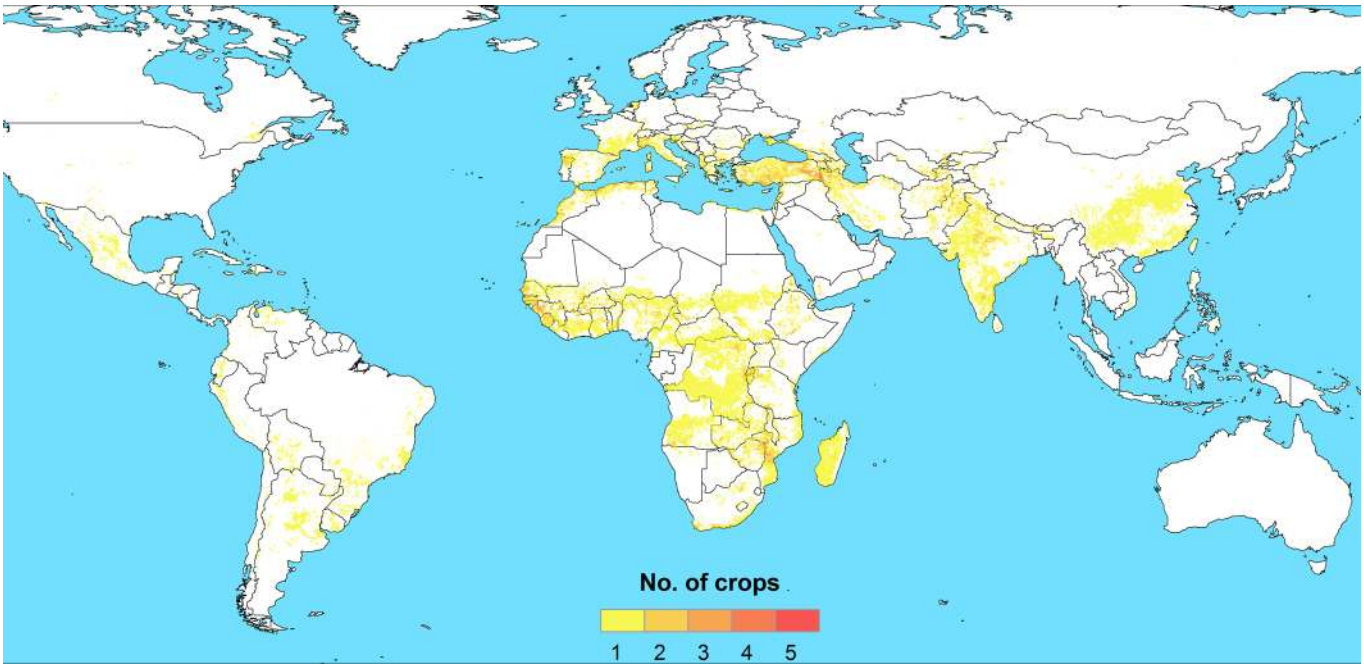
Extended Data Fig. 3 | Richness map of the predicted distributions of landrace groups of 9 pulse crops within their geographic regions of diversity. Richness map of the predicted distributions of landrace groups of 9 pulse crops within their geographic regions of diversity. Darker colors indicate greater numbers of crop landraces potentially overlapping in the same 2.5 arc-minute cells, quantified in terms of numbers of crops.



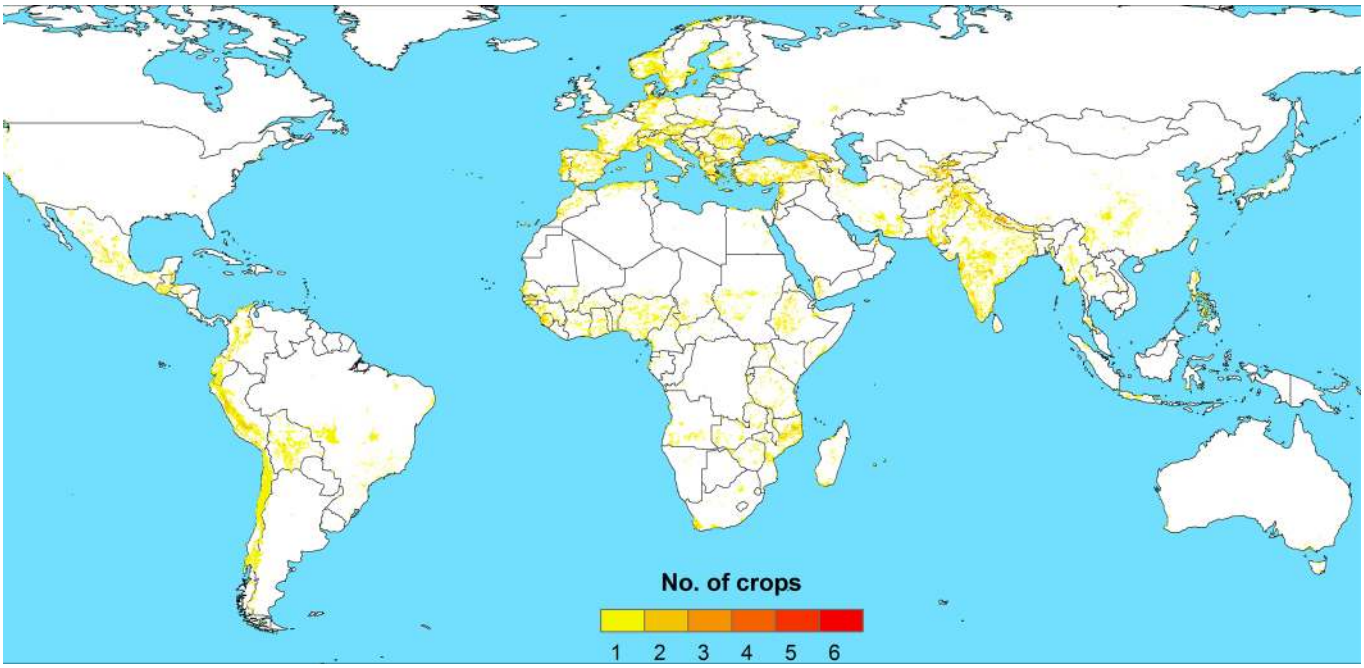
Extended Data Fig. 4 | Richness map of the predicted distributions of landrace groups of 7 starchy root, tuber, and fruit crops within their geographic regions of diversity. Richness map of the predicted distributions of landrace groups of 7 starchy root, tuber, and fruit crops within their geographic regions of diversity. Darker colors indicate greater numbers of crop landraces potentially overlapping in the same 2.5 arc-minute cells, quantified in terms of numbers of crops.



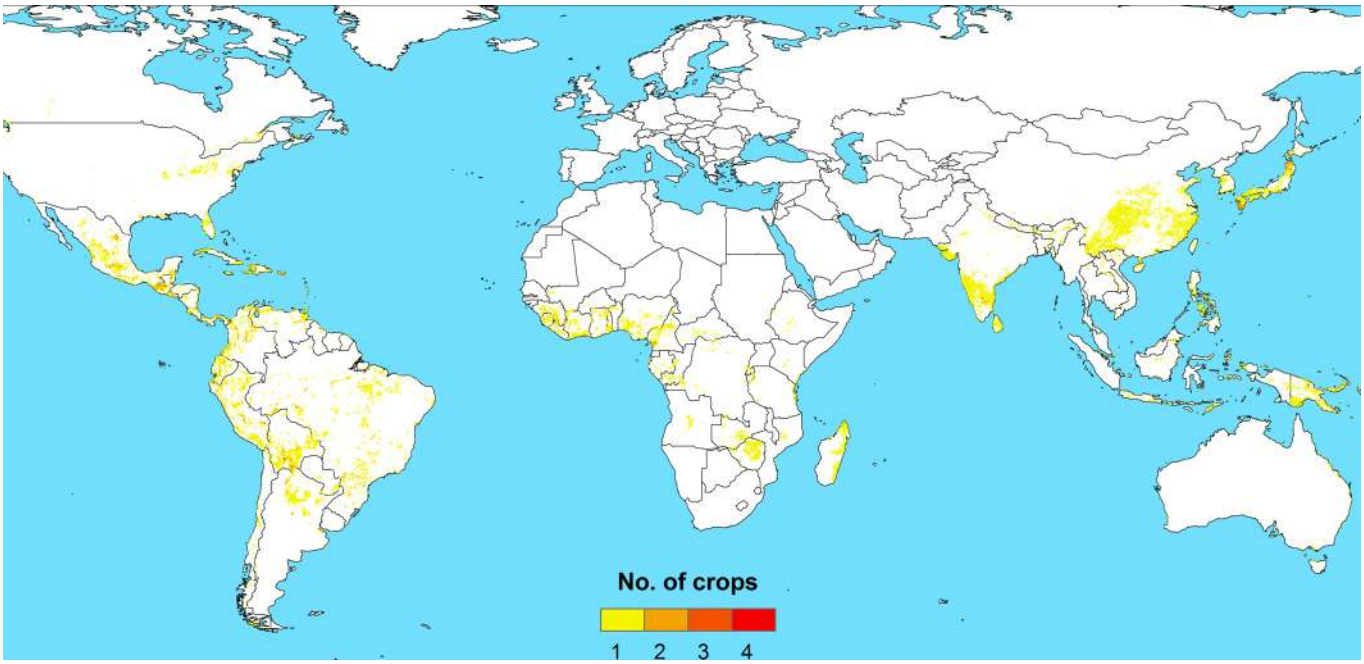
Extended Data Fig. 5 | Geographic hotspots for further collection for the *ex situ* conservation of crop landrace groups. Geographic hotspots for further collection for the *ex situ* conservation of crop landrace groups. The map displays 'gap richness' across the predicted worldwide distributions of 71 landrace groups of 25 cereal, pulse, and starchy root/tuber/fruit crops within their geographic regions of diversity, indicating where landrace groups are expected to occur and have not yet been collected and conserved in genebanks. Darker colors indicate greater numbers of un-collected crop landrace groups potentially overlapping in the same 2.5 arc-minute cells.



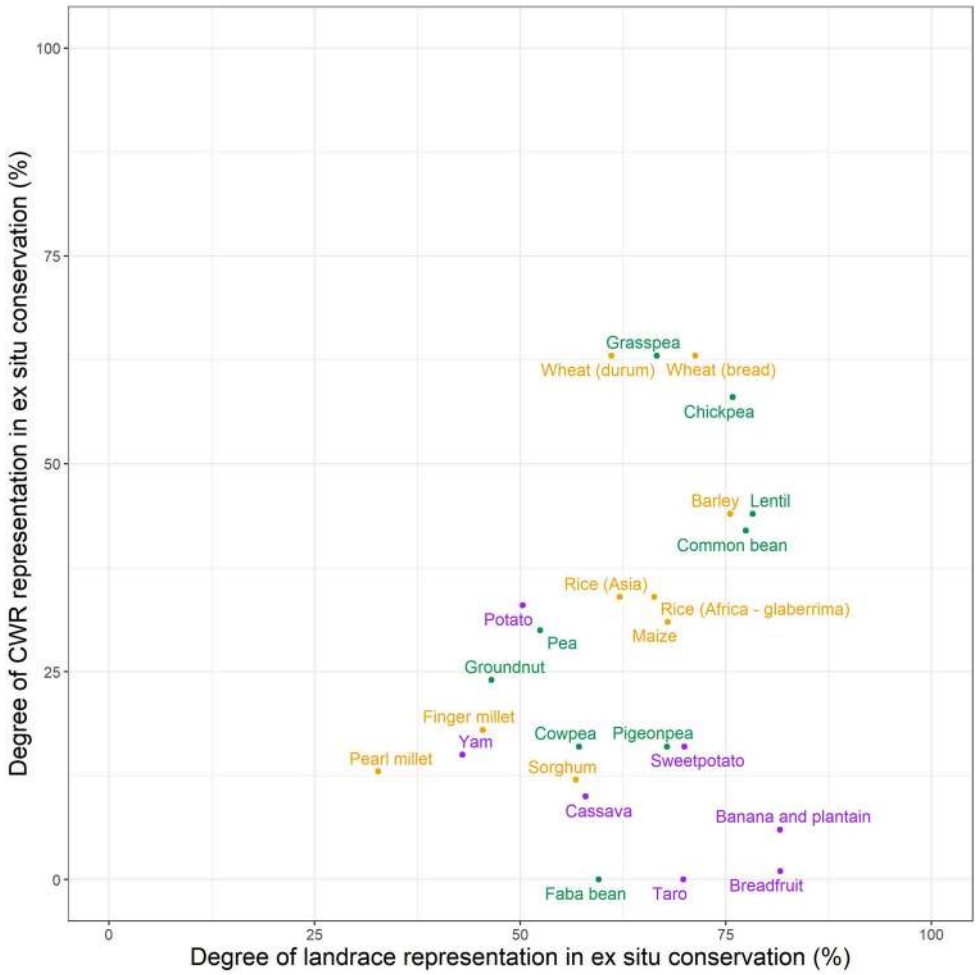
Extended Data Fig. 6 | Geographic hotspots for further collection for the *ex situ* conservation of landrace groups of cereal crops. Geographic hotspots for further collection for the *ex situ* conservation of landrace groups of cereal crops. The map displays 'gap richness' across the predicted distributions of landrace groups of 9 cereal crops within their geographic regions of diversity, indicating where landrace groups are expected to occur and have not yet been collected and conserved in genebanks. Darker colors indicate greater numbers of un-collected cereal crop landrace groups potentially overlapping in the same 2.5 arc-minute cells, quantified in terms of numbers of crops.



Extended Data Fig. 7 | Geographic hotspots for further collection for the *ex situ* conservation of landrace groups of pulse crops. Geographic hotspots for further collection for the *ex situ* conservation of landrace groups of pulse crops. The map displays 'gap richness' across the predicted distributions of landrace groups of 9 pulse crops within their geographic regions of diversity, indicating where landrace groups are expected to occur and have not yet been collected and conserved in genebanks. Darker colors indicate greater numbers of un-collected pulse crop landrace groups potentially overlapping in the same 2.5 arc-minute cells, quantified in terms of numbers of crops.



Extended Data Fig. 8 | Geographic hotspots for further collection for the *ex situ* conservation of crop landrace groups of starchy root, tuber, and fruit crops. Geographic hotspots for further collection for the *ex situ* conservation of crop landrace groups of starchy root, tuber, and fruit crops. The map displays 'gap richness' across the predicted distributions of landrace groups of 7 starchy root, tuber, and fruit crops within their geographic regions of diversity, indicating where landrace groups are expected to occur and have not yet been collected and conserved in genebanks. Darker colors indicate greater numbers of un-collected starchy root, tuber, and fruit crop landrace groups potentially overlapping in the same 2.5 arc-minute cells, quantified in terms of numbers of crops.



Extended Data Fig. 9 | Comparison of *ex situ* conservation representation of crop landrace groups and crop wild relative (CWR) for 25 cereal, pulse, and starchy root/tuber/fruit crops. Comparison of *ex situ* conservation representation of crop landrace groups and crop wild relative (CWR) for 25 cereal, pulse, and starchy root/tuber/fruit crops. For CWR, conservation representation results were first averaged across CWR taxa in each crop genepool[19]. The summary results were also averaged across related crops assessed here; for example, the results for three yam crop genepools were averaged to form a single result for the global yam genepool. The crop genepool results were then transformed to the crop landrace scale and format used here, and are compared to the crop aggregated-level conservation representation average (%) estimate. Crop wild relatives of taro were not assessed in Castaneda-Alvarez et al. (2016)19; for this figure the pertinent score was set to zero. Cereals are displayed in gold, pulses in green, and starchy roots, tubers, and fruits in purple.

nature portfolio

Corresponding author(s): Colin Khoury

Last updated by author(s): 2022_3_19

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software or code was used to collect data. Data collection is fully described in the Methods section.
Data analysis	We created custom code which is permanently available at: https://github.com/CIAT-DAPA/gap_analysis_landraces. A code availability statement is included in the Methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Occurrence data - For ex situ conservation records, occurrences marked as landraces were retrieved from two major online databases: the Genesys Plant Genetic Resources portal32 and the World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture (WIEWS) of the Food and Agriculture Organization of the United Nations33. Occurrences were also obtained directly from individual international genebank information systems: AfricaRice, the International Transit Centre and Musa Germplasm Information System of Bioversity International34, CePaCT, CIAT, CIMMYT, CIP, ICARDA, ICRISAT, IITA, and IRRI, as well as from the USDA Genetic Resources Information Network (GRIN)–Global35 and the Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO)36. Occurrences were compiled from the Global Biodiversity Information Facility (GBIF)37, with 'living specimen' records classified as ex situ conservation

records and the remaining serving as reference sightings for use in distribution modeling. Reference occurrences were also drawn from published literature (Supplementary Dataset 2 [Supplementary References]).

Spatial predictor data - We compiled and calculated spatially explicit gridded information for 50 potential environmental and cultural predictors of landrace distributions, including climatic, topographic, evolutionary history, and socioeconomic variables (Supplementary Dataset 1 [Supplementary Table 3])13. For climate data, we gathered or derived 39 variables, from WorldClim version 238 and Environmental Rasters for Ecological Modeling (ENVIREM)39. We included elevation from the Shuttle Radar Topography Mission (SRTM) dataset of the CGIAR-Consortium on Geospatial Information portal40,41. Two crop evolutionary history proxies were included: distance to human settlements before the year CE 150042, and distance to known wild progenitor populations13. The eight socioeconomic variables included population density43, distance to navigable rivers44, percentage of the area under irrigation45, population accessibility46,47, geographic distributions of ethnic or cultural groups48, and crop harvested area, production quantity, and yield49. All predictor data were scaled to 2.5 arc-minute resolution with World Geodetic System (WGS) 84 as a datum. Occurrence data, including spatial predictor variable results (at 2.5 arc minute resolution) for each occurrence (available at Supplementary Dataset 3). A global spatial predictor dataset (2.5 arc minute resolution, all 50 variables) (available at Supplementary Dataset 4).

Crop landrace group classification data - For each crop, we conducted an extensive literature review to identify recognized infraspecific groups with distinct morphological, physiological, chemical, genetic, nomenclatural, or other characteristics that could be tested for environmental and cultural associations (Supplementary Dataset 1 [Supplementary Table 1], Supplementary Dataset 2 [Supplementary Methods and References]). The nature of these groups varied by crop, and included genepools, races, genetic clusters, and geographic or environmental groupings. Crops often had more than one proposed grouping or classification.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Literature review and expert consultation were conducted to understand the various possible infraspecific genetic structures within assessed crops. Occurrence data from authoritative databases and published literature were compiled for these structures. This dataset was used to test the power of ecogeographic (spatial) information in distinguishing the various landrace groups within a crop (using the 50 ecogeographic spatial predictors, compiled in Ramirez-Villegas et al. 2020). Once the infraspecific structure with highest accuracy (ecogeographic signal) was identified, a total occurrence dataset for the landraces of each crop was compiled, with each occurrence attributed to a specific landrace group. These groups were then spatially modeled. Models were then subjected to the conservation gap analysis - essentially a comparison of these models to previous locations where germplasm for these landraces has been collected and is now conserved in genebanks. This gap analysis used three main approaches in determining conservation gaps-connectivity, accessibility, and environmental difference.
Research sample	Two main research samples are identified: Crop landrace group sample (an occurrence dataset with crop landrace group information attributed) - This dataset was used to test the power of ecogeographic signal in distinguishing the various landrace groups within a crop (using the 50 ecogeographic spatial predictors). This research sample was compiled by conducting an extensive literature review to identify recognized infraspecific groups with distinct morphological, physiological, chemical, genetic, nomenclatural, or other characteristics that could be tested for environmental and cultural associations (Supplementary Dataset 1 [Supplementary Table 1], Supplementary Dataset 2 [Supplementary Methods and References]). The nature of these groups varied by crop, and included genepools, races, genetic clusters, and geographic or environmental groupings. The total occurrence dataset for each crop - For ex situ conservation records, occurrences marked as landraces were retrieved from two major online databases: the Genesys Plant Genetic Resources portal32 and the World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture (WIEWS) of the Food and Agriculture Organization of the United Nations33. Occurrences were also obtained directly from individual international genebank information systems: AfricaRice, the International Transit Centre and Musa Germplasm Information System of Bioversity International34, CePaCT, CIAT, CIMMYT, CIP, ICARDA, ICRISAT, IITA, and IRRI, as well as from the USDA Genetic Resources Information Network (GRIN)–Global35 and the Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO)36. Occurrences were compiled from the Global Biodiversity Information Facility (GBIF)37, with 'living specimen' records classified as ex situ conservation records and the remaining serving as reference sightings for use in distribution modeling. Reference occurrences were also drawn from published literature (Supplementary Dataset 2 [Supplementary References]). Duplicated observations within or between data sources were eliminated, with a preference to utilize the most original data. Coordinates were corrected or removed when latitude and longitude were equal to zero or inverted, located in water bodies or in the wrong country, or had poor resolution (< 2 decimal places). Occurrences were clipped to study areas per crop. The complete occurrence dataset is available in Supplementary Dataset 3. After performing the steps described in a, all other occurrence data records not attributed to landrace groups were predicted. This complete dataset was then used for crop landrace distribution modeling and conservation gap analysis
Sampling strategy	Crop landrace group structures were tested using 15-fold cross-validation with 80% training and 20% testing. We accepted a given classification if each of its components was predicted with an average cross-validated accuracy of at least 80%. Spatial distribution models were fitted through five-fold (K = 5) cross-validation with 80% training and 20% testing. For each fold, we calculated the area under the receiving operating characteristic curve (AUC), sensitivity, specificity, and Cohen's kappa as measures of model performance. To create a single prediction that represents the probability of occurrence for the landrace group, we computed the

median across K models. Geographic areas in the form of pixels with probability values above the maximum sum of sensitivity and specificity were treated as the final area of predicted presence.

Data collection

Occurrence data - For ex situ conservation records, occurrences marked as landraces were retrieved from two major online databases: the Genesys Plant Genetic Resources portal32 and the World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture (WIEWS) of the Food and Agriculture Organization of the United Nations33. Occurrences were also obtained directly from individual international genebank information systems: AfricaRice, the International Transit Centre and Musa Germplasm Information System of Bioversity International34, CePaCT, CIAT, CIMMYT, CIP, ICARDA, ICRISAT, IITA, and IRRI, as well as from the USDA Genetic Resources Information Network (GRIN)–Global35 and the Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO)36. Occurrences were compiled from the Global Biodiversity Information Facility (GBIF)37, with 'living specimen' records classified as ex situ conservation records and the remaining serving as reference sightings for use in distribution modeling. Reference occurrences were also drawn from published literature (Supplementary Dataset 2 [Supplementary References]). Duplicated observations within or between data sources were eliminated, with a preference to utilize the most original data. Coordinates were corrected or removed when latitude and longitude were equal to zero or inverted, located in water bodies or in the wrong country, or had poor resolution (< 2 decimal places). Occurrences were clipped to study areas per crop. The complete occurrence dataset is available in Supplementary Dataset 3.

Spatial predictors - We compiled and calculated spatially explicit gridded information for 50 potential environmental and cultural predictors of landrace distributions, including climatic, topographic, evolutionary history, and socioeconomic variables (Supplementary Dataset 1 [Supplementary Table 3])13. For climate data, we gathered or derived 39 variables, from WorldClim version 238 and Environmental Rasters for Ecological Modeling (ENVIREM)39. We included elevation from the Shuttle Radar Topography Mission (SRTM) dataset of the CGIAR-Consortium on Geospatial Information portal40,41. Two crop evolutionary history proxies were included: distance to human settlements before the year CE 150042, and distance to known wild progenitor populations13. The eight socioeconomic variables included population density43, distance to navigable rivers44, percentage of the area under irrigation45, population accessibility46,47, geographic distributions of ethnic or cultural groups48, and crop harvested area, production quantity, and yield49. All predictor data were scaled to 2.5 arc-minute resolution with World Geodetic System (WGS) 84 as a datum. Extended descriptions of the sources and their justification for inclusion are provided in Ramirez-Villegas et al. (2020)13. For both datasets, data are provided in spreadsheets (i.e. Microsoft Excel).

Crop landrace group classification - for each crop, we conducted an extensive literature review to identify recognized infraspecific groups with distinct morphological, physiological, chemical, genetic, nomenclatural, or other characteristics that could be tested for environmental and cultural associations (Supplementary Dataset 1 [Supplementary Table 1], Supplementary Dataset 2 [Supplementary Methods and References]). The nature of these groups varied by crop, and included genepools, races, genetic clusters, and geographic or environmental groupings. Crops often had more than one proposed grouping or classification.

Timing and spatial scale

Occurrence data covers all time periods but the vast majority of data is from the past 50 years. Worldclim predictor data is annual average from 1970-2000; Envirem data is annual average from 1960-1990 (current); other predictors described in Ramirez-Villegas et al. 2020 refer to current conditions (e.g., road network, ethnic groups). Spatial scale is region and crop dependent; in total across this study, spatial scale is global, whereas the spatial resolution is 2.5 arc-min. Data collection for this work initiated in January 2017 and concluded in July 2021. We generally compiled this data crop by crop in collaboration with crop experts and often during in person collaborative workshops at each of the international research centers.

Data exclusions

Crop landrace group (infraspecific structure) data that was outperformed by other competing proposed structures was not used in the final combined analysis; results for these alternative structures are provided in Supplementary Table 2.

Reproducibility

All data is available through open access repositories; all code is available on Github (https://github.com/CIAT-DAPA/gap_analysis_landraces).

Randomization

Data: Occurrence data from authoritative databases and published literature were compiled for landrace group structures. This dataset was used to test the power of ecogeographic (spatial) information in distinguishing the various landrace groups within a crop (using the 50 ecogeographic spatial predictors, compiled in Ramirez-Villegas et al. 2020). Once the infraspecific structure with highest accuracy (ecogeographic signal) was identified, a total occurrence dataset for the landraces of each crop was compiled, with each occurrence attributed to a specific landrace group. We then built and tested classification models to determine how well the proposed groups could be predicted and distinguished based on spatial predictors, drawing from the occurrence database and training datasets compiled from the literature review. A random forest53, a support vector machine54, the K-nearest neighbor (KNN) algorithm55, and artificial neural networks56 were used to determine classification performance. The response variable was the group to which a given occurrence was assigned, whereas the explanatory variables were the spatial predictors. Models were combined into an ensemble using the mode—that is, the most frequent predicted value among the models—and tested using 15-fold cross-validation with 80% training and 20% testing (samples drawn at random). We accepted a given classification if each of its components was predicted with an average cross-validated accuracy of at least 80%. In the case of multiple classification proposals per crop, we selected the one with the best overall performance. Finally, we used the trained models to predict the corresponding group for occurrences missing such information. All landrace groups for all crops are provided in Supplementary Dataset 1 [Supplementary Table 2], with the best-performing groups identified.

Distribution modelling: To predict the probability of geographic occurrence for each landrace group within each crop, we generated MaxEnt models57,58 using the 'maxnet' R package59. Group-specific spatial predictors were selected using a combination of the variance inflation factor (VIF) and a principal component analysis (PCA) to control for excessive model complexity and variable collinearity60. We removed variables that did not contribute significantly to the variance in the PCA, defined as contributing less than 15% to the first component, and we further discarded variables with a VIF > 1061. The predictors and whether they were selected for the modeling of each landrace group are presented in Supplementary Dataset 1 (Supplementary Table 4). We generated a random sample of pseudo-absences as background points in areas that (a) were within the same ecological land units62 as the occurrence points, (b) were deemed potentially suitable according to a support vector machine classifier that uses all occurrences and predictor variables, and (c) were further than 5 km from any occurrence63. The number of pseudo-absences generated per crop group was ten times its number of unique occurrences. MaxEnt models were fitted through five-fold (K = 5) cross-validation with 80% training and 20% testing (with samples for these splits drawn at random each time). For each fold, we calculated the area under the receiving

operating characteristic curve (AUC), sensitivity, specificity, and Cohen's kappa as measures of model performance. To create a single prediction that represents the probability of occurrence for the landrace group, we computed the median probability across K models. Geographic areas in the form of pixels with probability values above the maximum sum of sensitivity and specificity were treated as the final area of predicted presence13.

Gap analysis validation - Spatial ex situ conservation gaps were determined from the conservation gap scores using a cross-validation procedure to derive a threshold for each score. We created synthetic gaps by removing existing genebank occurrences in five randomly chosen circular areas with a 100 km radius within the distribution model. We then tested whether these artificial gaps could be predicted by our gap analysis, identifying the threshold value of each score that would maximize the prediction of these synthetic gaps. Performance for each of the five gap areas was assessed using AUC, sensitivity, and specificity. The average cross-area threshold value was calculated for each score to discern pixels with a high likelihood of finding ex situ conservation gaps and which thus were higher priority for further field sampling. These were pixels with combined gap scores above the threshold, assigned a value of 1, as opposed to the relatively well-conserved areas below the threshold, which were assigned a value of 0.

Blinding

Blinding was not relevant to this study. All relevant data for pertinent crop landraces was acquired from authoritative databases and from published literature. These were used to test classification models, build distribution models, and perform gap analysis and validation based on statistical power/significance, as described above

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging