Received: 24 June 2019 | Revised: 31 January 2020 | Accepted: 4 February 2020

DOI: 10.1111/ddi.13046

BIODIVERSITY RESEARCH

Diversity and Distributions WILEY

A gap analysis modelling framework to prioritize collecting for ex situ conservation of crop landraces

Julian Ramirez-Villegas[1,2] | Colin K. Khoury[1,3,4] | Harold A. Achicanoy[1] | Andres C. Mendez[1] | Maria Victoria Diaz[1] | Chrystian C. Sosa[1] | Daniel G. Debouck[1] | Zakaria Kehel[5] | Luigi Guarino[6]

[1]International Center for Tropical Agriculture (CIAT), Cali, Colombia

[2]CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), c/o CIAT, Cali, Colombia

[3]United States Department of Agriculture, Agricultural Research Service, National Laboratory for Genetic Resources Preservation, Fort Collins, CO, USA

[4]Department of Biology, Saint Louis University, St. Louis, MO, USA

[5]International Center for Agricultural Research in the Dry Areas (ICARDA), Rabat, Morocco

[6]Global Crop Diversity Trust, Bonn, Germany

Correspondence
Julian Ramirez-Villegas, International Center for Tropical Agriculture (CIAT), Km 17 Recta Cali-Palmira, 763537, Cali, Colombia.
Email: j.r.villegas@cgiar.org

Editor: Martin Jung

Abstract

Aim: The conservation and effective use of crop genetic diversity are crucial to overcome challenges related to human nutrition and agricultural sustainability. Farmers' traditional varieties ("landraces") are major sources of genetic variation. The degree of representation of crop landrace diversity in ex situ conservation is poorly understood, partly due to a lack of methods that can negotiate both the anthropogenic and environmental determinants of their geographic distributions. Here, we describe a novel spatial modelling and ex situ conservation gap analysis modelling framework for crop landraces, using common bean (*Phaseolus vulgaris* L.) as a case study.

Location: The Americas.

Methods: The modelling framework includes five main steps: (a) determining relevant landrace groups using literature to develop and test classification models; (b) modelling the potential geographic distributions of these groups using occurrence (landrace presences) combined with environmental and socioeconomic predictor data; (c) calculating geographic and environmental gap scores for current genebank collections; (d) mapping ex situ conservation gaps; and (e) compiling expert inputs.

Results: Modelled distributions and conservation gaps for the two genepools of common bean (Andean and Mesoamerican) were robustly predicted and align well with expert opinions. Both genepools are relatively well conserved, with Andean ex situ collections representing 78.5% and Mesoamerican 98.2% of their predicted geographic distributions. Modelling revealed additional collection priorities for Andean landraces occur primarily in Chile, Peru, Colombia and, to a lesser extent, Venezuela. Mesoamerican landrace collecting priorities are concentrated in Mexico, Belize and Guatemala.

Conclusions: The modelling framework represents an advance in tools that can be deployed to model the geographic distributions of cultivated crop diversity, to assess the comprehensiveness of conservation of this diversity ex situ and to highlight geographic areas where further collecting may be conducted to fill gaps in ex situ conservation.

KEYWORDS
common bean, crop diversity, gap analysis, landrace, plant genetic resources

1 | INTRODUCTION

The effective use of crop genetic resources—including both traditional farmer varieties (or "landraces") and wild relatives—is important in efforts to overcome challenges related to human nutrition and agricultural sustainability (Burke, Lobell, & Guarino, 2009; Esquinas-Alcázar, 2005; Khoury et al., 2016). Progress in plant breeding and crop diversification is dependent on crop understanding and utilizing the available genetic resources (Glaszmann, Kilian, Upadhyaya, & Varshney, 2010; Hajjar & Hodgkin, 2007). The erosion of genetic diversity within many common crops has occurred over the last century through a combination of land use change, habitat degradation and the ongoing adoption of improved crop varieties or the substitution of crop species by farming communities (Hoisington et al., 1999; van de Wouw, Kik, Hintum, Treuren, & Visser, 2010). In some crops, only a fraction of the genetic diversity once present is still found today in farmers' fields, for example wheat landraces in the Fertile Crescent (Gepts, 2006; Harlan, 1975). Consequently, ex situ crop genebanks have become essential not only for distributing of genetic resources to various users (e.g. breeders, other genebanks), but also for their conservation of such resources (Gepts, 2006; Hoisington et al., 1999).

Understanding the representation of crop diversity in ex situ repositories provides a foundation for conservation planning (Castañeda-Álvarez et al., 2016; García, Parra-Quijano, & Iriondo, 2017; van Treuren, Engels, Hoekstra, & Hintum, 2009). Methods to assess the current degree of representation, and to inform further collecting efforts, have increasingly been developed for over more than a decade [e.g. Rodrigues et al., (2004); Maxted, Dulloo, Ford-Lloyd, Iriondo, and Jarvis (2008)]. Due to the general lack of genetic data, these methods are generally based on ecogeographic methodologies as a proxy for assessments of genetic diversity (Khoury et al., 2019; Ramirez-Villegas, Khoury, Jarvis, Debouck, & Guarino, 2010). Such methods have proved useful in estimating the representation of wild relatives and other wild species in genebanks in comparison with standing extant diversity in their natural environments (Castañeda-Álvarez et al., 2016; Khoury et al., 2019; Syfert et al., 2016). However, their application to cultivated plants, whose spatial distributions are determined by anthropogenic factors as well as environmental drivers, is limited (Fuller, 2007; Hilbert et al., 2017; Morris et al., 2013). This represents a critical gap, since cultivated materials are generally preferred over wild relatives for use by plant breeders (Camacho Villa, Maxted, Scholten, & Ford-Lloyd, 2005; Hammer, Knüpffer, Xhuveli, & Perrino, 1996).

Here, we present a conservation gap analysis modelling framework for cultivated crop diversity, that improves on current ecogeographic methods, using landraces of the common bean (*Phaseolus vulgaris* L.) as a case study. As opposed to previous analyses of the distributions of cultivated crop diversity [e.g. Upadhyaya, Reddy, Irshad Ahmed, and Gowda (2012), Upadhyaya et al. (2017)], our methods explicitly aim to include anthropogenic drivers in the modelling of the distributions of landraces. The results predict geographic areas that are likely gaps in ex situ landrace conservation collections and provide metrics that can be used to track conservation progress. These results are supplemented with expert knowledge, which is vital for elucidating spatial patterns and drivers of range change that are difficult to model.

Common bean is the most widely human-consumed grain legume, playing an essential role in food and nutritional security, particularly in Latin America and Sub-Saharan Africa (Beebe, 2012; Broughton et al., 2003). Two independent domestication events of wild *P. vulgaris* have been identified—one in Mexico and Central America, and the second in the Andes mountains of South America (Gepts, Osborn, Rashka, & Bliss, 1986). Significant movement of genetic material and gene exchange between genepools has occurred since domestication, with considerable overlap in current geographic distributions, both in the Neotropics and across other major cultivation areas (Singh, 1989; Singh, Gepts, & Debouck, 1991). These processes have resulted in recognized secondary regions of diversity in Brazil, Europe, Africa and Asia (Escribano & De Ron, 1991; Lobo Burle et al., 2011; Logozzo et al., 2007).

Globally, there are some 250 ex situ collections of cultivated *P. vulgaris*, with the largest and most diverse maintained at the International Centre for Tropical Agriculture (CIAT) with ~40,000 accessions, and the United States Department of Agriculture (USDA) National Genetic Resources Program with ~15,000 accessions (Debouck, 2014). Here, we assess the representation of common bean landraces in such major genebank collections, including estimating overall conservation and identifying gaps.

2 | MATERIALS AND METHODS

Our modelling framework first necessitates the defining of the study area, gathering of landrace occurrence and characterization data, and compilation of environmental and socioeconomic spatial predictor information. The modelling and conservation gap analysis is then performed, consisting of five main steps: (a) determining relevant landrace groups using the literature to develop and test classification models; (b) modelling the potential geographic distributions of these groups using the occurrence and predictor data; (c) calculating geographic and environmental gap scores for current genebank collections; (d) mapping ex situ conservation gaps; and (e) compiling expert inputs. The overall process is depicted in Figure 1.

2.1 | Study area

Crop landraces have been defined as "dynamic population(s) of a cultivated plant that has historical origin, distinct identity and

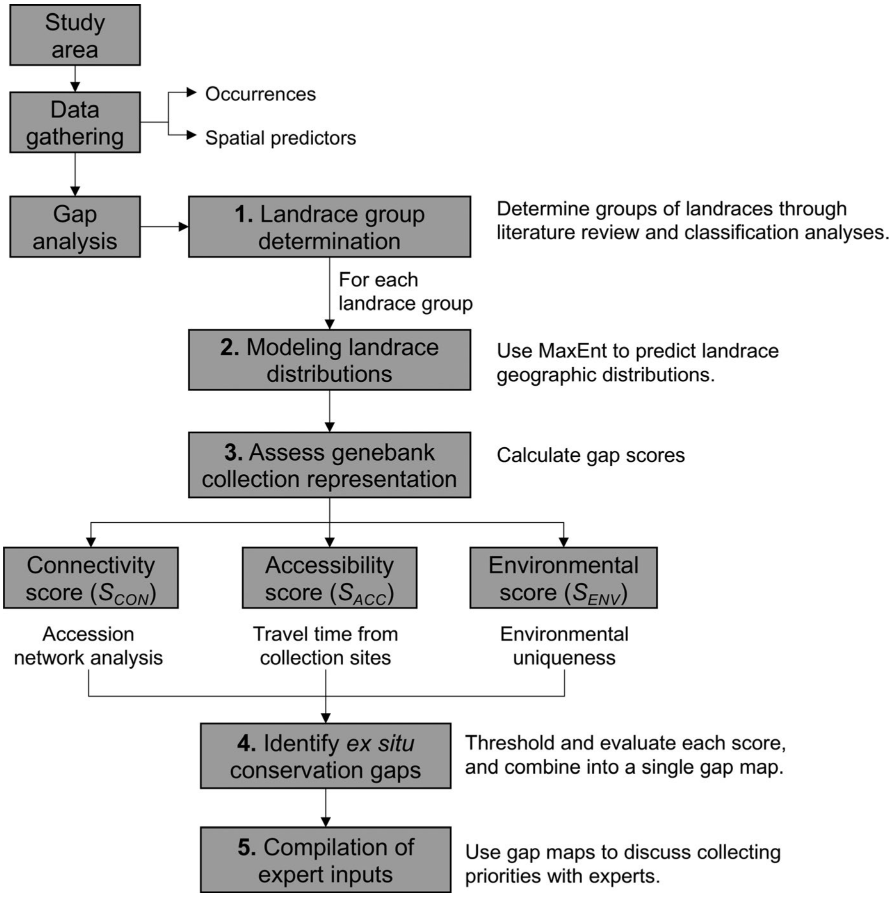
Study area
Data gathering
Occurrences
Spatial predictors
Gap analysis
1. Landrace group determination
Determine groups of landraces through literature review and classification analyses.
For each landrace group
2. Modeling landrace distributions
Use MaxEnt to predict landrace geographic distributions.
3. Assess genebank collection representation
Calculate gap scores
Connectivity score (S_{CON})
Accessibility score (S_{ACC})
Environmental score (S_{ENV})
Accession network analysis
Travel time from collection sites
Environmental uniqueness
4. Identify *ex situ* conservation gaps
Threshold and evaluate each score, and combine into a single gap map.
5. Compilation of expert inputs
Use gap maps to discuss collecting priorities with experts.

FIGURE 1 Conservation gap analysis modelling framework implemented in this study

lacks formal crop improvement, as well as often being genetically diverse, locally adapted and associated with traditional farming systems" (Camacho Villa et al., 2005; Casañas, Simó, Casals, & Prohens, 2017). A landrace can be further classified as autochthonous when grown in the original location where it developed its unique genetic and socioeconomic characteristics through grower selection and allochthonous when introduced from another region and then locally adapted. "Secondary" landraces may also be recognized, developed by the formal plant breeding sector but now maintained through repeated farmer selection and seed saving (Zeven, 1998).

While landraces cultivated over time in any given location may possess novel traits useful for plant breeding, our distribution modelling method rests on the premise that these varieties have distinct, local environmental adaptations (see 2.4.1–2.4.2). As adaptation to environment is developed over time, the geographic areas where landraces have occurred the longest—the origins and primary regions of diversity—would be considered to have the most significant association between environmental adaptation and genetic variation (Khoury et al., 2016). For this reason, landrace distribution modelling may focus foremost on autochthonous ranges.

For our case study, we focused on the Americas as the centre of domestication and primary region of diversity for *P. vulgaris* (Gepts et al., 1986). We included all areas extending from the southern United States to central Chile and northern Argentina, including the Caribbean, as this broadly includes the two reported domestication events and distributions of the progenitor and close relatives of the species (Chacon, Pickersgill, & Debouck, 2005; Gepts et al., 1986). We also included Brazil since it is geographically close to the putative regions of domestication and because existing evidence suggests clear relationships between Brazilian bean landraces and Andean and Mesoamerican types (Lobo Burle et al., 2011; Lobo Burle, Fonseca, Kami, & Gepts, 2010).

2.2 | Landrace occurrence and characterization data

Our distribution modelling and conservation gap analysis modelling framework requires geographic occurrence (presence) data for landraces and information on the locations where these landraces have been previously collected for conservation ex situ, as well as characterization data on the landrace accessions. To assess the world's common bean landrace collections, we compiled available genebank accession-level passport (i.e. site where collected) data from major online germplasm databases, including the Genesys plant genetic resources portal (Global Crop Diversity Trust, 2019) and the United Nations Food and Agriculture Organization World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture (WIEWS) (FAO, 2019). To ensure inclusion of the crop's major germplasm collections, we specifically gathered occurrence and characterization data from the CIAT database (CIAT, 2018), freely available at and from the United States Department of Agriculture (USDA) Genetic Resources Information Network (GRIN)–Global (USDA ARS NPGS, 2018).

Additional occurrences were gathered from the Global Biodiversity Information Facility (GBIF) (GBIF.org, 2019), which contained 25,670 observations from herbaria, botanic gardens and other plant repositories, to provide independent data from non-genebank sources. We compiled the datasets into a single database and performed a thorough quality check of all records. Duplicated observations were eliminated with preference to maintain original data, for example, USDA-GRIN or CGIAR records included in Genesys or WIEWS were discarded. Coordinates were corrected, or if not possible, eliminated, when latitude and longitude were equal to zero, located in inland water bodies or in the ocean, located in the wrong country, had an inverted sign in the latitude and/or longitude or had low coordinate precision (i.e. with less than 2 decimal places). Our full occurrence dataset for *P. vulgaris* is available in Dataset S1.

2.3 | Spatial predictors

With the aim of compiling a robust global dataset of important environmental and anthropogenic drivers of the geographic distributions of crop landraces, we gathered and/or calculated spatially explicit (gridded) information for a total of 50 potential predictors, including climate, topography, diversity and domestication and socioeconomic variables (Table S2.1). For climate, we used a total of 40 variables, derived from a combination of the WorldClim version 2 (Fick & Hijmans, 2017) and the Environmental Rasters for Ecological Modelling (ENVIREM) (Title & Bemmels, 2018) databases. We included topography from the Shuttle Radar Topography Mission (SRTM) dataset of the CGIAR-Consortium on Geospatial Information (CSI) portal (Jarvis, Reuter, Nelson, & Guevara, 2008; Reuter, Nelson, & Jarvis, 2007). Two crop genetic diversity and domestication proxy variables were included, namely the distance to known common bean wild relative populations and the distance to human settlements before year AD 1500. Regarding socioeconomic variables (8 in total), we included datasets on the geographic distribution of ethnic groups (Weidmann, Rød, & Cederman, 2010); crop yield, harvested area and crop production quantity (You et al., 2017); population density (CIESIN, 2018); population accessibility (Nelson, 2008); distance to navigable rivers (Natural Earth, 2019); and percentage of area under irrigation (Siebert, Henrich, Frenken, & Burke, 2013). All spatial predictor data were scaled to or computed on a common 2.5 arc-min grid, using the geographic coordinate system (GCS) with WGS84 as datum. A complete description of these data sources and their justification for inclusion is provided in Text S2.1 and Table S2.1. The full dataset of ecogeographic and socioeconomic variables is available in Dataset S1.

2.4 | Landrace distribution modelling and conservation gap analysis

2.4.1 | Determination of landrace groups

Crop landraces are domesticated, locally adapted varieties of crops, developed through farmer selection over time in specific agricultural ecosystems (Camacho Villa et al., 2005; Jones et al., 2008) and, for most crops, are considered to number in the thousands (Harlan, 1975; Jones et al., 2008). Crop landraces are associated with specific local adaptation traits and farmer preferences, and an understanding of these drivers is important to modelling their potential distributions. Given the large number of landraces and the knowledge necessary to distinguish their biocultural and ecological differences, our method seeks a compromise between the recognition of this complexity and performance of spatial modelling at scales which are feasible and permit comparison with existing genebank collections.

Therefore, the first step of our modelling method was to identify recognized groups within the crop that could be tested for whether they have distinct environmental and socioeconomic niches. We used Google Scholar™ to identify and review publications that, through morphological, physiological, chemical, genetic, nomenclatural or other characters, establish or propose groups of landraces (e.g. by identifying genepools, races, domestication centre(s), genetic clusters or other acknowledged groupings) (Table S2.2).

We then used classification models to test the significance of these classifications. The classification models allowed us to determine whether the classes identified could be predicted on the basis of the spatial predictors from Section 2.3. This process used data from the occurrence database (if the distinguishing characters of the identified landrace groups were reported in the database) or from training datasets containing both characters and geographic coordinates, compiled from the literature review. For this analysis, we used random forest (RF) (Pal, 2005), support vector machine (SVM) (Meyer, Leisch, & Hornik, 2003), K-nearest neighbour (KNN) (Guo, Wang, Bell, Bi, & Greer, 2003) and artificial neural networks (ANN) (Dreiseitl & Ohno-Machado, 2002). The response variable in all models was the group in which a given accession was assigned, whereas the explanatory variables were the spatial predictors. Models were combined into an ensemble using the mode (i.e. the most frequent predicted value amongst models) and tested using 15-fold cross-validation (80% training, 20% testing). We accepted a given classification if each of its classes was predicted with an average cross-validated accuracy of at least 80% (i.e. 8 of every 10 accessions are predicted correctly). Finally, we used the trained models to predict the corresponding class for any records in the database missing such information.

2.4.2 | Modelling landrace geographic distributions

The objective of this step was to develop a Landrace Distribution Model (LDM) which describes the probability of occurrence of the landrace groups derived from Section 2.4.1. To predict the probability of occurrence for each landrace group, we fitted a MaxEnt model (Elith et al., 2010; Phillips, Anderson, & Schapire, 2006) using the "maxnet" R package (Phillips, Anderson, Dudík, Schapire, & Blair, 2017). We chose MaxEnt as a standard and very commonly used tool for species distribution modelling (Costa, Nogueira, Machado, & Colli, 2010; Elith et al., 2006). MaxEnt has been demonstrated to yield robust results when compared with other species distributions modelling algorithms (Barbet-Massin, Jiguet, Albert, & Thuiller, 2012; Elith et al., 2006; Giovanelli, Siqueira, Haddad, & Alexandrino, 2010).

Variables used in the model were sub-selected from the environmental and socioeconomic predictors using a combination of the variance inflation factor (VIF) and a principal component analysis (PCA) to control for unwarranted model complexity and collinearity between explanatory variables (Warren & Seifert, 2011). We first removed any variables that did not contribute significantly (defined as contributing <15% to the first component) to the variance in the PCA and then discarded any variables with a VIF greater than 10 (Braunisch et al., 2013). The list of variables selected (or alternatively eliminated) for use in modelling are available in Table S2.1. We tried different model configurations (i.e. only climate, only non-climate and both) but present only the best-performing one (i.e. where all variables are used). Other results are presented in Text S2.2.

Background points (pseudo-absences) were generated based on the three-step method of Senay, Worner, and Ikeda (2013). In short, we took a random sample of pseudo-absences from areas that (a) were within the same ecological land units [as reported by Sayre et al. (2014)] as the occurrence points, (b) were deemed as potentially suitable according to a support vector machine (SVM) classifier that uses all occurrences and predictor variables and (c) were further than 5 km from any occurrence. The number of pseudo-absences drawn was equivalent to 10 times the total number of unique occurrences for a given landrace group.

MaxEnt models were fitted through a fivefold (K = 5) cross-validation process in which 80% of the occurrences (and pseudo-absences) were used to train the models, and the remaining 20% were used for 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 models. Finally, any areas above the probability value at the maximum sum of sensitivity and specificity were considered the final Landrace Distribution Model (LDM).

2.4.3 | Calculating geographic and environmental gap scores

We developed three scores that compare the geographic and environmental diversity in existing ex situ conservation collections against the LDM, revealing ex situ conservation gaps.

The accession connectivity score (S_{CON}) was formed with Delaunay triangulation (Lee & Schachter, 1980), that is, triangles linking every three (closest) accession occurrence locations, using the "deldir" R package (Turner, 2019). For each 2.5 arc-min pixel within each Delaunay triangle, we computed S_{CON} following Equation 1.

$$S_{\mathrm{CON}} = \frac{A_{T-i}}{\max\left(A_{T-i}, \cdots, A_{T-n}\right)} * \left(1 - D_{C-i}\right) * D_{NV-i} \qquad (1)$$

where, A_{T-i} is the area of the triangle (km^2) where the pixel is located (i.e. the i-th triangle), max (A_{T-i}, … A_{T-n}) is the area of the largest triangle amongst all triangles, D_{C-i} is the Euclidean distance from the pixel to the centroid of the triangle where it is located, normalized by the longest distance (using all pixels) within the given triangle, D_{NV-i} is the Euclidean distance from the pixel to the nearest vertex of the triangle where it is located, normalized by the longest distance (using all pixels) within the given triangle.

From Equation 1, it is clear that S_{CON} for any given pixel is largest (i.e. increases the likelihood of gaps) when the triangle is large (i.e. high area), when the pixel is close to the centroid of the triangle (i.e., where there are no accessions) and when the distance to the vertices (where the accessions are located) is high.

The accession accessibility score (S_{ACC}) was calculated by computing travel time from each pixel within the LDM to the nearest genebank accession, following Weiss et al. (2018). Travel time was in this case estimated through a product of the distance and the speed of travel (defined by a friction surface). Once the travel time from each location was computed, it was normalized by dividing pixel values by the longest travel time within the LDM, to derive a metric in the range 0–1, with high values reflecting long travel time.

The environmental score (S_{ENV}) measures how well the environments where the landraces are distributed are represented in ex situ collections. We first performed a hierarchical clustering analysis (Ward's method) for the pixels in the LDM using the predictor variables used to construct the LDM. On a per cluster basis, we computed the Mahalanobis distance between each pixel and the environmentally closest germplasm accession. The distance was finally normalized (0–1), with high values indicative of large distances to sites with similar environments that have previously been collected for ex situ conservation.

2.4.4 | Mapping ex situ conservation gaps

Spatial ex situ conservation gaps were calculated from the conservation gap scores using a cross-validation procedure to derive a threshold for each landrace group and each of the gap scores (S_{CON}, S_{ACC}, S_{ENV}). To do so, we created synthetic (artificial) gaps by removing genebank occurrences in five randomly chosen circular areas of 100 km radius within the LDM. We then tested whether these synthetic gaps could be predicted by our method and determined the threshold value of each gap score that would maximize the prediction of these synthetic gaps. Performance for each of the five synthetically created gaps was assessed using the AUC, sensitivity and specificity. Finally, the average threshold value of each gap score, maximizing the prediction of the synthetic gaps (balanced with minimizing false positives), was used to discretize the gap score datasets into areas with a high priority for further collecting (areas with gap score above the threshold, assigned a value of 1) as opposed to relatively well-conserved areas (areas with gap score below the threshold, assigned a value of 0).

We then summed the three binary gap score maps, resulting in a map with values from 0 to 3. Areas with a value of 0 indicate that there are no accession connectivity, accessibility or environmental gaps (i.e. well-conserved areas); areas with a value of 1 indicate gaps exist due any of accession connectivity, accessibility or environment (low confidence gaps); areas with a value of 2 indicate gaps exist due to two metrics (medium confidence gaps), and values of three indicate gaps for all metrics (highest confidence gaps). We termed this 3-value area our "final gaps map."

Once the final gaps map was calculated, we estimated the coverage of existing germplasm collections. The coverage is simply the area considered as gap divided by the total area of the LDM. We compute only the coverage resulting from the agreement of the three gap metrics, as an upper-level coverage estimation.

2.4.5 | Compilation of expert inputs

Gap analysis is a tool for assessing collection completeness as well as to plan collecting (García et al., 2017; Marinoni, Bortoluzzi, Parra-Quijano, Zabala, & Pensiero, 2015). Collecting based on model predictions may require extensive discussion with local institutions and crop experts including botanists, collectors, agronomists and breeders. This is because agricultural landscapes are highly dynamic, and areas predicted with gaps may have been subject to recent land use change, varietal replacement by improved or foreign material or significant genetic drift, resulting in loss of uncollected genetic material predicted to be of value (Hammer et al., 1996; van Heerwaarden, Hellin, Visser, & Eeuwijk, 2009; van de Wouw et al., 2010). This means that while the "final gaps map" resulting from Section 2.4.4 provides a detailed regional picture of collecting priorities, the planning of collecting missions will effectively require discussion with experts and further analysis (Greene et al., 1999a,1999b; Jarvis et al., 2005). In this sense, gap analysis results are a discussion support tool that aims at guiding, rather than prescribing where and how collecting may be done. Here, we illustrate this by conducting a semi-structured interview process with two relevant crop landrace experts. These inputs were used to add additional value to the model results.

3 | RESULTS

3.1 | Environmentally distinguishable groups of common bean landraces

Our literature review indicated that a single major classification system based on genetic, morphological and physiological characteristics has been accepted for common bean landraces. This system, first proposed by Singh et al. (1991), classifies beans into two genepools—Andean and Mesoamerican. The Andean genepool, derived from the domestication event proposed to have occurred around Peru, Chile and Bolivia, is composed of typically larger-seeded genotypes. The Mesoamerican genepool, derived from the domestication event in Mexico and Central America, is typically composed of smaller-seeded genotypes (Singh et al., 1991). These and subsequent authors divide these genepools into races according to morphological criteria, agro-ecological adaptation and genetic data (see Table S2.2 for a complete list of publications reviewed). The Andean genepool is divided into races Chile, Nueva Granada and Peru, whereas the Mesoamerican genepool contains races Guatemala, Durango–Jalisco and Mesoamerica (Blair, Díaz, Hidalgo, Díaz, & Duque, 2007; Blair, Díaz, Buendía, & Duque, 2009; Singh et al., 1991).

We tested a variety of accession-level data pertinent to common bean genepools, including seed protein type; seed weight, colour shape and brightness; and landrace names. Based on degree of acceptance in published literature and availability of accession-level data with geographic coordinates, we ultimately based our training data on genepool designations given in the CIAT accessions dataset and specific accession numbers gathered from the reviewed literature (Table S2.2).

Our average classification accuracy at the genepool level was 86% (88.3% for Andean and 85% for Mesoamerican landraces), indicating that these two genepools have distinct environmental and socioeconomic signatures, with Mesoamerican beans being present in lower, drier and hotter places compared to Andean beans. Identified predictors (see Figure S2.1) for the classification models agree with previously reported predictors of domesticated and wild bean distributions (Cortes, Monserrate, Ramirez-Villegas, Madrinan, & Blair, 2013; Ramirez-Villegas et al., 2010). At the race level, the classification accuracy was low 58.5% as a mean across all races and hence deemed not informative. Based on these results, we concluded that the genepool level was the most appropriate for all subsequent distribution modelling and conservation gap analysis steps. Hence, in all following sections we show results separately for Andean and Mesoamerican common bean landrace groups.

3.2 | Geographic distributions of common bean landrace groups

Figure 2 shows the predicted geographic distributions of Andean (Figure 2a) and Mesoamerican (Figure 2b) landraces. Cross-validated MaxEnt models performed well with mean AUC values of 0.973 (Andean) and 0.996 (Mesoamerican). The MaxEnt-based LDMs also indicated that 23 variables were important for the geographic prediction of landrace presence. Importantly, seven of these are non-climatic variables (Table S2.1), and amongst these, we find that accessibility and the geographic distribution of ethnic groups contribute substantially to the model.

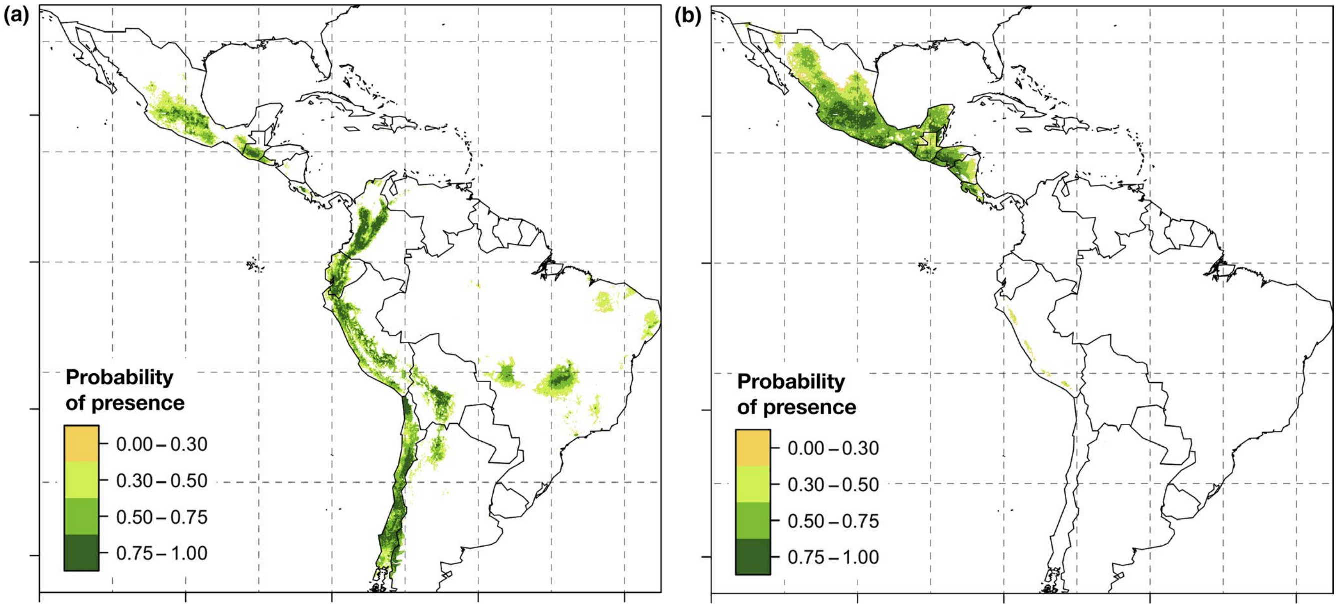

FIGURE 2 Predicted geographic distributions of Andean (a) and Mesoamerican (b) common bean (*Phaseolus vulgaris* L.) landrace groups

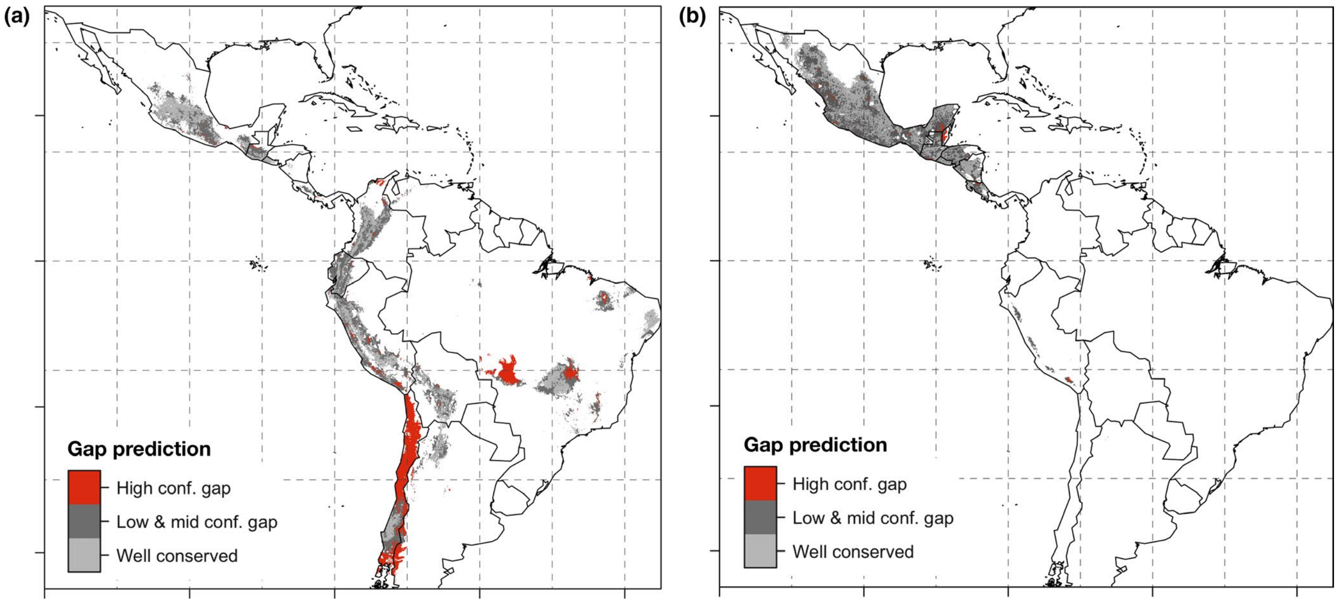

FIGURE 3 Final gaps map for Andean (a) and Mesoamerican (b) common bean (*Phaseolus vulgaris* L.) landrace groups. Red indicates areas where the three gap scores (S_{CON}, S_{ACC}, S_{ENV}) agree in identifying a gap

As expected, Andean landraces were predicted to be mostly distributed across the Andes mountains and to a lesser extent in Mexico and Central America. The converse was true for Mesoamerican landraces. Andean landraces were also predicted to occur in Brazil, which is considered a secondary diversity centre for common beans (Lobo Burle et al., 2010, 2011). Overlap was particularly evident in the geographic intermediate zone in Central America, (Beebe, Rengifo, Gaitan, Duque, & Tohme, 2001; Beebe et al., 2000) and in some areas of Peru.

3.3 | Conservation gap maps for common bean landraces

Conservation gap maps, displaying the overlap of results for the three gap scores per pixel, are shown in Figure 3. Figure S2.2 shows the individual gap scores, whereas Figure S2.3 shows model performance and coverage estimation. Overall gaps are larger for Andean compared to Mesoamerican beans, with representation of their distributions in genebanks estimated at 78.5% for the Andean and 98.2% for the Mesoamerican genepool. There is significant agreement amongst the gap areas identified by the accessibility, connectivity and environmental scores, and all performed well at predicting gaps.

For Andean beans, overlapping gap areas were found in the northern Venezuelan Andes, the Santander department in Colombia, specific pockets in the Andean hillsides between the Central and East cordillera in Colombia, the highlands of Ecuador, several areas in western and southern Peru, a major area in northern and central Chile, and central Brazil. In Mexico and Central America, noting that Andean bean variation is considered less diverse compared to South America (Becerra Velasquez & Gepts, 1994; Beebe et al., 2001), gaps were identified in the states of Oaxaca and to a lesser extent in Chiapas. Gaps were also predicted for Andean beans in Guatemala and Panama.

For the Mesoamerican genepool, the largest overlapping predicted gap was found in the area around Belize–Guatemala–southern Mexico (state of Campeche). Smaller overlapping gap areas were predicted in the states of San Luis Potosi, Jalisco and Sinaloa in Mexico. Across South America, southern Peru is predicted to be a gap.

3.4 | Expert inputs for common bean landrace distributions and conservation gaps

To illustrate how gap analysis results may be used to discuss collecting priorities, semi-structured interviews were carried out with two national and international *Phaseolus* scientists from the study region. One expert, Daniel G. Debouck (DGD), member of many collecting missions for the genus across many countries in the Americas, and expert in bean taxonomy, ecology, domestication and diversity and conservation (both in situ and ex situ) (Freytag & Debouck, 2002). He discussed both Andean and Mesoamerican beans for the entire Americas. The second expert was Eduardo Peralta (EP), a retired scientist, bean expert and breeder from Ecuador, pioneer in bean breeding in the Andean region that helped consolidate the National Legumes Program in Ecuador. He discussed Andean beans in the Andes. Detailed maps are shown in Figure S2.4.

Regarding areas of interest for collection for the Mesoamerican genepool, the experts indicated collecting should be prioritized in predicted gaps in San Luis Potosí, Oaxaca and Chiapas (Mexico), as well as in Belize and Ecuador. Notably, Ecuador is not predicted to be a gap by our method. For Andean landraces, the experts suggested collecting in the Venezuelan Andes and in the Santander department of Colombia. For the Colombian and Ecuadorian Andes, however, they indicated that collecting work would need to be done with precision (i.e. collecting only in specific sites and genotypes) rather than in an extensive manner.

Many areas were also identified by the two experts as unlikely to be considered collecting priorities. There were many areas, especially for Andean beans, where the experts indicated that it is likely that landraces are already lost due to traditional cropping practice replacement. This is the case in northern Chile and in southern and coastal Peru, where beans have been replaced by grape cropped for wine and pisco. Other areas were considered by experts to not be collecting priorities since these are mostly "documentation" gaps (e.g. central Brazil for Andean beans); this is because these materials are mostly in national collections, and passport information (including coordinates of collection sites) from these collections was not available or had insufficient quality for inclusion in our analyses.

4 | DISCUSSION

Here, we documented the development of a novel modelling framework to predict the distributions of crop landraces and to identify gaps in ex situ germplasm collections with relation to geographic and environmental variation in their distributions. We base our framework on the rationale that the distributions of landraces can be predicted using environmental and socioeconomic drivers, and that important conservation gaps can be identified by characterizing the geographic (accessibility and connectivity) and environmental space across which previous collecting has been carried out. Previous studies assessing gaps in landrace collections only used climate drivers and did not explicitly assess gap prediction robustness (Upadhyaya et al., 2012, 2017; Upadhyaya, Reddy, Irshad Ahmed, Gowda, & Haussmann, 2010) nor introduce expert inputs to prioritize collecting.

Our analysis suggests that both genepools of *P. vulgaris* are relatively well conserved and that progress towards comprehensive representation ex situ may be relatively fast if targeted collecting is performed in the areas outlined in the results. This contrasts with results for common bean wild relatives, for which research indicates that about two-thirds of the wild species in the genus need further conservation action, and about half are considered high priority for further collecting (Castañeda-Álvarez et al., 2016; Ramirez-Villegas et al., 2010).

For Andean beans, gaps were predicted throughout most bean-producing countries in South America, with the highest priority being Chile, Peru, Colombia and specific spots in the Venezuelan Andes. For Mesoamerican landraces, the results target regions of Mexico, Belize, Guatemala and to a lesser extent South America (mostly Peru) for further collecting. While current common bean collections already hold substantial diversity from across the Americas (Beebe et al., 2000, 2001), our results, supplemented by expert opinion, indicated that further collecting is warranted, especially where valuable traits such as phosphorous use efficiency (Beebe, Lynch, Galwey, Tohme, & Ochoa, 1997) or heat stress tolerance (CGIAR, 2015) may be found.

Our ongoing review of other crop landraces indicates that the classification approach, based on recognized groups, can be widely applicable to other crops (van Heerwaarden et al., 2011; Lasky et al., 2015; Ndjiondjop et al., 2018). Moreover, the continuous generation of new genetic diversity data and related knowledge

(Crossa et al., 2016; Halewood et al., 2018) will facilitate the further application of our methods, which are ultimately dependent on the availability of robust classification, occurrence and characterization data.

While our framework contributes to revealing existing gaps in current germplasm collections and to highlighting geographic areas where novel diversity may be collected, the question remains as to the extent to which the results can support on-the-ground collecting work. Our discussion with experts indicates that priorities for collecting can be drawn using our predicted gap maps. Moreover, previous ecogeographic analyses have proven useful for collecting planning (García et al., 2017; Jarvis et al., 2005; Marinoni et al., 2015). To further translate our results for action, designing tools for real-time collecting mission support (e.g. route tracing) that combine the outputs with existing technologies for map visualization and navigation would be advantageous.

4.1 | Challenges and limitations to landrace distribution modelling and conservation gap analysis

Predicting the distributions of cultivated plants, whose ranges are determined by anthropogenic along with environmental drivers, presents a challenge that has not been fully resolved in geospatial sciences. While we attempted to gather the widest range of quality input occurrence and predictor data and used state of the art approaches to ensure high species distribution model (SDM) performance, several further improvements can be suggested.

With regard to occurrence information, particularly for genebank collections, we incorporated data from the two central global repositories for such information (Genesys and WIEWS) and in addition (due to our focus here on common bean) insured the full compilation of data from the world's two largest *P. vulgaris* collections (CIAT and USDA). This said, these sources are not fully representative of all common bean collections worldwide, including collections such as the Agricultural Research Institute (CIAP) in Cuba. Ongoing initiatives, such as Genesys that list in a single location passport and (eventually) characterization data for many genebanks (Global Crop Diversity Trust, 2019), may help resolve this data challenge in the future. On the other hand, national policies influencing germplasm distributions hinder the international accessibility of many such "low-visibility" collections (Castiñeiras, Esquive, Lioi, & Hammer, 1991; Lobo Burle et al., 2011).

We also note that coordinate information, which is an essential input into our methods, is missing for many current genebank accessions. Further efforts to georeference records missing coordinates but possessing locality information, and to make this information easily available online, will facilitate a more robust assessment of the state of conservation of crop landraces ex situ.

Distributions of crop landraces are influenced by factors beyond the environmental and socioeconomic predictors used here. These may include other abiotic (e.g. soil parent material and other edaphic characteristics), biotic (e.g. mycorrhizae, pathogens and pollinators), and agriculturally relevant socioeconomic (e.g. farm sizes and farming systems) factors. Further development of high-resolution global datasets will be needed to incorporate such information into our analyses. Similarly, we note that model uncertainty can be a challenge and highlight the need to use model results as a "discussion support" tool to prioritize collecting. Finally, while we employ a widely used distribution modelling algorithm, it is possible that incorporating other methods, or forming ensembles of multiple methods, could improve our prediction of gaps (Grenouillet, Buisson, Casajus, & Lek, 2011).

4.2 | Landrace conservation gap analysis for global targets

The high value of crop landrace diversity in breeding programmes and for farm-level resilience (Camacho Villa et al., 2005; van Etten et al., 2019; van de Wouw et al., 2010), and the evident erosion of these resources in their primary and secondary centres of diversity (van Heerwaarden et al., 2009; Mekbib, 2008) justify urgent action to secure ex situ the diversity of landrace still cultivated by farmers and in addition (though not discussed in this article) to invest in farmer-based (i.e. in situ/on farm) conservation (Bellon, Dulloo, Sardos, Thormann, & Burdon, 2017). The United Nations Sustainable Development Goal (SDG) 2.5, the Convention on Biological Diversity (CBD), Strategic Plan for Biodiversity 2011–2020, Aichi Biodiversity Target 13 (CBD, 2010a) and Global Strategy for Plant Conservation (GSPC) Target 9 (CBD, 2010b) and Article 5 of the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) (FAO, 2002) all discuss and/or establish targets for the maintenance of genetic diversity of cultivated plants and their wild relatives, both in situ and ex situ. Recently, Khoury et al. (2019) proposed an indicator to track the conservation of useful wild plants, which furthers tested gap analysis methodologies for wild flora (Ramirez-Villegas et al., 2010). Here, we developed a coverage metric that, if implemented for a sufficiently large number of crops, could be used to track progress towards the conservation of cultivated plants for SDG 2.5, Aichi 13 and other important international goals.

ACKNOWLEDGEMENTS

This work was carried out under the CGIAR Genebanks Platform (https://www.genebanks.org). The CGIAR Genebanks Platform enables CGIAR Research Centers to fulfil their legal obligation to conserve and make available accessions of crops and trees on behalf of the global community under the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA). Authors thank Andy Jarvis from CIAT, Chris Richards from USDA and Paul Evangelista from Colorado State University for input on the methodology during early stages of this project, and Angela M. Hernández from CIAT for providing bean passport and phenotypic characterization data for the analyses. Authors also thank CGIAR Genebank

Managers and Scientists present at the Annual Genebanks Meetings (AGM) in 2017 (Brussels) and 2018 (Fortaleza) for their feedback on the methodology.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Occurrence and predictor data are provided in Dataset S1. The data used as input of this research are available open access at the original sources, all of which are cited in the respective sections of the text. The R code for performing the analyses is available at: https://github.com/CIAT-DAPA/gap_analysis_landraces.

ORCID

Julian Ramirez-Villegas https://orcid.org/0000-0002-8044-583X
Colin K. Khoury https://orcid.org/0000-0001-7893-5744

REFERENCES

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Becerra Velasquez, V. L., & Gepts, P. (1994). RFLP diversity of common bean (*Phaseolus vulgaris*) in its centres of origin. *Genome*, *37*, 256–263.

Beebe, S. (2012). *Common bean breeding in the tropics. Plant breeding reviews* (pp. 357–426). Hoboken, NJ, USA: John Wiley & Sons Inc.

Beebe, S., Lynch, J., Galwey, N., Tohme, J., & Ochoa, I. (1997). A geographical approach to identify phosphorus-efficient genotypes among landraces and wild ancestors of common bean. *Euphytica*, *95*, 325–338.

Beebe, S., Rengifo, J., Gaitan, E., Duque, M. C., & Tohme, J. (2001). Diversity and Origin of Andean Landraces of Common Bean. *Crop Science*, *41*, 854. https://doi.org/10.2135/cropsci2001.413854x

Beebe, S., Skroch, P. W., Tohme, J., Duque, M. C., Pedraza, F., & Nienhuis, J. (2000). Structure of genetic diversity among common bean landraces of Middle American origin based on correspondence analysis of RAPD. *Crop Science*, *40*, 264. https://doi.org/10.2135/cropsci2000.401264x

Bellon, M. R., Dulloo, E., Sardos, J., Thormann, I., & Burdon, J. J. (2017). In situ conservation-harnessing natural and human-derived evolutionary forces to ensure future crop adaptation. *Evolutionary Applications*, *10*, 965–977. https://doi.org/10.1111/eva.12521

Blair, M. W., Díaz, J. M., Hidalgo, R., Díaz, L. M., & Duque, M. C. (2007). Microsatellite characterization of Andean races of common bean (*Phaseolus vulgaris* L.). *Theoretical and Applied Genetics*, *116*, 29–43. https://doi.org/10.1007/s00122-007-0644-8

Blair, M. W., Díaz, L. M., Buendía, H. F., & Duque, M. C. (2009). Genetic diversity, seed size associations and population structure of a core collection of common beans (*Phaseolus vulgaris* L.). *Theoretical and Applied Genetics*, *119*, 955–972. https://doi.org/10.1007/s00122-009-1064-8

Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H., & Bollmann, K. (2013). Selecting from correlated climate variables: A major source of uncertainty for predicting species distributions under climate change. *Ecography*, *36*, 971–983. https://doi.org/10.1111/j.1600-0587.2013.00138.x

Broughton, W. J., Hernández, G., Blair, M., Beebe, S., Gepts, P., & Vanderleyden, J. (2003). Beans (*Phaseolus* spp.) – model food legumes. *Plant and Soil*, *252*, 55–128. https://doi.org/10.1023/A:1024146710611

Burke, M. B., Lobell, D. B., & Guarino, L. (2009). Shifts in African crop climates by 2050, and the implications for crop improvement and genetic resources conservation. *Global Environmental Change*, *19*, 317–325. https://doi.org/10.1016/j.gloenvcha.2009.04.003

Camacho Villa, T. C., Maxted, N., Scholten, M., & Ford-Lloyd, B. (2005). Defining and identifying crop landraces. *Plant Genetic Resources: Characterization and Utilization*, *3*, 373–384. https://doi.org/10.1079/PGR200591

Casañas, F., Simó, J., Casals, J., & Prohens, J. (2017) Toward an evolved concept of landrace. *Frontiers in Plant Science*, *8*, 145. https://doi.org/10.3389/fpls.2017.00145

Castañeda-Álvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., … Toll, J. (2016). Global conservation priorities for crop wild relatives. *Nature Plants*, *2*, 16022. https://doi.org/10.1038/nplants.2016.22

Castiñeiras, L., Esquive, M., Lioi, L., & Hammer, K. (1991). Origin, diversity and utilization of the Cuban germplasm of common bean (*Phaseolus vulgaris* L.). *Euphytica*, *57*, 1–8.

CBD (2010a) Aichi biodiversity targets. https://www.cbd.int/sp/targets/. Accessed August 8, 2018.

CBD (2010b) Global strategy for plant conservation. The targets, 2011–2020. https://www.cbd.int/gspc/targets.shtml. Accessed August 8, 2018.

CGIAR (2015). *Developing beans that can beat the heat*. Cali, Colombia: CGIAR.

Chacon, M. I., Pickersgill, B., & Debouck, D. G. (2005). Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. *Theoretical and Applied Genetics*, *110*, 432–444. https://doi.org/10.1007/s00122-004-1842-2

CIAT (2018). CIAT Genebank Database, version 2018.

CIESIN (2018). *Gridded population of the world, version 4 (GPWv4): Population density, revision 11*. Palisades, USA.

Cortes, A. J., Monserrate, F. A., Ramirez-Villegas, J., Madrinan, S., & Blair, M. W. (2013). Drought tolerance in wild plant populations: The case of common beans (*Phaseolus vulgaris* L.). *PLoS ONE*, *8*, e62898. https://doi.org/10.1371/journal.pone.0062898

Costa, G., Nogueira, C., Machado, R., & Colli, G. (2010). Sampling bias and the use of ecological niche modeling in conservation planning: A field evaluation in a biodiversity hotspot. *Biodiversity and Conservation*, *19*, 883–899. https://doi.org/10.1007/s10531-009-9746-8

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., … Singh, S. (2016). Genomic prediction of gene bank wheat landraces. *G3, Genes|genomes|genetics*, *6*, 1819–1834. https://doi.org/10.1534/g3.116.029637

Debouck, D. G. (2014). *Conservation of Phaseolus beans genetic resources: A strategy*. Rome, Italy: Global Crop Diversity Trust.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*, 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Elith, J., Graham, C., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., … Zimmermann, N. (2006). Novel methods improve prediction of species distributions from occurrence data. *Ecography*, *29*, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2010). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*, 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Escribano, M., & De Ron, A. M. (1991). Taxonomical relationships among common bean populations from northern Spain. *Anales de la Estación Experimental de Aula Dei*, *20*, 17–27.

Esquinas-Alcázar, J. (2005). Protecting crop genetic diversity for food security: Political, ethical and technical challenges. *Nature Reviews Genetics*, *6*, 946–953. https://doi.org/10.1038/nrg1729

FAO (2002). *The international treaty on plant genetic resources for food and agriculture*. Rome, Italy: FAO.

FAO (2019). *United Nations food and agriculture organization world information and early warning system on plant genetic resources for food and Agriculture (WIEWS)*. Rome, Italy: Food and Agriculture Organization of the United Nations (FAO).

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*, 4302–4315. https://doi.org/10.1002/joc.5086

Freytag, G. F., & Debouck, D. G. (2002). *Taxonomy, distribution, and ecology of the genus* Phaseolus *in North America, Mexico and Central America*. Forth Worth, TX, USA: Botanical Research Institute

Fuller, D. Q. (2007). Contrasting patterns in crop domestication and domestication rates: Recent archaeobotanical insights from the old world. *Annals of Botany*, *100*, 903–924. https://doi.org/10.1093/aob/mcm048

García, R. M., Parra-Quijano, M., & Iriondo, J. M. (2017). A multispecies collecting strategy for crop wild relatives based on complementary areas with a high density of ecogeographical gaps. *Crop Science*, *57*, 1059. https://doi.org/10.2135/cropsci2016.10.0860

GBIF.org (2019). Global Biodiversity Information Facility (GBIF) occurrence download.

Gepts, P. (2006). Plant genetic resources conservation and utilization. *Crop Science*, *46*, 2278.

Gepts, P., Osborn, T. C., Rashka, K., & Bliss, F. A. (1986). Phaseolin-protein variability in wild forms and landraces of the common bean (*Phaseolus vulgaris*): Evidence for multiple centers of domestication. *Economic Botany*, *40*, 451–468. https://doi.org/10.1007/BF02859659

Giovanelli, J. G. R., de Siqueira, M. F., Haddad, C. F. B., & Alexandrino, J. (2010). Modeling a spatially restricted distribution in the Neotropics: How the size of calibration area affects the performance of five presence-only methods. *Ecological Modelling*, *221*, 215–224. https://doi.org/10.1016/j.ecolmodel.2009.10.009

Glaszmann, J., Kilian, B., Upadhyaya, H., & Varshney, R. (2010). Accessing genetic diversity for crop improvement. *Current Opinion in Plant Biology*, *13*, 167–173. https://doi.org/10.1016/j.pbi.2010.01.004

Global Crop Diversity Trust (2019) *Genesys-PGR: A gateway to genetic resources*. Bonn, Germany: Global Crop Diversity Trust

Greene, S. L., Hart, T. C., & Afonin, A. (1999a). Using geographic information to acquire wild crop germplasm for ex situ collections: I. Map development and field use. *Crop Science*, *39*, 836. https://doi.org/10.2135/cropsci1999.0011183X003900030037x

Greene, S. L., Hart, T. C., & Afonin, A. (1999b). Using geographic information to acquire wild crop germplasm for ex situ collections: II. Post-collection analysis. *Crop Science*, *39*, 843. https://doi.org/10.2135/cropsci1999.0011183X003900030038x

Grenouillet, G., Buisson, L., Casajus, N., & Lek, S. (2011). Ensemble modelling of species distribution: The effects of geographical and environmental ranges. *Ecography*, *34*, 9–17. https://doi.org/10.1111/j.1600-0587.2010.06152.x

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In: R. Meersman, Z. Tari, D.C. Schmidt (eds) *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science*, *2888*. (pp. 986–996). Berlin, Heidelberg: Springer.

Hajjar, R., & Hodgkin, T. (2007). The use of wild relatives in crop improvement: A survey of developments over the last 20 years. *Euphytica*, *156*, 1–13. https://doi.org/10.1007/s10681-007-9363-0

Halewood, M., Lopez Noriega, I., Ellis, D., Roa, C., Rouard, M., & Sackville Hamilton, R. (2018). Using genomic sequence information to increase conservation and sustainable use of crop diversity and benefit-sharing. *Biopreservation and Biobanking*, *16*, 368–376. https://doi.org/10.1089/bio.2018.0043

Hammer, K., Knüpffer, H., Xhuveli, L., & Perrino, P. (1996). Estimating genetic erosion in landraces — two case studies. *Genetic Resources and Crop Evolution*, *43*, 329–336. https://doi.org/10.1007/BF00132952

Harlan, J. R. (1975). Our vanishing genetic resources. *Science*, *188*, 617–621. https://doi.org/10.1126/science.188.4188.617

Hilbert, L., Neves, E. G., Pugliese, F., Whitney, B. S., Shock, M., Veasey, E., … Iriarte, J. (2017). Evidence for mid-Holocene rice domestication in the Americas. *Nature Ecology & Evolution*, *1*, 1693–1698. https://doi.org/10.1038/s41559-017-0322-4

Hoisington, D., Khairallah, M., Reeves, T., Ribaut, J.-M., Skovmand, B., Taba, S., & Warburton, M. (1999). Plant genetic resources: What can they contribute toward increased crop productivity? *Proceedings of the National Academy of Sciences of the United States of America*, *96*, 5937–5943. https://doi.org/10.1073/pnas.96.11.5937

Jarvis, A., Reuter, H. I., Nelson, A., & Guevara, E. (2008). Hole-filled seamless SRTM data V4.

Jarvis, A., Williams, K., Williams, D., Guarino, L., Caballero, P., & Mottram, G. (2005). Use of GIS for optimizing a collecting mission for a rare wild pepper (*Capsicum flexuosum* Sendtn.) in Paraguay. *Genetic Resources and Crop Evolution*, *52*, 671–682. https://doi.org/10.1007/s10722-003-6020-x

Jones, H., Lister, D. L., Bower, M. A., Leigh, F. J., Smith, L. M., & Jones, M. K. (2008). Approaches and constraints of using existing landrace and extant plant material to understand agricultural spread in prehistory. *Plant Genetic Resources: Characterization and Utilization*, *6*, 98–112. https://doi.org/10.1017/S1479262108993138

Khoury, C. K., Achicanoy, H. A., Bjorkman, A. D., Navarro-Racines, C., Guarino, L., Flores-Palacios, X., … Struik, P. C. (2016). Origins of food crops connect countries worldwide. *Proceedings of the Royal Society B: Biological Sciences*, *283*, 20160792. https://doi.org/10.1098/rspb.2016.0792

Khoury, C. K., Amariles, D., Soto, J. S., Diaz, M. V., Sotelo, S., Sosa, C. C., … Jarvis, A. (2019). Comprehensiveness of conservation of useful wild plants: An operational indicator for biodiversity and sustainable development targets. *Ecological Indicators*, *98*, 420–429. https://doi.org/10.1016/j.ecolind.2018.11.016

Lasky, J. R., Upadhyaya, H. D., Ramu, P., Deshpande, S., Hash, C. T., Bonnette, J., … Morris, G. P. (2015). Genome-environment associations in sorghum landraces predict adaptive traits. *Science Advances*, *1*, e1400218–e1400218. https://doi.org/10.1126/sciadv.1400218

Lee, D. T., & Schachter, B. J. (1980). Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer & Information Sciences*, *9*, 219–242. https://doi.org/10.1007/BF00977785

Lobo Burle, M. L., Fonseca, J. R., Jose del Peloso, M., Melo, L. C., Temple, S. R., & Gepts, P. (2011). Integrating phenotypic evaluations with a molecular diversity assessment of a Brazilian collection of common bean landraces. *Crop Science*, *51*, 2668. https://doi.org/10.2135/cropsci2010.12.0710

Lobo Burle, M., Fonseca, J. R., Kami, J. A., & Gepts, P. (2010). Microsatellite diversity and genetic structure among common bean (*Phaseolus vulgaris* L.) landraces in Brazil, a secondary center of diversity. *Theoretical and Applied Genetics*, *121*, 801–813. https://doi.org/10.1007/s00122-010-1350-5

Logozzo, G., Donnoli, R., Macaluso, L., Papa, R., Knüpffer, H., & Zeuli, P. S. (2007). Analysis of the contribution of mesoamerican and andean gene pools to European common bean (*Phaseolus vulgaris* L.) germplasm and strategies to establish a core collection. *Genetic Resources and Crop Evolution*, *54*, 1763–1779. https://doi.org/10.1007/s10722-006-9185-2

Marinoni, L., Bortoluzzi, A., Parra-Quijano, M., Zabala, J. M., & Pensiero, J. F. (2015). Evaluation and improvement of the ecogeographical representativeness of a collection of the genus Trichloris in Argentina.

Genetic Resources and Crop Evolution, *62*, 593–604. https://doi.org/10.1007/s10722-014-0184-4

Maxted, N., Dulloo, E., Ford-Lloyd, V. B., Iriondo, J. M., & Jarvis, A. (2008). Gap analysis: A tool for complementary genetic conservation assessment. *Diversity and Distributions*, *14*, 1018–1030. https://doi.org/10.1111/j.1472-4642.2008.00512.x

Mekbib, F. (2008). Genetic erosion of sorghum (*Sorghum bicolor* (L.) Moench) in the centre of diversity, Ethiopia. *Genetic Resources and Crop Evolution*, *55*, 351–364.

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, *55*, 169–186. https://doi.org/10.1016/S0925-2312(03)00431-4

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., … Kresovich, S. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 453–458. https://doi.org/10.1073/pnas.1215985110

Natural Earth (2019). *Rivers + lake centerlines*.

Ndjiondjop, M. N., Semagn, K., Sow, M., Manneh, B., Gouda, A. C., Kpeki, S. B., … Warburton, M. L. (2018) Assessment of genetic variation and population structure of diverse rice genotypes adapted to lowland and upland ecologies in Africa using SNPs. *Frontiers in Plant Science*, *9*, 446. https://doi.org/10.3389/fpls.2018.00446

Nelson, A. (2008) A global map of Accessiblity. Published by the Office for Official Publications of the European Communities. Luxembourg. https://doi.org/10.2788/95835. ISBN: 978-92-79-09771-3

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*, 217–222. https://doi.org/10.1080/01431160412331269698

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography*, *40*, 887–893. https://doi.org/10.1111/ecog.03049

Phillips, S., Anderson, R., & Schapire, R. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Ramirez-Villegas, J., Khoury, C., Jarvis, A., Debouck, D. G., & Guarino, L. (2010). A gap analysis methodology for collecting crop genepools: A case study with *Phaseolus* beans. *PLoS ONE*, *5*, e13497. https://doi.org/10.1371/journal.pone.0013497

Reuter, H. I., Nelson, A., & Jarvis, A. (2007). An evaluation of void-filling interpolation methods for SRTM data. *International Journal of Geographical Information Science*, *21*, 983–1008. https://doi.org/10.1080/13658810601169899

Rodrigues, A. S. L., Akçakaya, H. R., Andelman, S. J., Bakarr, M. I., Boitani, L., Brooks, T. M., … Yan, X. (2004). Global gap analysis: Priority regions for expanding the global protected-area network. *BioScience*, *54*, 1092–1100. https://doi.org/10.1641/0006-3568(2004)054[1092:GGAPRF]2.0.CO;2

Sayre, R., Dangermond, J., Frye, C., Vaughan, R., Aniello, P., Breyer, S., … Comer, P. (2014). *A new map of global ecological land units — An ecophysiographic stratification approach*. Washington, DC.

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PLoS ONE*, *8*, e71218. https://doi.org/10.1371/journal.pone.0071218

Siebert, S., Henrich, V., Frenken, K., & Burke, J. (2013) Update of the Global Map of Irrigation Areas to version 5. Project report.

Singh, S. (1989). Patterns of variation in cultivated common bean (*Phaseolus vulgaris* Fabaceae). *Economic Botany*, *43*, 39–57.

Singh, S., Gepts, P., & Debouck, D. (1991). Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Economic Botany*, *45*, 379–396. https://doi.org/10.1007/BF02887079

Syfert, M. M., Castaneda-Alvarez, N. P., Khoury, C. K., Sarkinen, T., Sosa, C. C., Achicanoy, H. A., … Knapp, S. (2016). Crop wild relatives of the brinjal eggplant (*Solanum melongena*): Poorly represented in genebanks and many species at risk of extinction. *American Journal of Botany*, *103*, 635–651.

Title, P. O., & Bemmels, J. B. (2018). ENVIREM: An expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, *41*, 291–307. https://doi.org/10.1111/ecog.02880

Turner, R. (2019). deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation. R package version 0.1-16.

Upadhyaya, H. D., Reddy, K. N., Irshad Ahmed, M., & Gowda, C. L. L. (2012). Identification of gaps in pearl millet germplasm from East and Southern Africa conserved at the ICRISAT genebank. *Plant Genetic Resources*, *10*, 202–213. https://doi.org/10.1017/S1479262112000275

Upadhyaya, H. D., Reddy, K. N., Irshad Ahmed, M., Gowda, C. L. L., & Haussmann, B. I. G. (2010). Identification of geographical gaps in the pearl millet germplasm conserved at ICRISAT genebank from West and Central Africa. *Plant Genetic Resources*, *8*, 45–51. https://doi.org/10.1017/S147926210999013X

Upadhyaya, H. D., Reddy, K. N., Vetriventhan, M., Krishna Gumma, M., Irshad Ahmed, M., Thimma Reddy, M., & Singh, S. K. (2017). Status, genetic diversity and gaps in sorghum germplasm from South Asia conserved at ICRISAT genebank. *Plant Genetic Resources*, *15*, 527–538. https://doi.org/10.1017/S147926211600023X

USDA ARS NPGS (2018). USDA GRIN Global.

van de Wouw, M., Kik, C., van Hintum, T., van Treuren, R., & Visser, B. (2010). Genetic erosion in crops: Concept, research results and challenges. *Plant Genetic Resources*, *8*, 1–15. https://doi.org/10.1017/S1479262109990062

van Etten, J., de Sousa, K., Aguilar, A., Barrios, M., Coto, A., Dell'Acqua, M., … Steinke, J. (2019). Crop variety management for climate adaptation supported by citizen science. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 4194–4199. https://doi.org/10.1073/pnas.1813720116

van Heerwaarden, J., Doebley, J., Briggs, W. H., Glaubitz, J. C., Goodman, M. M., de Jesus Sanchez Gonzalez, J. & Ross-Ibarra, J. (2011). Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 1088–1092. https://doi.org/10.1073/pnas.1013011108

van Heerwaarden, J., Hellin, J., Visser, R., & van Eeuwijk, F. (2009). Estimating maize genetic erosion in modernized smallholder agriculture. *TAG Theoretical and Applied Genetics*, *119*, 875–888. https://doi.org/10.1007/s00122-009-1096-0

van Treuren, R., Engels, J. M. M., Hoekstra, R., & van Hintum, T. J. L. (2009). Optimization of the composition of crop collections for ex situ conservation. *Plant Genetic Resources*, *7*, 185–193. https://doi.org/10.1017/S1479262108197477

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. *Ecological Applications*, *21*, 335–342. https://doi.org/10.1890/10-1171.1

Weidmann, N. B., Rød, J. K., & Cederman, L.-E. (2010). Representing ethnic groups in space: A new dataset. *Journal of Peace Research*, *47*, 491–499. https://doi.org/10.1177/0022343310368352

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., … Gething, P. W. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, *553*, 333–336. https://doi.org/10.1038/nature25181

You, L., Wood-Sichra, U., Fritz, S., Guo, Z., See, L., & Koo, J. (2017). Spatial Production Allocation Model (SPAM) 2005 v3.2.

Zeven, A. C. (1998). Landraces: A review of definitions and classifications. *Euphytica*, *104*, 127–139.

BIOSKETCH

Dr. Julian Ramirez-Villegas is a senior scientist at the International Centre for Tropical Agriculture (CIAT), based at their headquarters in Cali, Colombia. He has a PhD from the School of Earth and Environment at the University of Leeds (UK). He leads a group on agricultural and climate modelling, doing research on climate information services, climate change impacts and adaptation, and crop agro-biodiversity conservation. The research presented here is part of the "Landrace Gap Analysis of 22 Crops" activity, under the Conservation Module of the CGIAR Genebanks Platform (https://www.genebanks.org).

Authors Contributions: J.R.-V., C.K.K., L.G. and Z.K. conceived the idea and designed the modelling framework. H.A.A., A.C.M., M.V.D. and C.C.S. collected and analysed the data. D.G.D. provided critical inputs to correctly interpret results and to adjust models. J.R.-V. and C.K.K. wrote the manuscript. All authors reviewed and edited the manuscript.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Ramirez-Villegas J, Khoury CK, Achicanoy HA, et al. A gap analysis modelling framework to prioritize collecting for ex situ conservation of crop landraces. *Divers Distrib*. 2020;00:1–13. https://doi.org/10.1111/ddi.13046