Global conservation priorities for crop wild relatives

Nora P. Castañeda-Álvarez, Colin K. Khoury, Harold A. Achicanoy, Vivian Bernau, Hannes Dempewolf, Ruth J. Eastwood, Luigi Guarino, Ruth H. Harker, Andy Jarvis, Nigel Maxted, Jonas V. Müller, Julian Ramirez-Villegas, Chrystian C. Sosa, Paul C. Struik, Holly Vincent, & Jane Toll

Supplementary Methods

Selection of crops and identification of wild related taxa. The 81 crops included in this study were selected according to their importance to food security (as measured in terms of importance in food supply and agricultural production systems worldwide[34]), acknowledged relevance to income generation for smallholder farmers, and/or inclusion in the Multilateral System of Access and Benefit Sharing (as listed in the Annex 1) of the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA)[29], whose objectives are to enhance conservation, use and benefit sharing of plant genetic resources for food security and sustainable agriculture (Supplementary Table 1). Lack of sufficient data led to the exclusion of analyses for the genepools of coffee, tea, and avocado; all other targeted crop genepools were included.

For the comparative analyses of further collecting priorities per crop genepool and crop type (Fig. 3; Supplementary Fig. 2), a mean importance score for each crop was produced based on contribution to four measures of global aggregate food supply [calories (kcal/capita/day), protein (g/capita/day), fat (g/capita/day), and food weight (g/capita/day)], and three measures of total global agricultural production [production quantity (tonnes), harvested area (ha), and production value (current million $USD)], using FAO statistical data[34] data averaged over three recent years

(2009-2011). Values for crops listed within aggregated commodity classes in food supply data were disaggregated by dividing the total aggregated commodity values equally across all listed crops, aside from the sugar commodity, for which 70% of value was assigned to sugarcane, and 30% to sugar beet[35]. The importance score was produced for each crop by first dividing its food supply/agricultural production metric by the maximum existing value across all crops per metric. The four food supply and three agricultural production metrics were then averaged separately, with a final importance score derived by averaging the mean food supply and mean agricultural production values. This score is presented on a scale from 0 (low importance) to 10 (high importance).

The wild relatives of the 81 crops were selected according to their potential to cross with their associated domesticated crop species. This potential was estimated using the crop genepool concept[36], which categorizes crop wild relatives by primary, secondary and tertiary genepools. The first two categories include taxa that have potential to produce viable offspring, although technical tools (e.g., embryo rescue, chromosome doubling) may be required. Introgression of genes from taxa in the tertiary genepool generally requires more advanced biotechnological tools[31,36]. Taxonomic relationships were used as proxy when crossability information was not available[19,37]. In this analysis, we included taxa listed in the primary and secondary genepools of target crops, as well as any more distantly related taxa with confirmed and/or potential utilization in crop breeding reported in the literature[19].

Data gathering and preparation. Reference data (i.e., occurrence information associated with specimens collected in botanical expeditions, reported sightings and historical or inactive

germplasm accessions) for crop wild relatives were acquired from online biodiversity, herbarium, and genebank databases; through communications with herbarium and genebank managers and crop researchers; scientific literature; and via direct recording of provenance data during visits to selected herbaria (Supplementary Table 3). Germplasm accession data was obtained from digital repositories that provide access to genetic resources and associated data to the global research and crop breeding community through online information systems. The occurrence data were then compiled in a standardized format, and recognized duplicate records were deleted. Nomenclature was verified against GRIN Taxonomy for Plants[38]; the "Taxonomic Name Resolution Service (TNRS)[39], and The Plant List[40,41], with GRIN Taxonomy used as the preferred nomenclature in the case of more than one proposed accepted name. Existing coordinates were verified as being on land and within the reported country of collection following Warren et al.[42]. If an inconsistency was detected, coordinates were recalculated for records with detailed locality descriptions, or deleted for records with insufficient locality data. Records with locality information but no coordinates were geo-referenced using an automated system based on the Google Maps Geocoder V.3 application programming interface. Occurrence data were mapped, iteratively evaluated for correctness, and further processed in order to form a final dataset of improved taxonomic and spatial accuracy. In total, we utilized 768,298 occurrence records.

Species distribution modelling. We used the MaxEnt method for calculating species distribution models due to its wide application in conservation studies and ability to discriminate the environmental niches where a species is likely to occur[43]. MaxEnt is a machine-learning algorithm that uses the maximum entropy principle to estimate the suitability of a species to

occur in certain environment[44]. We trained our models using the default settings of MaxEnt as they are considered adequate for studies at the global level[44]. We used as inputs all the unique occurrence records (germplasm accessions and reference data combined) with verified coordinates, and the nineteen bioclimatic variables from the Worldclim database[45] (Supplementary Table 5). These bioclimatic variables are produced using high resolution global layers based upon monthly temperature and precipitation information from terrestrial meteorological stations and satellite-derived images (e.g., The Shuttle Radar Topography Mission and GTOPO3)[45]. Ongoing improvements to global soils, land cover, topography, and other spatial datasets will in future analyses help to further refine such species distribution models.

The background extent for each taxon was defined by overlapping the occurrence records of the taxon on a global map divided into six regions (e.g., North America, South America, Europe, Asia, Africa and Oceania). Any regions containing occurrence records were included in the background extent. This broad background characterizes the environment within the study area of each taxon, and tends to produce optimistic predictions of the potential geographic distributions of species, which may be useful particularly in encouraging the discovery of unrecorded populations on or beyond the known edges of distribution[46,47]. Ten thousand random background points were created for each taxon within the background extent and were used for training each model.

Trained models were projected onto bioclimatic layers at a spatial resolution of 2.5 min (ca. 4.6 km × 4.6 km at the equator) to estimate the potential distributions of taxa. Adapted from

Ramirez-Villegas et al.[25], we used the cross-validation option (k=5) for assessing the accuracy and adequacy of each model for use in the gap analysis. Models considered adequate for the gap analysis assessment met the following conditions: i) the five-fold average of the test sample Area Under the Receiver Operating Characteristic (ROC) curve (ATAUC) was greater than 0.7, ii) the standard deviation of the ATAUC for the five different folds was lower than 0.15, and iii) the proportion of the potential distribution where the standard deviation was greater than 0.15 was less than 10%. We used the shortest distance to the upper left corner of the ROC curve[48] as the threshold to produce binomial (presence-absence) distribution maps. Each potential distribution model was further restricted by clipping it to its known native distribution (at the country level) as reported in the literature[38], or to a convex hull (i.e., a polygon surrounding the outermost occurrence records for the taxon under analysis) when native distribution descriptions were not available. For the taxa whose MaxEnt models did not meet the three-fold validation, a convex hull was used to represent the potential distribution model in the gap analysis assessment. When neither a MaxEnt model nor a convex hull were produced due to lack of data (i.e., only one or two unique records with coordinates available), a circular buffer of 50 km around each coordinate was used to estimate potential distribution[49].

Assessing the extent of representation of crop wild relatives in genebanks. We derived three quantitative measures to determine the extent of representation of taxa in genebanks: the sampling representativeness score (SRS), the geographical representativeness score (GRS); and the ecological representativeness score (ERS). Adapted from the gap analysis method of Ramirez-Villegas et al.[25], all measures were fit in a numeric range between zero and ten.

The SRS is a general indicator of sufficiency of accessions in genebanks, comparing the total number of reference records against the current number of germplasm accessions available in genebanks[25]. The SRS provides a gross estimation of sufficiency, with the benefits of making use of all compiled reference and germplasm records (regardless of whether or not they possess verified geographical coordinates), and providing a general sufficiency metric relative to the extent of distribution of taxa (estimated by total number of records), as the number of accessions sufficient to capture the diversity of a taxon is partly dependent upon the extent of distribution of the taxon. The GRS and ERS estimate geographic and ecological variation encompassed in genebank collections (i.e., within a 50 km radius surrounding the original site of collection of each current genebank accession) in comparison to the variation exhibited in the potential distribution models of taxa[25]. For the purpose of our analysis, the layer used for estimating the ERS contained 867 distinct terrestrial ecoregions[50] as a proxy for ecological diversity and potential adaptation to distinct ecological characteristics.

A final priority score (FPS) was produced by averaging SRS, ERS and GRS per taxon.
High priority for further collecting to improve representation in genebanks was assigned for taxa where FPS $\geq$ 7; medium priority where 5 $\leq$ FPS < 7; low priority where 2.5 $\leq$ FPS < 5; and sufficiently represented for taxa whose FPS < 2.5 (Ramirez-Villegas et al.[25]). While an adequate total number of accessions relative to the overall range of each taxon (SRS score) and thus better accounting for differences in genetic structure between taxa[51] was considered preferable for analyses of taxa with sizeable germplasm collections, the 10 accession threshold may be useful as an objective baseline of adequacy across all taxa.

Mapping richness patterns. A wild relative taxon richness map was prepared by overlapping potential distribution models for all taxa (Fig. 1). A map with proposed hotspots for further collecting was produced by subtracting the areas of circular buffers of 50 km radius surrounding the original sites of collection of existing genebank accessions for each taxon from its potential distribution model, and then overlapping the resulting "collecting gap" models for all taxa assessed as high priority for further collecting.

Expert evaluation of results. Forty-four experts assessed the input data, potential distribution models, and gap analysis results based on their experience in botany, plant taxonomy, biodiversity, plant genetic resources conservation, and plant breeding (Supplementary Table 4). First, experts provided a numeric prioritization score for each wild relative taxon based solely on their knowledge of the sufficiency of existing accessions in genebanks worldwide. We called this score the comparable experts priority score (comparable EPS). Second, experts prioritized taxa based on their knowledge of wild relatives (including threats to taxa in their natural habitats as well as relative value of taxa in plant breeding). This score was called contextual EPS and was useful for providing additional information for collecting prioritization efforts. Both contextual EPS and comparable EPS were provided on a scale from 0 to 10 in alignment with the gap analysis scores. Following these steps, the gap analysis final priority score (FPS) was revealed to the experts, and they qualitatively evaluated their agreement with the results.

In addition, experts were asked to comment on occurrence data, potential distribution models, and maps of proposed collecting priorities. Following these contributions by experts, input occurrence data were further refined by eliminating clearly incorrect points and adjusting

country-level native areas, and the potential distribution modelling and gap analyses were re-run using the refined datasets in order to improve the quantitative and spatial results. The final runs of the analyses are presented in this article.

34. Food and Agriculture Organization of the United Nations (FAO). *FAOSTAT* http://faostat.fao.org (2013).

35. Khoury, C. K. *et al.* Estimation of countries' interdependence in plant genetic resources provisioning national food supplies and production systems. International Treaty on Plant Genetic Resources for Food and Agriculture, Research Study 8 (Rome: FAO) http://www.planttreaty.org/content/research-paper-8 (2015).

36. Harlan, J. R. & de Wet, J. M. J. Toward a rational classification of cultivated plants. *Taxon* **20,** 509–517 (1971).

37. Maxted, N., Ford-Lloyd, B. V., Jury, S., Kell, S. & Scholten, M. Towards a definition of a crop wild relative. *Biodivers. Conserv.* **15,** 2673–2685 (2006).

38. USDA ARS National Genetic Resources Program. Germplasm Resources Information Network - (GRIN). *National Germplasm Resources Laboratory Beltsville, Maryland* http://www.ars-grin.gov/~sbmljw/cgi-bin/taxcrop.pl?language=en (2014).

39. Boyle, B. *et al.* The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics* **14,** 16 (2013).

40. Cayuela, L., Granzow-de la Cerda, Í., Albuquerque, F. S. & Golicher, D. J. taxonstand: An R package for species names standardisation in vegetation databases. *Methods Ecol. Evol.* **3,** 1078–1083 (2012).

41. The Plant List. *The Plant List Version 1.1*. http://www.theplantlist.org/ (2013).

42. Warren, R. *et al.* Quantifying the benefit of early climate change mitigation in avoiding biodiversity loss. *Nat. Clim. Chang.* **3,** 1–5 (2013).

43. Elith, J. *et al.* Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29,** 129–151 (2006).

44. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecol. Modell.* **190,** 231–259 (2006).

45. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25,** 1965–1978 (2005).

46. VanDerWal, J., Shoo, L. P., Graham, C. & Williams, S. E. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecol. Modell.* **220,** 589–594 (2009).

47. Raxworthy, C. J. *et al.* Predicting distributions of known and unknown reptile species in Madagascar. *Nature* **426,** 837–841 (2003).

48. Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* **28,** 385–393 (2005).

49. Hijmans, R. J. & Spooner, D. M. Geographic distribution of wild potato species. *Am. J. Bot.* **88,** 2101–2112 (2001).

50. Olson, D. M. *et al.* Terrestrial ecoregions of the world: a new map of life on Earth. *Bioscience* **51,** 933–938 (2001).

51. Camadro, E. L. Is the genetic integrity of natural plant populations ex situ preserved with the current sampling, conservation and regeneration approaches? *J. Basic Appl. Genet.* **25,** 41–44 (2014).