An ex-situ conservation assessment diversity index (ECADI) for evaluating the coverage and diversity in germplasm collections

Chrystian C. Sosa[1], María Victoria Diaz[1], Norma Giraldo[1], Andrés Mendez[1], Lizeth Llanos[1], Julián Ramírez-Villegas[2]

[1]Alliance of Bioversity and CIAT, Cali, Colombia

[2]Alliance of Bioversity and CIAT, Rome, Italy.

1. Introduction

Germplasm banks constitute one of the main strategies to conserve key species able to be used for habitat restoration, create new crop varieties able to overcome challenges related to food security, and economic development by creating research repositories for global crops (Jago et al., 2024; Badri & Ludidi, 2022; Offord, 2017). This conservation is possible by obtaining material from plants out of their natural environments (Ex Situ conservation) with the goal to protect the genetic diversity (range of differences among individuals in DNA sequenced) of populations of cultivated species and their wild relatives in the form of accessions (Offord, 2017).

Plant genetic resources management of a species includes prospecting, collection, conservation, characterization, assessment, and use (Badri & Ludidi, 2022). Collecting aims to maximize species diversity under the premise of obtaining a minimum number of samples with minimum expenditure of resources (Maxted et al., 2020). Tools based on accession passport information for improving this management have been developed over the past few decades. Some examples are: (i) crop wild relatives and landraces gap analysis identify geographic areas for further collecting using ecogeographical approaches (Ramirez-Villegas et al., 2022, 2020; Khoury et al., 2019; Castañeda-Álvarez et al., 2016; Ramírez-Villegas et al., 2010), (ii) the composition imbalance index (CCII) optimize collecting of target subsets of species hierarchically structured known as end-groups (Van Treuren et al., 2009) or (iii) the Passport Data Completeness Indicator (PDCI) quantifies the level of completeness of passport data in a collection (Van Hintum et al., 2011).

Nevertheless, these tools are focused on supporting prospecting and data quality rather than diagnosing the coverage and diversity in germplasm collections properly.

To date there is a gap of methodologies that help diagnose diversity reported in plant germplasm collections which require consider complex components such as: (i) Collection Composition, (What taxa are in a collection and how abundant they are?) (ii) Passport Completeness (How well is the passport information used for research?), (iii) Ecogeography, (How many ecoregions or geographical regions covered wild relatives and landraces in the collection?), (iv) Genetic diversity (How effective is genetic diversity represented in plant materials collected?). Thus, as part of the initiative on genebanks led by CIAT, this report introduces a conceptual framework with a numerical index named ex-situ conservation assessment diversity index (ECADI) that includes the four mentioned components together and that can be reproducible across different germplasm collections, using as case studies the Beans, Cassava and Forages collections of CIAT.

2. Methodology

The ex-situ conservation assessment diversity index (ECADI) contemplates four main components (i) Collection composition and taxonomy, (ii) Documentation completeness, (iii) Ecogeographic representativeness, and (iv) Genetics diversity and genetic usability information availability.

2.1. Assumptions used for the ex-situ conservation assessment diversity index (ECADI) calculation

For the creation of the index related to the four components, the following assumptions were considered: (i) There are some approaches as the End-groups which consider the crop gene pool and cultivars in consideration which follows core collection concepts based on experts and farmers. (Van Treuren et al., 2009). These kinds of tools help the crop collection accessions prioritize or compare collections. (Hanson et al., 2024). Nevertheless, there is not always the availability of this hierarchical information for crops and germplasm collections, and the process of building a diversity tree takes time and collaborative efforts, making the use of this approach among genebanks extensively hard.

Consequently, with a need for simplicity in calculations, the following assumptions were considered: Each collection works as a community with three subsets represented by crop wild relatives, traditional cultivars, and interspecific hybrids which provide insights into wild diversity, and agrobiodiversity related to the subsets. (ii) a taxon found for each collection subset even if it is found in another collection subset is considered as an individual taxon (e.g., *Phaseolus vulgaris* found in landraces and crop wild relatives are considered as two different taxonomic units for the analysis). (iii) No cleaning procedure for georeferencing is done given that it is beyond the scope of this work. The geographical information quality used is evaluated in the documentation completeness component and the accessions with sufficient quality are used for the ecogeographic component when it is possible, (iv) As complementary information is required, only information used for the subsequent analyses here come from well-established taxonomy databases such as GRIN taxonomy and WorldFlora (Miller & Ulate, 2017), and Terrestrial Ecoregions of the World from Olson et al., (2001) (v) collection subsets (crop wild relatives, landraces, and interspecific hybrids) can be analyzed using geographical units such as countries. This is convenient for creating collecting and diagnostic planning and visualizing data for collecting data users including collection subsets and accessions with no country. (vi) This workflow wants to create a diagnostic framework for collection curators, thus, the taxonomy provided in each collection is the basis for all the analyses and the workflow will suggest taxonomy matches with current taxonomy databases.

2.2. Case studies used to evaluate ECADI

The three main collections of the International Center for Tropical Agriculture (CIAT) were used as case studies to assess the pertinence and accuracy of the ex-situ collection diversity index. These collections correspond to 66550 accessions in total, where 37936 (57%) are from the Beans collection, 5957 (8.95%) are from the Cassava collection, and 22657 accessions (34%) are from the Forages collection. The information on the CIAT's genebank was downloaded from the Genesys database (https://www.genesys-pgr.org/) and excludes historical accessions to evaluate active accessions only.

2.3. Preprocessing

The preprocessing step consisted of three stages i) Remove plant breeding material, ii) split collection data into subsets and iii) Taxonomic verification, geography, and extinction risk extraction.

2.3.1. Remove plant breeding material

All accessions with biological status of conservation (SAMPSTAT) with the following labels were not considered for subsequent analyses: Hybrid, Founder stock/base population, Inbred line (parent of hybrid cultivar), Segregating population, Clonal selection, Genetic stock, Mutant (e.g. induced/insertion mutants, tilling populations), Cytogenetic stocks (e.g. chromosome addition/substitution, aneuploids, amphiploids), Other genetic stocks (e.g. mapping populations), Advanced or improved cultivar (conventional breeding methods), and GMO (by genetic engineering) (Figure 1). This filtering step was performed to avoid artifacts such as excess of low values in the four different components of the ECADI given that the accessions above mentioned mostly do not possess geographical information, genetic data, or even do not have a clear taxonomy explained in a database.

2.3.2. Split collection data into subsets

Each collection was of (i) crop wild relatives (wild) expressed as S_W, (ii) traditional cultivars or landraces (landraces) expressed as S_L, and (iii) interspecific hybrids that are not part of plant breeding programs (interspecific) (e.g. *P. vulgaris* x *P. dumosus*) expressed as S_H, respectively using the biological hybrid status of accession as criteria (see Figure 1 for details). Records whose biological status was wild, natural, Semi-natural/wild, or Semi-natural/sown were assumed as crop wild relatives, and records cataloged as Traditional cultivar/landrace were considered landraces. The hybrids were discarded to reduce noise in subsequent analyses.

Each collection subset was represented into count tables with each country reported in the total collection as a column and each taxon as a row (Figure 1). Also, records without countries reported were included to facilitate subsequent analyses.

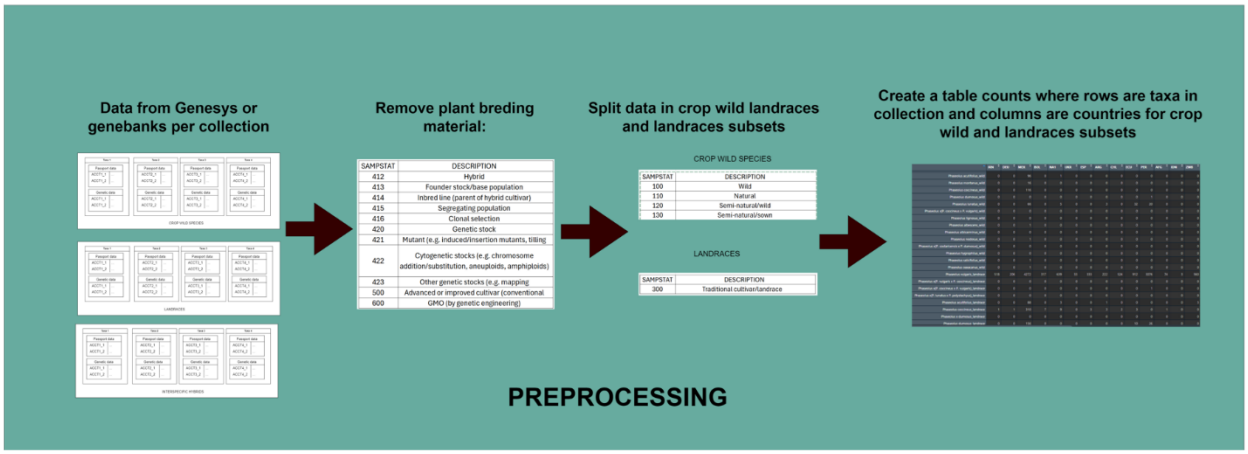


Figure 1. Preprocessing workflow proposed as the first step to calculate the ex-situ conservation assessment diversity index (ECADI). Each collection is divided in S_W, S_L, S_H representing crop wild relatives, landraces, and interspecific hybrids. Further, each collection subset is represented in table counts where each country is a column including the accessions without country in their passport data.

2.3.3. Taxonomic verification, geography, and extinction risk extraction

The taxonomy of each subset was verified to be used in further steps in the completeness component using GRIN taxonomy as main taxonomy and main source of native countries information, and World of Flora online (WoF) (Miller & Ulate, 2017) as an alternative taxonomy source. Extinction risk was extracted from the International Union for Conservation of Nature (IUCN) red list. All Preprocessing steps were performed in R; taxonomic matches with WoF were performed using WorldFlora v1.14.4 R package (Kindt, 2020). IUCN red list extinction risks, GRIN taxonomy and native areas extraction and count matrix were extracted the R code available at: https://github.com/alliance-datascience/genebanks_diversity_index.

2.4. Collection composition component

The composition and the taxonomic dimension of a collection were evaluated from two different perspectives: (i) Diversity expressed in effective number of species or Hill's numbers (focus on common species abundance), or Margalef index (focus on the number of species and the number of total individuals), ii) inequalities in records.

2.4.1. Diversity indexes

2.4.1.1.Hill-Simpson index

The effective number of species represents the number of equally abundant species that is needed to get a diversity value equal to D. This N is part of the unifying approach widely used in ecology and linearly represents better changes in diversity than measures based on entropy such as *Shannon or Gini Simpson represented* commonly as D (Ricotta & Feoli, 2024; Chao, Chiu, et al., 2014; Leinster & Cobbold, 2012).

D is the Simpson index, $D = (\sum_{i=1}^{R} {p_i}^2)$, where p_i, is the record proportion of the i[th] taxon in a subset (S_W, S_L, $or\ S_H$), or country of the subset (e.g. S_{W1}, S_{W2}, representing a country 1 or 2 in the crop wild relatives subset) This index is easily transformed into the Gini – Simpson index $H_{GS} = 1 - \sum_{i=1}^{R} {p_i}^2$, that represents the probability of two randomly chosen individuals belonging to different species where values near zero indicate high dominance and low diversity while values near one show high diversity (Chao, Chiu, et al., 2014) . To follow the premise of true diversity to obtain effective number of species proposed by Jost, (2006), D was converted to the effective number of species using the following formula: $H_2 = 1/(\sum_{i=1}^{R} {p_i}^2)$ (Jost, 2006) known also as Hill-Simpson, and can be interpreted as the effective number of dominant or very abundant species in the assemblage. And it has the following features : (i) It is referred as unit of species, (ii) It is sensitive to low relative abundance of common species due to it favored abundant species and rare species are discounted, (iii) Hill Simpson is under the replication principle: If N equally large and completely distinct assemblages (i.e., no species in common) have identical relative abundance distributions, then the Hill number of the pooled assemblage will be N times the diversity of an individual assemblage and ranges from $0 \leq H_2 \leq \infty$ (Roswell et al., 2021; Chao, Gotelli, et al., 2014; Chao, Chiu, et al., 2014).

For the ECADI, Hill Simpson (H_2) was calculated for each of the subsets and countries of the subsets, A normalization step using a sigmoid transformation was employed to: (i) obtain a normalized H_2 expressed as $H_{2(N)}$ that ranging from 0 to 1; $0 \leq H_{2(N)} \leq 1$ using the following equation $x_y = \frac{1}{1+\exp\left(\frac{-x_y-\bar{x}}{\sigma_x}\right)}$, and (ii) compare the diversity of each country and a subset. We suggest use H_2 for the following scenarios:

- Where a user wants to know the number of dominant species in a collection subset
- Where the evenness concept is important for the user, for instance when record numbers among species need to be similar
- Where a user wants to know if a collection subset or country part of a subset changed in time with a focus on evenness

2.4.1.2. Margalef index

The Margalef index, $d = \frac{S-1}{lnN}$, where S is the total number of species observed, N is the total number of counts in a sample, and ranges from $0 \leq d \leq \infty$. This index does not consider evenness or relative abundances as other indexes like Hill Simpson. Margalef index is based on that few species are heavily represented while other species have decreased numbers and consider the sample size on species richness reducing the bias from differences in sample sizes across communities and allow compare different communities (Peng et al., 2018; Yeom & Kim, 2011; Margalef, 1958).

For the ECADI index, the Margalef index is sensitive to the number of organisms collected and increases as the number of organisms sampled increases (Death, 2008). This property is convenient for ECADI for the following scenarios:

- Compare countries of a subset and a subset regardless of the different taxa number collected
- Given that the Margalef index is sensitive to species number and number of accessions collected, if a new taxon is collected, the index will change easily. Thus, the comparison in time for countries of a subset or subset is possible and it can show the differences in sampling effort in time and space more efficiently than the Hill Simpson index
- A user wants a fast indicator of composition assessment irrespective of the sample size

Similarly to Hill Simpson, a normalization step using the Sigmoid transformation method was employed to get a $d_{(N)}$ ranging from 0 to 1; $0 \leq d_{(N)} \leq 1$.

2.4.2. Inequalities in records

Due to diversity indexes affected by disparities in records, the Atkinson inequality index (A) was used instead of the Gini coefficient as the measurement of inequality of accessions records in the subsets and countries. Gini is a common index used in economy, but it is sensitive to changes in the middle part of the distribution, and not sensitive at bottom or top of incomes distribution (De Maio, 2007). In consequence, the Atkinson index (Atkinson, 1970) that is well-known used to measure the inequality of incomes in economics was used instead. This index is consistent of $A = 1 - \left[\sum_{i=1}^{n} \left(\frac{Y_i}{\bar{Y}}\right)^{1-\varepsilon} f_i\right]^{\frac{1}{1-\varepsilon}}, if\ \varepsilon \neq 1$, and $A = 1 - exp\left[\sum_{i=1}^{n} f_i log_e \frac{Y_i}{\bar{Y}}\right], if\ \varepsilon = 1$, where Y_i is the income of individuals in the ith h income range (N ranges altogether), f_i is the proportion of the population with income in the ith range, $\bar{Y}$ is the mean household income. In addition, the index introduces an explicit parameter (ε), this parameter increases the sensitivity to unequal incomes the higher the value, the more sensitive the Atkinson index becomes to inequalities at the bottom of the income distribution, and the commonest values are 0.5, 1, 1.5, or 2 (De Maio, 2007; Schlör et al., 2012). The interpretation of the Atkinson index is this: a value of 0.20 suggests that we could achieve the same level of social welfare with only 1- 0.20 = 80% of income. The theoretical range of Atkinson values is 0 to 1, with 0 being a state of equal distribution (De Maio, 2007).

In the case of the collection subsets or countries of a subset, the income is replaced by the number of accessions per taxon in a country or subset, respectively. Thus, the interpretation of A would be equivalent to what sampling effort is required to obtain a subset, or subsets in a country with an equal proportion of accessions per taxon. Additionally, 1-A may be considered as a measurement of equality (White, 2007) and in consequence, it was used in the composition index as a representation of the equality in taxon sampling per country and subset, where the higher the A value, the most equal the proportion of accession records is in the unit of analysis (e.g. countries of a subset, or the subset). The Atkinson index was calculated with an ε=0.5 to represent a moderate level of inequality aversion of taxa accessions inequal numbers in countries and the collection subsets. This index was calculated as a form to weigh up the inequalities of accessions collecting across taxa and obtain a composition index able to show diversity insights and consider that few taxa can have a high weight in a collection.

220

221 ### 2.4.3. Collection composition index calculation

222 The Composition index (CI) (Equation 1) is applied to each of the subsets: crop wild relatives,
223 landraces, and each of the countries of these subsets:

224 $$CI = D * (1 - A) \quad \text{(Equation 1)}$$

225 where D, and $(1 - A)$, represents diversity in the form of Hill Simpson or Margalef normalized
226 indexes ($H_{2(N)}$ or $d_{(N)}$), and equity of accession records (expressed as 1- Atkinson index)
227 respectively. This index ranges from $0 \leq CI \leq 1$, with values near zero corresponding to the low
228 diversity of taxa and a high inequality of records in a country for a given collection subset
229 ($S_{W1}, S_{W2,\ldots}$) or a collection subset itself ($S_W, S_L,$). The average CI for the wild $\left(CI_{S_W}\right)$, and
230 landrace (CI_{S_L}) subsets, was considered the main indicator of composition and taxonomy given
231 that interspecific hybrids do not have a consistent taxonomy due to being materials from distinct
232 species. In consequence, the composition index for the collection (Equation 2) is calculated as
233 follows:

234 $$CI_{collection} = \frac{CI_{S_W} + CI_{S_L}}{2} \quad \text{(Equation 2)}$$

235 ## 2.5. Documentation completeness component

236 Documentation completeness refers to the completeness of the information reported for each
237 accession in the form of a passport with taxonomy, collection site including country, administrative
238 levels, biological status of conservation, holding institute, and acquisition date among other
239 valuable information. To evaluate the documentation completeness, three points of view were
240 considered: (i) passport data completeness, in the form of the Passport data completeness index
241 (PDCI) (Van Hintum et al., 2011), (ii) Geographical quality score, and (iii) a taxonomical quality
242 score.

243 ### 2.5.1. Passport Data Completeness Index (PDCI)

244 The Passport Data Completeness Index (PDCI) was proposed by Van Hintum et al., (2011) and
245 indicates the degree of completeness of information based on the Multi-crop Passport Descriptors
246 (MCPD) (Alercia et al., 2015). PDCI is defined as the sum of values of MCDP fields with available

information divided by one hundred. Nevertheless, this index itself does not reflect the quality of an accession for taxonomic information or geographical information reported but it works as a starting point that is enriched with a Taxonomic completeness score (TCS), and the Geographical completeness score (GQS). With this regard, to calculate a completeness index, the PDCI reported for each of the accessions in each collection subset, was converted to PDCI normalized value using the following formula: $PDCI_{(N)} = \frac{PDCI}{10}$ which ranges $0 \leq PDCI_{(N)} \leq 1$.

For each of the subsets and countries of subsets was calculated the average of the average $PDCI_{(N)}$ values per taxon per country or collection subset respectively to obtain the value to be used for further steps.

2.5.2. Taxonomic completeness score (TCS)

The taxonomical quality score (TCS) was inspired by Van Hintum et al., (2011) and it is calculated for each accession in a collection, and it is presented as follows (Equation 4):

$$TCS = \sum Taxscores \quad \text{(Equation 4)}$$

Where $Taxscores$ are values added when the fields available in https://github.com/alliance-datascience/genebanks_diversity_index/tree/main are evaluated in the decision tree available at genebank-general/scripts/python/01_taxonomic_score_calc.ipynb at dev · alliance-datascience/genebank-general. $Taxscores$ works as follows: (i) Firstly, the GENUS field is evaluated to be if it is available in the accession. After this step, the SUBTAUTHOR field is evaluated in combination with SUBTAXA field to create two branches of options using SPAUTHOR and SPECIES information. The first branch is related to cases where SUBTAXA field information is available, and the second branch is where SUBTAXA field information is not available. Thus, the purpose of the two branches is to weigh up the absence of information in the author's fields: SUBTAUTHOR, and SPAUTHOR to provide an index that ranges from

$0 \leq TCS \leq 3$, where high values indicate accession passport data with good taxonomy information. In consequence, to provide curators with traffic lights to enhance accessions passport information a traffic light system with the following values was implemented:

- Red: (0-1 TCS value) indicates accessions with deficiencies such as no species name, genus name, species author or all together

- Yellow (1-2 TCS value) indicates accessions with no author information for species or sub taxa
- Green (2-3 TCS value) indicates records with accurate taxonomic information reported

In addition, a warning system was implemented suggesting taxonomic information according to GRIN taxonomy and WorldOfFlora databases to suggest taxonomic information to curators and users of the ECADI.

2.5.2.1.Taxonomic completeness index usage in ECADI

TCS was standardized as $TcS_{(N)} = {}^{TCS}/_{3}$, to take values from 0 to 1 and be used for the completeness component. The TCS was calculated in Python v3.11.

2.5.2.2. Taxonomic completeness score for a collection ($TCS_{collection}$)

The taxonomic quality score for the collection $TCS_{collection}$ is an average of the average taxonomy quality score of crop wild records $average(TCS_{S_W})$ and the average taxonomy completeness score of landraces $average(TCS_{S_L})$ (Equation 5).

$$TCS_{collection} = \frac{average(TCS_{S_W}) + average(TCS_{S_L})}{2} \quad \text{(Equation 5)}$$

2.5.3. Geographical Quality Score index (GQS)

A geographical quality index was developed in-house to evaluate geographical information provided by accessions passport data. This geographical quality score index was built under the premise that accurate geographical information is possible if the collecting site, origin country, administrative information, and altitude are mostly complete, and the coordinates do not lie near country borderlines, in the sea and if the administrative information matches with coordinates. This index consists of sixty-nine combinations of fields. The index takes a maximum value of 12 and it is calculated similarly to the taxonomic completeness score index as: $GQS = \sum_{i=1}^{n} val_i$, where val_i represents the number of validations in the decision tree used for the approach. This GQS ranges $0 \leq \text{GQS} \leq 12$, with the following bands as a traffic light for curators: GQS values: 11 - 12 (High geographical quality information), GQS values: 5 - 9 (Moderate geographical quality information), and GQS values: 0 - 4 (Low geographical quality information).

2.5.3.1. Geographical Quality Score index usage in ECADI

GQS was standardized as $GCS_{(N)} = {}^{GQS}/_{12}$, to take values from 0 to 1 and be used for the completeness component. A complete explanation of how GQS is calculated is presented is available at the following URL: GitHub - alliance-datascience/genebank-general at dev.

2.5.4. Documentation completeness index

The Documentation Completeness Index (DCI) is defined for each country for a collection subset or the collection subset itself as shows in Equation 6 as the geometric mean of PDCI, GQS, and TCS. The DCI values range from 0 to 1, $0 \leq DCI \leq 1$, where values near 0 indicate poor geographical and taxonomical data reported performance, and the passport data fields lack information. For each subset the: $PDCI_{collection}, GQS_{collection}, and\ TCS_{collection}$ are calculated as weighted means of these indexes using as weights the proportion of accessions of species in the subset collection. On the other hand, the completeness indexes for a country, for instance, the indexes for a country 1 in the crop wild relatives subset (S_{W1}) are expressed as $PDCI_{S_{W1}}, GQS_{S_{W1}}, and\ TCS_{S_{W1}}$, and they are the weighted averages of the averages values of indexes of the species available in S_{W1}, where weights are the proportion of taxa in the country S_{W1}. The values of PDCI, GQS, and TCS per country in combination with the values in a subset are used to visualize how similar the countries are regarding the documentation completeness index.

$$DCI = \sqrt[3]{PDCI * GQS * TCS} \qquad \text{(Equation 6)}$$

Finally, for a collection, the Documentation completeness index is defined in Equation 7. In general, the DCI represents how well represented is the quality of information in passport, geography, and taxonomic information is and it will take high values when: An accurate passport, geographical coordinates and the complete fields of taxonomy in the passport data in a collection are available.

$$DCI_{Collection} = \frac{DCI_{S_W} + DCI_{S_L}}{2} \qquad \text{(Equation 7)}$$

2.6. Ecogeography component

The ecogeographical component refers to the aspect of diversity that is driven by variation across different ecological and geographical zones. This considers how environmental and geographical factors influence the distribution and variation of species across broader landscapes, linking diversity not just to species counts, but to the underlying environmental and geographical contexts that support it. To evaluate this component, only accessions with Geographical Quality Score (GQS) high or moderate were used, due to the 'higher' accuracy in the location information such as coordinates, collecting sites and origin countries, reducing the probability of finding mismatches in the information.

2.6.1. Preprocessing

Given that the ecogeographic component corresponds to the ecosystems represented in a collection that can shape the genetic diversity observed in either phenotypic or genotypic traits in a plant taxon, the following premises were considered:

- The main information used must be extracted from the accessions to capture the information of the genebank collection.
- The information would be enriched using complementary information from literature or sources such as world terrestrial ecosystems.
- The main unit evaluated is the ecosystems to evaluate the adaptation and diversity of environments represented in a collection.
- The calculation should be done for each country and each of the subsets of a collection to provide an approach able to diagnose how well taxa have been collected in the geographical space.
- Only accurate geographical coordinates are used to work as a proxy of the whole subset collection and ensure the reproducibility of the approach across germplasm collections and use for future comparisons.
- Similarly as showed by Castañeda-Álvarez et al., (2016), A taxon with less than 10 accessions indicates a poor performance in sampling representation in germplasm system. The approach presented here only considers taxa with more than 10 accessions with the

goal of using accurate ecogeographical information reducing the bias in the ecogeographical representativeness index.

Ecosystems provide a natural framework that consider a range of interacting environmental factors, that is why this component is focused on finding how the different taxon in a germplasm collection are distributed across different ecoregions around the world, and how this distribution varies depending on the biological status of the accessions and the countries that hold them (either if they are considered as part of their native area or not). For the extraction of the ecoregions, Terrestrial Ecoregions of the World (Olson et al., 2001) was used as input using the sf R package to extract the ecosystems information per country.

The ecogeographic component is calculated as follows:

2.6.2. At collection level:

As the main unit evaluated are the ecosystems, the main indicator for countries of a subset or a collection is represented in the ecosystems coverage for a taxon which is determined by the coverage of ecosystems in native countries ($COVER_{NC}$), which is calculated as follows (Equation 8):

$$COVER_{NC} = \frac{\#\ unique\ ecoregions\ inside\ native\ countries\ where\ the\ taxon\ is}{\#\ unique\ ecoregions\ inside\ native\ countries} \quad \text{(Equation 8)}$$

This equation represents the proportion of unique ecosystems found in the native countries reported for a given taxon and its values range $0 \leq COVER_{NC} \leq 1$ where high values represent that a collection subset or a taxon that has a broad range of ecosystems collected. As this indicator only considers the ecoregions uniqueness, a weight is consider and it is the proportion of taxa with at least 10 accessions available in the subset n ($P_{n>10} = \frac{\sum_{i=1}^{N} x_i}{N}$), where x is a taxon that has more than 10 accessions in the subset, and N the total number of taxa in the subset

The combination of the coverage of ecosystems in native countries ($COVER_{NC}$), and the proportion of taxa with at least 10 accessions available is considered to obtain an indicator of ecogeographical representativeness index ($ECO_{Collection}$) for a germplasm collection as follows (Equation 9):

$$ECO_{COL} = \frac{[mean(COVER_{NC}) * P_{n>10}]_{landraces} + [mean(COVER_{NC}) * P_{n>10}]_{wild}}{2} \quad \text{(Equation 9)}$$

ECO_{COL} values range $0 \leq ECO_{COL} \leq 1$ where high values indicate a greater ecogeographical diversity of environments represented in the collection.

2.6.3. At country level (Complementary information):

Similarly to the approach used at landraces and crop wild relatives' collection subset. The ecogeographical representativeness index approach can be used to obtain a coverage of ecosystems diagnosis per country ($COVER_{Country}$) (Equation 10).

$$COVER_{Country} = \frac{\#\ unique\ ecoregions\ inside\ the\ country\ where\ the\ diff\ taxa\ are}{\#\ unique\ ecoregions\ inside\ the\ country} \quad \text{(Equation 10)}$$

For each subset (either landrace or wild biological status), the proportion of taxa with at least 10 accessions available $P_{n>10} = \frac{\sum_{i=1}^{N} x_i}{N}$ is used but for this case, *x* is the taxon that has more than 10 accessions in the country and N the total number of taxa in the country. Thus, the ecogeographical representativeness per country ($ECO_{Country}$) regardless of the subset is defined as follows (Equation 11):

$$ECO_{Country} = \frac{[mean(COVER\ C) * P_{n>10}]s_W + [mean(COVER\ C) * P_{n>10}]s_L}{2} \quad \text{(Equation 11)}$$

2.7. Usability and genetic diversity component

Genetic diversity is a vital component conserved in germplasm collections. Including genetics data into a diversity index is a complex task that contemplates collecting plant material, DNA extraction, and sequencing using genotyping by sequencing techniques such as DaRTSeq making it difficult to be included in an evaluation of diversity or under the convention of biological diversity (CBD) framework (Carvajal-Yepes et al., 2024; Hoban et al., 2024, 2021). Nevertheless, for all wild species, population genetics data is lacking. This lack of data jeopardizes its use for examinate the diversity. Thus, approaches such as macrogenetics or include proxies such as IUCN Red List status arises to be used as proxy to evaluate genetic diversity (Schmidt et al., 2023). Thus, two approaches were considered for this report: (i) Median of genetic distances evaluated (ii) Use a blend of different proxies for genetic diversity such as the IUCN Red List status, combined with the availability of sequenced accessions for each of the CIAT germplasm collections to cover this component when genetic information is not available.

2.7.1. Genetic diversity for CIAT germplasm collections

Genetic diversity is conceived in this report such as: (i) geometric average of genetic distance (e.g., Nei, Rogers) among individuals of the same species or a core. Thus, the Genetic distance for a species is reported as $GD_{species} = Median(x)$, where x represents the genetic distances for a pairwise of individuals of a given species or taxon. This Genetic distance can be obtained for a collection subset S_W, or S_L or a country in a subset $GD_{S_{W1}}, GD_{S_{L1}}$ and it is represented as $GD_{SW} = Median(GD_{species})$, and $GD_{S_{W1}} = Median(GD_{species\ in\ a\ country\ for\ the\ subset})$, and (ii) The expected heterozygosity, and it is calculated using the Gini – Simpson index $H_{GS} = 1 - \sum_{i=1}^{R} {p_i}^2$, which represents the probability that a pair of randomly sampled allele copies p_i from a population are different (Harris & DeGiorgio, 2017). Nevertheless, this genetic diversity needs to be studied deeply in further steps beyond the scope of this report.

2.7.2. Genetic information Usability

2.7.2.1. Proportion of accessions and taxa sequenced

Given that genetic information available to calculate a genetic diversity approach is limited in a realistic scenario, the proportion of records per taxon that have been sequenced, and proportion of taxa sequenced were considered as indicator of usability in terms of sequenced accessions coverage in a collection subset or country part of collection subset: the first approach is represented by ($pRseq$). It is calculated using a weighted average (Equation 12), which represents the proportion of accessions sequenced in each taxon weighted by the proportion of taxon records in a collection subset, where each subset corresponds to crop wild relatives, and landraces.

$$pRseq = \frac{\sum_{i=1}^{n} \frac{nRecords\ taxon(sequenced)}{nRecords\ taxon} * \frac{nRecords\ taxon}{nRecords\ taxon_{collection}}}{\sum_{i=1}^{n} \frac{nRecords\ taxon}{nRecords\ taxon_{collection}}} \quad \text{(Equation 12)}$$

The second approach corresponds to the proportion of taxa in a country of a subset or a collection subset that possesses at least one record sequenced ($pTseq$), and this is calculated in the following equation (Equation 13):

$$pTseq = \frac{ntaxa(sequenced)}{ntaxa\ total\ _{collection}} \quad \text{(Equation 13)}$$

2.7.2.2. Proportion of taxa reported as threatened in IUCN red list

As an alternative to replace the genetic diversity component, a proportion of taxa in the categories: Extinct: (EX), Extinct in the Wild (EW), Critically Endangered (CR), Endangered (EN), and Vulnerable (VU) in the IUCN red list of threatened species. This IUCN information alone does not inform about the genetic diversity but it is a possible indicator of genetic diversity erosion and it is used as recommended by the Convention on Biological Diversity (CBD) as status of genetic diversity easily obtained for a country (Schmidt et al., 2023; Hoban et al., 2021). Thus, the proportion of vulnerable taxa $pTaxaIUCN$ reflects this approach (Equation 14).

$$pTaxaIUCN = \frac{ntaxa\ in\ (VU,EN,CR,EW)}{ntaxa\ total_{\ collection}} \quad \text{(Equation 14)}$$

2.7.3 Usability index

The usability index used as an alternative to genetic diversity expressed either the median of genetic distances or the expected heterozygosity. This usability index is defined as the average of the Proportion of accessions or taxa sequenced ($pRseq$), Proportion of accessions or taxa sequenced ($pTseq$), and Proportion of taxa reported as threatened in IUCN red list ($pTaxaIUCN$) (Equation 15). This usability index USI ranges $0 \leq USI \leq 1$ and it is applied to each country in a subset, and collection subsets. The USI takes high values when there are a high number of accessions sequenced, a high proportion of taxa sequenced and there are taxa threatened in IUCN. Thus, a high value can indicate to curators how well a subset, country of subset or collection about DNA sequenced available and if a collection works as a reservoir of genetic information for threatened taxa.

$$USI = \frac{pRseq + pTseq + pTaxaIUCN}{3} \quad \text{(Equation 15)}$$

2.8. ex-situ conservation assessment diversity index (ECADI)

Finally, the ECADI is calculated as the average of the four main components analyzed in this report (Equation 16) and it is the main output of the workflow introduced in this report. This index is an indicator of composition, documentation completeness, ecogeographic representativeness and

usability applied to subsets, countries of a collection subset. ECADI values take values from 0 to 1, 0≤ECADI ≤1. ECADI works as diagnostics of the diversity expressed in the composition (CI), and coverage of a collection: (i) Documentation completeness, (ii) Ecogeographical representativeness, and (iii) Usability. Given this complex nature, ECADI must be interpreted with the four main components together. For ECADI, a high value indicates a more complete collection in terms of composition, documentation completeness, ecogeographical representativeness and DNA availability or threatened taxa.

$$ECADI = \frac{CI+DCI+ECO+USI}{4} \quad \text{(Equation 16)}$$

For a collection, the ECADI is the average of the value obtained for crop wild relatives, and landraces collection subsets and it is calculated as follows (Equation 17).

$$ECADI_{Collection} = \frac{ECADI_{S_W}+ECADI_{S_L}}{2}. \quad \text{(Equation 17)}$$

All calculations were performed using R ≥4.2. and the R scripts created are documented and available in the following GitHub repository: https://github.com/alliance-datascience/genebanks_diversity_index/tree/dev. For the composition, documentation completeness, usability, and ECADI, Non-linear Iterative Partial Least Squares (NIPALS) were performed as visualization of results obtained using the plsdepot, and ggplot2 R packages. Finally, spider plots using ggradar R package were obtained to provide a complete visualization of ECADI and its components together.

3. Results

3.1. Composition index for CIAT collections

3.1.1. A higher composition diversity for crop wild relatives subsets in the Beans collection

Beans collection had a total of 37077 accessions analyzed. A total of 94% of accessions corresponded to five landraces taxa, 6% of accessions corresponded to 47 taxa. In accordance with the landrace's dominance, averages of 18.74, and 306.4 accessions were sampled for crop wild relatives and landraces subsets, respectively. The crop wild relative's subset collection showed a

497 greater inequality and diversity than the landraces subset collection respectively with Margalef,
498 Hill-Simpson indexes.

499 A total of three dominant taxa were found for the crop wild relatives' subset whilst one species
500 was found dominant for the total landraces subset which was *P. vulgaris* landraces. In general, the
501 indexes proposed showed greater values for landraces subset due to this subset having better
502 equality (higher 1-A value) than the crop wild relatives subset (Supplementary Table 1), suggesting
503 that Beans landraces had been collected more equally. This premise is supported by the
504 composition index values found lower for crop wild relatives than landraces (Supplementary Table
505 1).

506

507 Regarding countries in the crop wild relatives' collection subset, Mexico, Guatemala, and Costa
508 Rica were the countries with highest values for diversity and equity (Supplementary Figure 1). On
509 the other hand, for the landraces subset, Mexico, Guatemala, and Puerto Rico had the highest
510 values of equity, and Morrocco, Dominican Republic, Puerto Rico had the highest values of Hill
511 Simpson index with almost two dominant landraces taxa (Supplementary Figure 2A & 2B). A
512 disperse trend was observed in NIPALS graphic with highest values of equality for Mexico, Puerto
513 Rico, Colombia, Guatemala, El Salvador. But also, Morrocco, Indonesia, Jamaica, Zimbabwe,
514 Argentina, or Philippines possessed high values of the composition index calculated with Hill-
515 Simpson index and Margalef index which suggest that those countries possessed a similar number
516 of Beans species landraces and similar accession counts.

517

518 **3.1.2. A higher composition diversity for crop wild relatives subsets in the Cassava**
519 **collection**

520 The CIAT Cassava germplasm collection evaluated for ECADI consisted of twenty-four species
521 of *Manihot* genus, and one unique landrace taxon (*Manihot esculenta)* that overall contains 5329
522 accessions together. The crop wild relatives collection subset possessed almost four dominant
523 species, which indicated less diversity than observed in the Beans crop wild relatives with almost
524 three species (Supplementary Table 2), and this less diversity but with a higher equality which is
525 observed in collection composition values of 0.125, and 0.187 using the Margalef and Hill-

Simpson diversity indexes respectively (Supplementary Table 2). Given that only one taxon was observed in the landraces subset, the Hill-Simpson diversity had a value of one, which means that there is not diversity and the unavailability of calculating Atkinson index. Finally, only three countries had information for Cassava crop wild relatives (Brazil, Colombia, and Mexico), Brazil had the highest value for the composition index using either Margalef or Hill-Simpson index (Supplementary Figure 3A & B).

3.1.3. Forages have the most diverse collection for crop wild relatives in CIAT germplasm system

CIAT Forages collection is the most diverse collection from the three case studies for the crop wild relatives subset with a total of 749 taxa, and no taxa were labeled as landraces representing 22534 accessions. Forages collection had the highest diversity in terms of species richness with sixty dominant species, and on average almost nine dominant Forages species per country, and a forty-three dominant for the accessions without reported country.

Equality (1-Atkinson) values were like Beans crop wild relatives collection subset (Supplementary Table 1-3), but on average the equality values per country were almost zero indicating that overall few countries had more collected accessions (Supplementary Table 3). The composition index values for the collection were 0.195, and 0.193 respectively due to the lack of landraces information.

Colombia, Mexico, Venezuela, and Brazil where the countries with the highest values of composition index using Margalef and Hill-Simpson indexes respectively that means that those countries had the highest diversity and good equity in accessions sampling (Supplementary Figure 4A & 4B).

Overall, composition index values indicated that Beans collection had greater values of composition due to its balance of diversity and equality. In comparison Cassava, and Forages collections had a greater diversity for Cassava and Forages crop wild relatives subset, nevertheless these collections have low and inexistent diversity for landraces subset (Table 1).

3.2. Documentation Completeness indexes for CIAT collections

In general passport for the crop wild relatives subsets in the CIAT germplasm collections, the passport data completeness index (PDCI), Geographical quality score index (GQS), and Taxonomic completeness score index (TCS) values were higher for Beans than Cassava and Forages collections. This behavior was corroborated by the Documentation completeness index (DCI) that indicated that Beans collection had greater values followed by Forages and Cassava collections respectively for crop wild relatives subset (Supplementary Table 4). For the case of landraces, a similar trend was observed where Beans landraces subset had higher completeness indexes values than Cassava collection DCI values (Supplementary Table 4 & Table 1). In addition, as a validation of the documentation completeness results, the accessions with no reported country of a collection had lower values for PDCI, and GQS whilst Forages collection had the poorest performance in taxonomy quality score given the high number of accessions without a proper epithet reported (71 taxa in total with the epithet *sp*., and one taxon labeled as *Leguminosa indet.faboideae*).

3.3. Ecographical representativeness index for CIAT germplasm collections

The ecogeographical representativeness index only considered accessions with medium and high Geographical Quality Score categories ($5 \leq \mathrm{GQS} \leq 12$) as a way to avoid use data without proper geographical coordinates to extract ecoregions from the and Terrestrial Ecoregions of the World (Olson et al., 2001). A total of 36 countries for Beans landraces subset had complete ecogeographic repetitiveness ($ECO_{Country}$ =1) indicating a good sampling effort. In contrast, the crop wild relatives subset had a complete ecogeographical representativeness for Belize, and Bermudas while Honduras, El Salvador, Bolivia had moderate to high representativeness ($ECO_{Country} \geq 0.4$).

For Cassava collection, a total of 18 countries located in the Latin America region had high and complete ecogeographic representativeness ($0.896 \leq ECO_{Country} \leq 1$) for the landrace subset represented *M. esculenta.* In comparison, the crop wild relatives subset only covered two countries (Brazil, and Colombia) whose taxa had more than 10 accessions and their representativeness were

of 0.0805, and 1 respectively). These $ECO_{Country}$ values indicate that the Brazil ecogeographical representation of crop wild relatives in the CIAT germplasm system would be poor in comparison with Colombia.

Regarding Forages crop wild relatives subset (Forages landraces are not reported), a total of fourteen countries located in Oceania, Africa, North America, and Latinoamerica had a complete ecogeographic representativeness ($ECO_{Country} = 1$). China, Nigeria, Rwanda, Uganda, Laos, Oman, Paraguay, and Trinidad and Tobago had moderate to high values ($ECO_{Country} \geq 0.5$).

The index suggests that on average, the Beans and Cassava landraces collection subsets possessed the highest ecogeographical representation in their respective collections showing that the landraces of the five *Phaseolus* landraces species, and *Manihot esculenta* had a greater sampling effort. On the other hand, Beans, and Forages crop wild relatives had higher ecogeographical representativeness than Forages. Thus, obtaining the average per collection, Cassava followed by Beans collection had the highest values of ecogeographical representation given the landraces high values found (Table 1).

3.4. Genetic diversity or usability index for CIAT collections

The usability component was evaluated for the three CIAT collections as an alternative of the genetic diversity approach. This usability index was considered since only Cassava landraces had reported genetic data information to calculate the genetic diversity indicator. This genetic distance had a value of 0.184 which reflects redundance in Cassava Landraces as observed in Carvajal-Yepes et al., (2024).

The usability index calculated for the three CIAT germplasm collections showed higher value for the landraces of Beans and Cassava than the crop wild relatives subsets, respectively. Furthermore, the highest value of usability was obtained for Cassava collection. This high value is due to the high proportion of taxa and accessions per taxa sequenced which was also observed in the accessions without country reported (Supplementary Table 6). Also, it is important to highlight that none of the three CIAT collections registered more than the 17.5% of taxa threatened in IUCN

red list, which means that CIAT collections possessed a low level of threatened taxa and the sampling has been focused on landraces genetic information extraction.

For Beans crop wild relatives collection subset, the countries with highest values for the usability index were Mexico, Bermudas, and Guatemala, whilst Colombia, Mexico and Guatemala had the highest values for landrace subset collection (Supplementary figure 5A & 5B). Similarly, for Cassava collection, Brazil, Mexico, and Colombia had highest values for crop wild relatives subset and Colombia, Brazil, and Peru had the highest values for usability for landrace subset. (Supplementary figure 6A & 6B). Finally, for Forages, Burkina Faso, Oman, and Zimbabwe had the highest values for the usability index which have the highest proportion of accessions sequenced for Forages crop wild relatives (Supplementary figure 7). Thus, the usability index was able to identify how has been the DNA sequencing effort focused.

3.5. Summary of four components in one unique diversity index (ECADI)

The ECADI was calculated using the four main components considered previously, and it was applied for each of the countries of crop wild relatives and landraces where they were available to be used. The results showed that the Beans crop wild relatives tended to have low values of ECADI rather than the landraces subsets (Supplementary Table 7 & Table 1). The collections with the highest values for crop wild relatives and landraces ECADI values were Cassava followed by Beans and finally Forages (Figure 2A, 2B, and 2C).

The reason for this ECADI high values for Cassava were due to the high ecogeographical representativeness of Cassava landrace combined with a good quality in the passport data and a good performance of sequenced accessions reported even if the composition value reflected a low value which indicates a low taxa diversity (Figure 2B). On the other hand, in the case of Beans collection, landrace collection exhibited higher ecogeographical representativeness as well as usability, and composition. This composition index for Beans was higher for landraces due to the higher equality shown in landraces (Table 1, Supplementary Table 1, Figure 2A). Finally, the

Forages collection had the lowest ECADI value. This is due to landraces were not available for this collection. This was the most diverse collection of the three analyzed (65 dominant taxa, see Table 1). But also, usability, and ecogeographical representativeness values were lowest than Beans and Cassava collection (Figure 2C).

For the crop wild relatives subset for Beans collection, Mexico, Bolivia, Guatemala, Costa Rica, Honduras, Belize, Bermudas had values ($ECADI_{Country} > 0.3$). For the landraces subset of Beans collection a list of 66 countries located mainly in Latin America, Europe, and Sub-Saharan Africa possessed values greater than 0.3 ($ECADI_{Country} > 0.3$) (Supplementary figures 8A & 8B). The ECADI for collection (Average of crop wild relatives and landraces) had values greater than 0.3 for countries in Latin America region.

For the Cassava crop wild relatives subset collection, Only Colombia had values greater than 0.3 ($ECADI_{Country} > 0.3$) . On the other hand, a total of sixteen countries worldwide, mostly located in Latin America (14 countries, 87.5%) had values greater than 0.3 ($ECADI_{Country} > 0.3$) for the Cassava landrace subset. Consequently, Brazil and Colombia were the countries with ECADI values for the collection (Average of crop wild relatives and landraces) (Supplementary figure 9 A & 9B).

For the Forages crop wild relatives subset, a list of 13 countries (30.7% of these countries were located in Latin America, and 30.7% of countries were located in Sub-Saharan Africa) gotten values greater than 0.3 ($ECADI_{Country} > 0.3$). Given that no landrace was available for this collection, ECADI for the collection are low ($ECADI_{Country} < 0.2$) (Supplementary Figure 10).

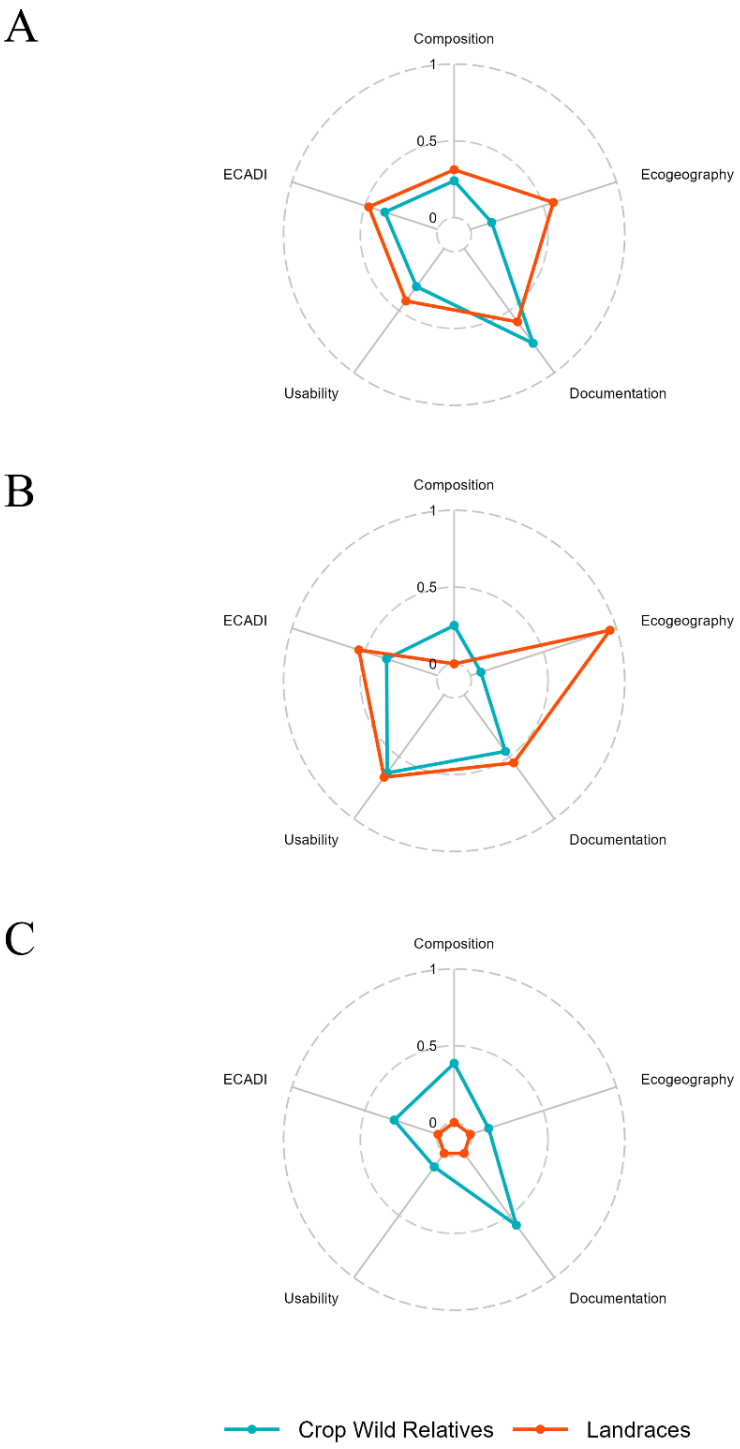


Figure 2. Radar chart with the main metrics calculated for the four main components of the ex-situ conservation assessment diversity index (ECADI) (i) Collection composition and taxonomy (Composition), (ii) Documentation completeness (Documentation), (iii) Ecogeographical representativeness (Ecogeography), and (iv) Genetics diversity and genetic usability information availability (Usability). The blue color represents the values obtained for Crop Wild Relatives and red represents the values obtained for landraces for: A) Beans CIAT collection, B) Cassava CIAT collection, and C) Forages CIAT collection.

Table 1. Summary table for the ECADI and each of the four components evaluated per collection in the CIAT collections. [a] Numbers in parentheses represent the number of accessions analyzed). [b] Numbers in parentheses represent the effective number of species, and equality index represented as 1-A). [c] Numbers in parentheses represent the median of the genetic distance. In addition, interspecific hybrids are not displayed given that they are not contributing to the index. N/A displayed not available information. Values highlighted are the highest values for the subsets, and collection per component.

Component	Collection	Crop Wild Relatives	Landraces	Overall
Species count [a]	Beans	47 (2137)	**5 (34940)**	52
	Cassava	24 (382)	1 (4947)	25
	Forages	**749 (22534)**	0 (0)	**749**
(i) Composition [b]	Beans	0.24 (3.175/0.339)	**0.312 (1.297/0.487)**	**0.276**
	Cassava	0.25 (3.77/0.498)	0 (0/1)	0.15
	Forages	**0.386 (60.69/0.39)**	0 (0/N/A)	0.193
(ii) Documentation completeness	Beans	**0.764**	**0.591**	**0.6775**
	Cassava	0.457	0	0.2285
	Forages	0.579	N/A	0.579
(iii) Ecogeography	Beans	0.0874	0.568	0.257
	Cassava	0.72	**0.955**	0.447
	Forages	0.125	0	0.0625
(iv) Genetic diversity or usability index [c]	Beans	0.307	0.423	0.365
	Cassava	0.63	**0.665 (0.184)**	0.648
	Forages	0.116	N/A	0.055
ECADI	Beans	**0.364**	0.474	0.419
	Cassava	0.352	0.**542**	0.447
	Forages	0.3	0	0.15

4. Conclusions and next steps

This research is ongoing and aims to develop a more comprehensive, efficient method for assessing germplasm collection diversity. The ECADI pipeline provides a flexible, accessible approach to understanding genetic resources, with clear potential for further refinement and broader application in agricultural research. ECADI possessed the following strengths:

(i) Uses readily information available MCDP format obtained from Genesys

(ii) ECADI used biodiversity metrics widely used such as the Effective Number of Species obtained from Simpson dominance index (Hill- Simpson index).

(iii) ECADI can provide a quick diagnostic of germplasm collection by dividing it into two subsets (crop wild relatives and landraces) that do not require the computation of species distribution models or complex algorithms to obtain results and evaluate the collection.

(iv) ECADI provides a first glance at the genetic diversity bottleneck assessment in germplasm collection by the introduction of an indicator based on the IUCN red list and the inclusion of the number of accessions and taxa sequenced. This is not properly a proxy to examine the genetic diversity of a collection, but its use provides a vision of the potential use in future of a collection to examine its genetic diversity.

(v) ECADI prioritizes information collected in native countries. This is observed in the results obtained for ecographical representativeness, composition, and ECADI.

Also, the methodology here proposed needs the following next steps to be more robust:

- Enhance the standardization applied in the composition method to use better effective number of species or Margalef indexes results together with the inequality metrics.
- Add information from the previous analysis made for Crop Wild Relatives and Landraces conservation gap made by CIAT as alternative to the ecogeographical representativeness score introduced here.
- Use macrogenetics approaches to interpolate genetic data when little genetic data is available using machine learnings approaches (Schmidt et al., 2023; Sosa et al., 2023).
- Use phylogenetic diversity (PD) data as an alternative to the genetic component to obtain insights of the functional diversity given that the PD measures the evolutionary relationship

of species using the tree of life, and its use with crop wild relatives ecogeographic have been used (González-Orozco et al., 2021).

- Implement the code in Python to integrate the pipeline with the Genesys database.
- Evaluate with other CGIAR centers and get feedback from curators.

5. Acknowledgements

The authors are grateful to Monica Carvajal-Yepes for her help in providing the data used in the genetic and usability component of the ECADI framework as well as guidance to enhance this research work.

6. References

Alercia, A., Akita, S., & Mackay, M. (2015, December). *FAO/Bioversity Multi-Crop Passport Descriptors V.2.1 [MCPD V.2.1]*. Bioversity International. https://cgspace.cgiar.org/items/e2aab885-8e08-4a0a-86eb-4494974c5199

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, *2*(3), 244–263. https://doi.org/10.1016/0022-0531(70)90039-6

Badri, M., & Ludidi, N. (2022). Germplasm Conservation for Biotechnology and Plant Breeding. In Kamaluddin, U. Kiran, & M. Z. Abdin (Eds.), *Technologies in Plant Biotechnology and Breeding of Field Crops* (pp. 67–80). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-5767-2_4

Carvajal-Yepes, M., Ospina, J. A., Aranzales, E., Velez-Tobon, M., Correa Abondano, M., Manrique-Carpintero, N. C., & Wenzl, P. (2024). Identifying genetically redundant accessions in the world's largest cassava collection. *Frontiers in Plant Science*, *14*, 1338377. https://doi.org/10.3389/fpls.2023.1338377

Castañeda-Álvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., Guarino, L., Harker, R. H., Jarvis, A., Maxted, N., Müller, J. V., Ramirez-Villegas, J., Sosa, C. C., Struik, P. C., Vincent, H., & Toll, J. (2016). Global conservation priorities for crop wild relatives. *Nature Plants*, *2*(4), 16022. https://doi.org/10.1038/nplants.2016.22

Chao, A., Chiu, C.-H., & Jost, L. (2014). Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 297–324. https://doi.org/10.1146/annurev-ecolsys-120213-091540

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45–67. https://doi.org/10.1890/13-0133.1

De Maio, F. G. (2007). Income inequality measures. *Journal of Epidemiology & Community Health*, *61*(10), 849–852. https://doi.org/10.1136/jech.2006.052969

Death, R. (2008). Margalef's Index. In S. E. Jørgensen & B. D. Fath (Eds.), *Encyclopedia of Ecology* (pp. 2209–2210). Academic Press. https://doi.org/10.1016/B978-008045405-4.00117-8

González-Orozco, C. E., Sosa, C. C., Thornhill, A. H., & Laffan, S. W. (2021). Phylogenetic diversity and conservation of crop wild relatives in Colombia. *Evolutionary Applications*, *14*(11), 2603–2617. https://doi.org/10.1111/eva.13295

Hanson, J., Lusty, C., Furman, B., Ellis, D., Payne, T., & Halewood, M. (2024). Opportunities for strategic decision making in managing *ex situ* germplasm collections. *Plant Genetic Resources: Characterization and Utilization*, *22*(4), 195–200. https://doi.org/10.1017/S1479262123000357

Harris, A. M., & DeGiorgio, M. (2017). An Unbiased Estimator of Gene Diversity with Improved Variance for Samples Containing Related and Inbred Individuals of any Ploidy. *G3 Genes|Genomes|Genetics*, *7*(2), 671–691. https://doi.org/10.1534/g3.116.037168

Hoban, S., Campbell, C. D., Da Silva, J. M., Ekblom, R., Funk, W. C., Garner, B. A., Godoy, J. A., Kershaw, F., MacDonald, A. J., Mergeay, J., Minter, M., O'Brien, D., Vinas, I. P., Pearson, S. K., Pérez-Espona, S., Potter, K. M., Russo, I.-R. M., Segelbacher, G., Vernesi, C., & Hunter, M. E. (2021). Genetic diversity is considered important but interpreted narrowly in country reports to the Convention on Biological Diversity: Current actions and indicators are insufficient. *Biological Conservation*, *261*, 109233. https://doi.org/10.1016/j.biocon.2021.109233

Hoban, S., Da Silva, J. M., Hughes, A., Hunter, M. E., Kalamujić Stroil, B., Laikre, L., Mastretta-Yanes, A., Millette, K., Paz-Vinas, I., Bustos, L. R., Shaw, R. E., Vernesi, C., the Coalition for Conservation Genetics, Funk, C., Grueber, C., Kershaw, F., MacDonald, A., Meek, M., Mittan, C., … Segelbacher, G. (2024). Too simple, too complex, or just right? Advantages, challenges, and guidance for indicators of genetic diversity. *BioScience*, *74*(4), 269–280. https://doi.org/10.1093/biosci/biae006

Jago, S., Elliott, K. F. V. A., Tovar, C., Soto Gomez, M., Starnes, T., Abebe, W., Alexander, C., Antonelli, A., Baldaszti, L., Cerullo, G., Cockel, C., Collison, D., Cowell, C., Delgado, R., Demissew, S., Devenish, A., Dhanjal-Adams, K., Diazgranados, M., Drucker, A. G., … Borrell, J. S. (2024). Adapting wild biodiversity conservation approaches to conserve agrobiodiversity. *Nature Sustainability*. https://doi.org/10.1038/s41893-024-01427-2

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Khoury, C. K., Amariles, D., Soto, J. S., Diaz, M. V., Sotelo, S., Sosa, C. C., Ramírez-Villegas, J., Achicanoy, H. A., Velásquez-Tibatá, J., Guarino, L., León, B., Navarro-Racines, C., Castañeda-Álvarez, N. P., Dempewolf, H., Wiersema, J. H., & 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

Kindt, R. (2020). WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. *Applications in Plant Sciences*, *8*(9), e11388. https://doi.org/10.1002/aps3.11388

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: The importance of species similarity. *Ecology*, *93*(3), 477–489. https://doi.org/10.1890/10-2402.1

Margalef, R. (1958). Temporal succession and spatial heterogeneity in Phytoplankton. In A. A. Buzzati-Traverso (Ed.), *Perspectives in Marine Biology* (DGO-Digital original, 1, pp. 323–350). University of California Press; JSTOR. https://doi.org/10.2307/jj.8441698.27

Maxted, N., Hunter, D., & Ortiz Ríos, R. (2020). Germplasm Collecting. In *Plant Genetic Conservation* (pp. 320–352). Cambridge University Press.

Miller, C., & Ulate, W. (2017). World Flora Online Project: An online flora of all known plants. *Proceedings of TDWG*, *1*, e20529. https://doi.org/10.3897/tdwgproceedings.1.20529

Offord, C. A. (2017). Germplasm Conservation. In *Encyclopedia of Applied Plant Sciences* (pp. 281–288). Elsevier. https://doi.org/10.1016/B978-0-12-394807-6.00046-0

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, *51*(11), 933. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

Peng, Y., Fan, M., Song, J., Cui, T., & Li, R. (2018). Assessment of plant species diversity based on hyperspectral indices at a fine scale. *Scientific Reports*, *8*(1), 4776. https://doi.org/10.1038/s41598-018-23136-5

Ramírez-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*(10), e13497. https://doi.org/10.1371/journal.pone.0013497

Ramirez-Villegas, J., Khoury, C. K., Achicanoy, H. A., Diaz, M. V., Mendez, A. C., Sosa, C. C., Kehel, Z., Guarino, L., Abberton, M., Aunario, J., Awar, B. A., Alarcon, J. C., Amri, A., Anglin, N. L., Azevedo, V., Aziz, K., Capilit, G. L., Chavez, O., Chebotarov, D., … Zavala, C. (2022). State of ex situ conservation of landrace groups of 25 major crops. *Nature Plants*, *8*(5), 491–499. https://doi.org/10.1038/s41477-022-01144-8

Ramirez-Villegas, J., Khoury, C. K., Achicanoy, H. A., Mendez, A. C., Diaz, M. V., Sosa, C. C., Debouck, D. G., Kehel, Z., & Guarino, L. (2020). A gap analysis modelling framework to prioritize collecting for ex situ conservation of crop landraces. *Diversity and Distributions*, *26*(6), 730–742. https://doi.org/10.1111/ddi.13046

Ricotta, C., & Feoli, E. (2024). Hill numbers everywhere. Does it make ecological sense? *Ecological Indicators*, *161*, 111971. https://doi.org/10.1016/j.ecolind.2024.111971

Roswell, M., Dushoff, J., & Winfree, R. (2021). A conceptual guide to measuring species diversity. *Oikos*, *130*(3), 321–338. https://doi.org/10.1111/oik.07202

Schlör, H., Fischer, W., & Hake, J.-F. (2012). Measuring social welfare, energy and inequality in Germany. *Applied Energy*, *97*, 135–142. https://doi.org/10.1016/j.apenergy.2012.01.036

Schmidt, C., Hoban, S., & Jetz, W. (2023). Conservation macrogenetics: Harnessing genetic data to meet conservation commitments. *Trends in Genetics*, *39*(11), 816–829. https://doi.org/10.1016/j.tig.2023.08.002

Sosa, C. C., Arenas, C., & García-Merchán, V. H. (2023). Human Population Density Influences Genetic Diversity of Two Rattus Species Worldwide: A Macrogenetic Approach. *Genes*, *14*(7), 1442. https://doi.org/10.3390/genes14071442

Van Hintum, T., Menting, F., & Van Strien, E. (2011). Quality indicators for passport data in *ex situ* genebanks. *Plant Genetic Resources*, *9*(3), 478–485. https://doi.org/10.1017/S1479262111000682

Van Treuren, R., Engels, J. M. M., Hoekstra, R., & Van Hintum, Th. J. L. (2009). Optimization of the composition of crop collections for ex situ conservation. *Plant Genetic Resources*, *7*(02), 185–193. https://doi.org/10.1017/S1479262108197477

White, T. J. (2007). Sharing resources: The global distribution of the Ecological Footprint. *Ecological Economics*, *64*(2), 402–410. https://doi.org/10.1016/j.ecolecon.2007.07.024

Yeom, D.-J., & Kim, J. H. (2011). Comparative evaluation of species diversity indices in the natural deciduous forest of Mt. Jeombong. *Forest Science and Technology*, *7*(2), 68–74. https://doi.org/10.1080/21580103.2011.573940

7. Supplementary data

7.1. Supplementary tables

Supplementary Table 1. Summary table for the composition component indexes calculated for CIAT Beans crop wild relatives, and landraces collection. Values referred as average are reported as mean ± sd. In addition, interspecific hybrids are not displayed given that they are not contributing to the composition index.

Indicator	*Crop Wild Relatives subset*			*Landraces subset*			*Range*
	Average for countries	**N/A country**	**Subset**	**Average for countries**	**N/A country**	**Subset**	
Taxa number	0.83 ± 3.494	4	47	1.85±1.04	4	5	0-∞
Accession counts	18.74±99.25	38	2137	306.49±727.048	810	34940	0-∞
1-Atkinson	0.069±0.069	0.057	0.339	0.272±0.094	0.449	0.487	0-1
Margalef index	1.083±1.237	0.825	6	0.208±0.236	0.448	0.382	0-∞
Hill-Simpson (Effective number of species)	2.008±1.311	1.553	3.175	1.144±0.254	1.513	1.297	0-∞
Composition index (Margalef)	0.05±0.065	0.024	0.317	0.152±0.109	0.328	0.328	0-1
Composition index (Hill-Simpson)	0.045±0.059	0.023	0.240	0.144±0.11	0.362	0.312	0-1

Supplementary Table 2. Summary table for the composition component indexes calculated for CIAT Cassava crop wild relatives, and landraces collection. Values referred to as average are reported as mean ± sd. In addition, interspecific hybrids are not displayed given that they are not contributing to the composition index. N/A displayed not available information.

Indicator	*Crop Wild Relatives subset*			*Landraces subset*			*Range*
	Average for countries	**N/A country**	**Subset**	**Average for countries**	**N/A country**	**Subset**	
Taxa number	0.607±2.657	15	24	0.964±0.189	1	1	0-∞
Accession counts	11.25±55.3	24	382	176.62±402.432	1	4947	0-∞
1-Atkinson	0.12±0.107	0.5275	0.498	1	N/A	1	0-1
Margalef index	0.949±1.193	3.329	3.86	0	0	0	0-∞
Hill-Simpson (Effective number of species)	1.730±0.69	9.737	3.77	1	1	0	0-∞
Composition index (Margalef)	0.055 ± 0.067	0.362	0.374	N/A	N/A	N/A	0-1
Composition index (Hill-Simpson)	0.047 ± 0.046	0.447	0.250	N/A	N/A	N/A	0-1

Supplementary Table 3. Summary table for the composition component indexes calculated for CIAT Forages crop wild relatives, and landraces collection. Values referred to as average are reported as mean ± sd. In addition, interspecific hybrids are not displayed given that they are not contributing to the composition index. N/A displayed not available information.

Indicator	***Crop Wild Relatives subset***			***Landraces subset***			***Range***
	Average for countries	**N/A country**	**Subset**	**Average for countries**	**N/A country**	**Subset**	
Taxa number	28.324 ± 46.286	297	749	N/A	N/A	N/A	0-∞
Accession counts	244.838 ± 766.393	4416	22534	N/A	N/A	N/A	0-∞
1-Atkinson	0.025 ± 0.033	0.2	0.39	N/A	N/A	N/A	0-1
Margalef index	6.087 ± 5.978	35.268	74.63	N/A	N/A	N/A	0-∞
Hill-Simpson (Effective number of species)	8.862 ± 8.984	43.387	60.689	N/A	N/A	N/A	0-∞
Composition index (Margalef)	0.018 ± 0.027	0.185	0.389	N/A	N/A	N/A	0-1
Composition index (Hill-Simpson)	0.017 ± 0.027	0.189	0.386	N/A	N/A	N/A	0-1

Supplementary Table 4. Summary table for the documentation completeness component indexes calculated for CIAT collections. Values referred to as average are reported as mean ±sd. In addition, interspecific hybrids are not displayed given that they are not contributing to the completeness index. N/A displayed not available information.

Collection	Indicator	*Crop Wild Relatives subset*			*Landraces subset*		
		Average for countries	**N/A country**	**Subset**	**Average for countries**	**N/A country**	**Subset**
Beans	Passport Data completeness index (PDCI)	0.677±0.023	0.541	0.663	0.633±0.029	0.533	0.648
	Geographical quality score index (GQS)	0.729±0.311	0	0.82	0.335±0.256	0.002	0.481
	Taxonomic quality score index (TQS)	0.666±0.002	0.667	0.82	0.663±0.019	0.666	0.662
	Documentation completeness index (DCI)	0.665±0.147	0	0.764	0.484±0.145	0.087	0.591
Cassava	Passport Data completeness index (PDCI)	0.593±0.025	0.504	0.597	0.625±0.032	0.545	0.651
	Geographical quality score index (GQS)	0.231±0.246	0	0.4	0.305±0.219	0	0.465
	Taxonomic quality score index (TQS)	0.52±0.035	0.59	0.4	0.582±0.081	0.667	0.547
	Documentation completeness index (DCI)	0.384±0.155	0	0.457	0.452±0.13	0	0.549
Forages	Passport Data completeness index (PDCI)	0.585±0.037	0.502	0.598	N/A	0	0
	Geographical quality score index (GQS)	0.373±0.286	0.001	0.57	N/A	0	0
	Taxonomic quality score index (TCS)	0.487±0.029	0.496	0.57	N/A	0	0
	Documentation completeness index (DCI)	0.439±0.135	0.057	0.579	N/A	N/A	N/A

Supplementary Table 5. Summary table for the ecogeographic component indexes calculated for CIAT collections. Values referred to as average are reported as mean±sd. In addition, interspecific hybrids are not displayed given that they are not contributing to the index. N/A displayed not available information. ECO represents the ecogeographical representativeness index.

Collection	Indicator	*Crop Wild Relatives subset*		*Landraces subset*		*Collection value*	Range
		Average for countries	**Subset**	**Average for countries**	**Subset**		
Bean	ECO	0.409±0.274	0.146	0.806±0.216	0.568	0.257	0-1
	$P_{n>10}$	N/A	0.0874	N/A	0.8846		0-1
Cassava	ECO	0.06±0.235	0.072	0.994±0.024	0.955556	0.4473	0-1
	$P_{n>10}$	N/A	0.0688	N/A	0.9311		0-1
Forages	ECO	0.541±0.327	0.125	N/A	0	0.0625	0-1
	$P_{n>10}$	N/A	1	N/A	0		0-1

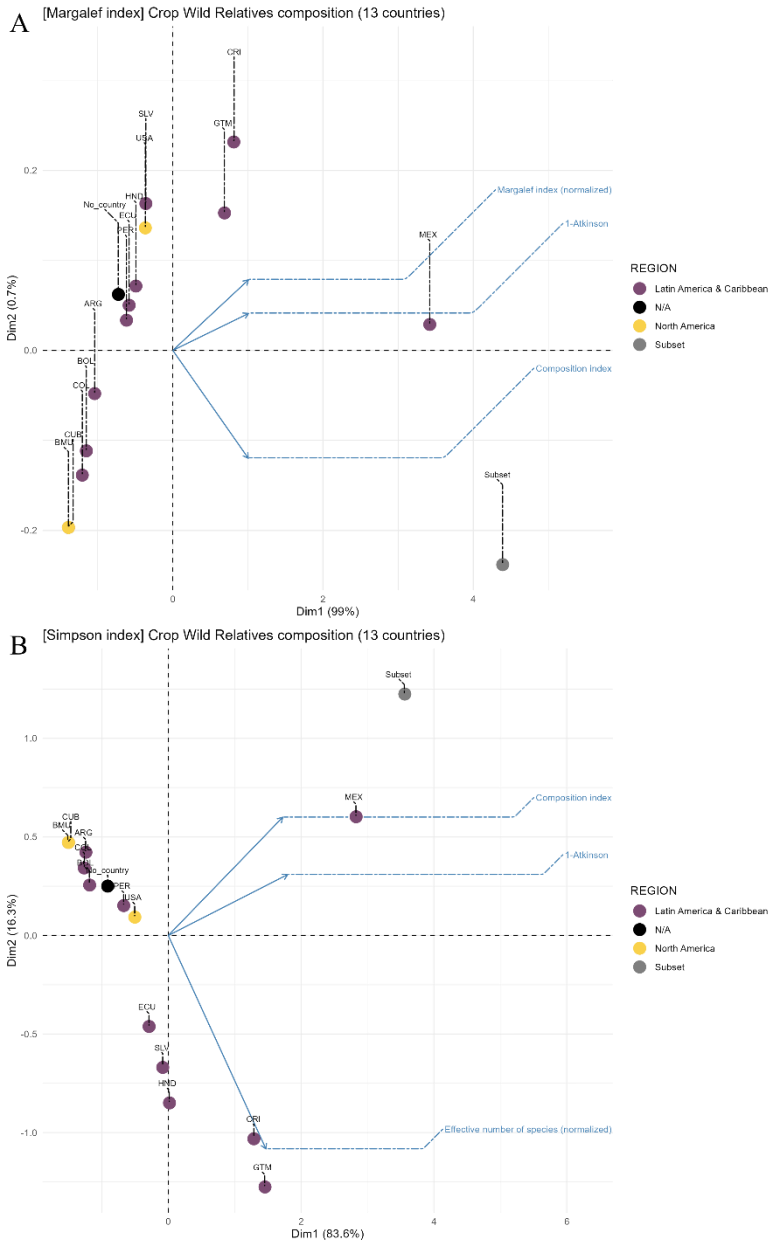
Supplementary Table 6. Summary table for the genetic diversity and usability component indexes calculated for CIAT collections. Values referred to as average are reported as mean ± sd. In addition, interspecific hybrids are not displayed given that they are not contributing to the index. N/A displayed not available information.

Collection	Indicator	*Crop Wild Relatives subset*			*Landraces subset*			*Collection*
		N/A country	**Average for countries**	**Subset**	**N/A country**	**Average for countries**	**Subset**	
Bean	Proportion of taxa sequenced	0	0.171 ± 0.283	0.237	0	0.004 ± 0.008	0.254	0.75
	Proportion of accessions per taxon sequenced	N/A	0.006 ± 0.024	0.511	N/A	0.014 ± 0.015	1	0.245
	Proportion of taxa reported as threatened in IUCN red list	0	0.002 ± 0.013	0.175	0	0.001 ± 0.005	0.175	0.174
	Usability index	0	0.009 ± 0.044	0.307	0	0.006 ± 0.008	0.423	0.3653
Cassava	Proportion of taxa sequenced	0.56	0.**024** ± 0.106	0.958	0.04	0.039 ± 0.008	1	0.979
	Proportion of accessions per taxon sequenced	0.660	0.761 ± 0.217	0.932	0	0.037 ± 0.082	0.996	0.964
	Proportion of taxa reported as threatened in IUCN red list	0	0 ± 0	0	0	0 ± 0	0	0
	Usability index	0.407	0.035 ± 0.108	0.630	0.013	0.025 ± 0.027	0.665	0.648
Forages	Proportion of taxa sequenced	0.013	0.001 ± 0.002	0.035	N/A	N/A	N/A	0.017
	Proportion of accessions per taxon sequenced	0.333	0.24 ± 0.326	0.290	N/A	N/A	N/A	0.145
	Proportion of taxa reported as threatened in IUCN red list	0	$5.3 \times 10^{-5} \pm 2.6 \times 10^{-4}$	0.004	N/A	N/A	N/A	0.002
	Usability index	0.116	0.021 ± 0.064	0.109	N/A	N/A	N/A	0.055

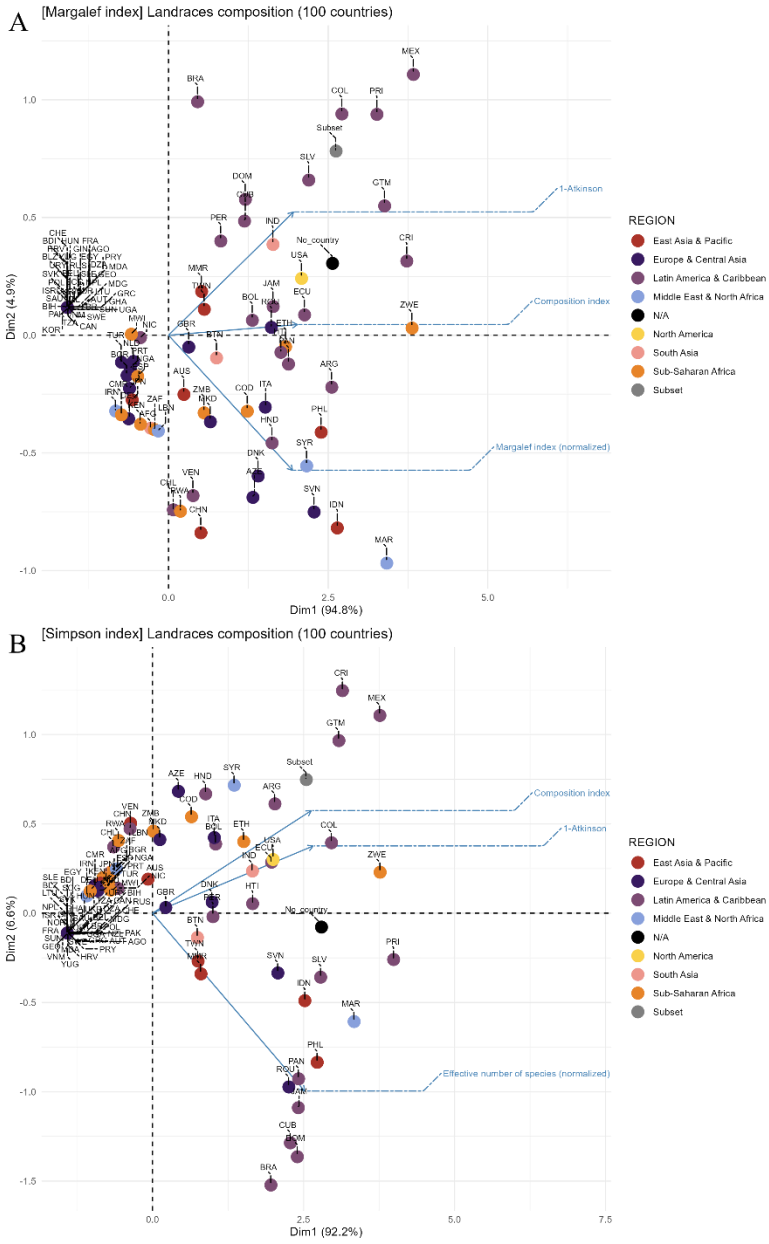
Supplementary Table 7. ECADI values per country, subset and collection obtained for the three CIAT collections analyzed in this report.

Collection	**Indicator**	***Crop Wild Relatives subset***			***Landraces subset***			***Collection***
		N/A country	**Average for countries**	**Subset**	**N/A country**	**Average for countries**	**Subset**	
Bean	ECADI	0.006	0.044 ± 0.11	0.364	0.112	0.282 ± 0.145	0.474	0.419
Cassava		0.213	0.03 ± 0.093	0.352	0.003	0.257 ± 0.159	0.542	0.447
Forages		0.091	0.195 ± 0.107	0.3	0	0 ± 0	0	0.15

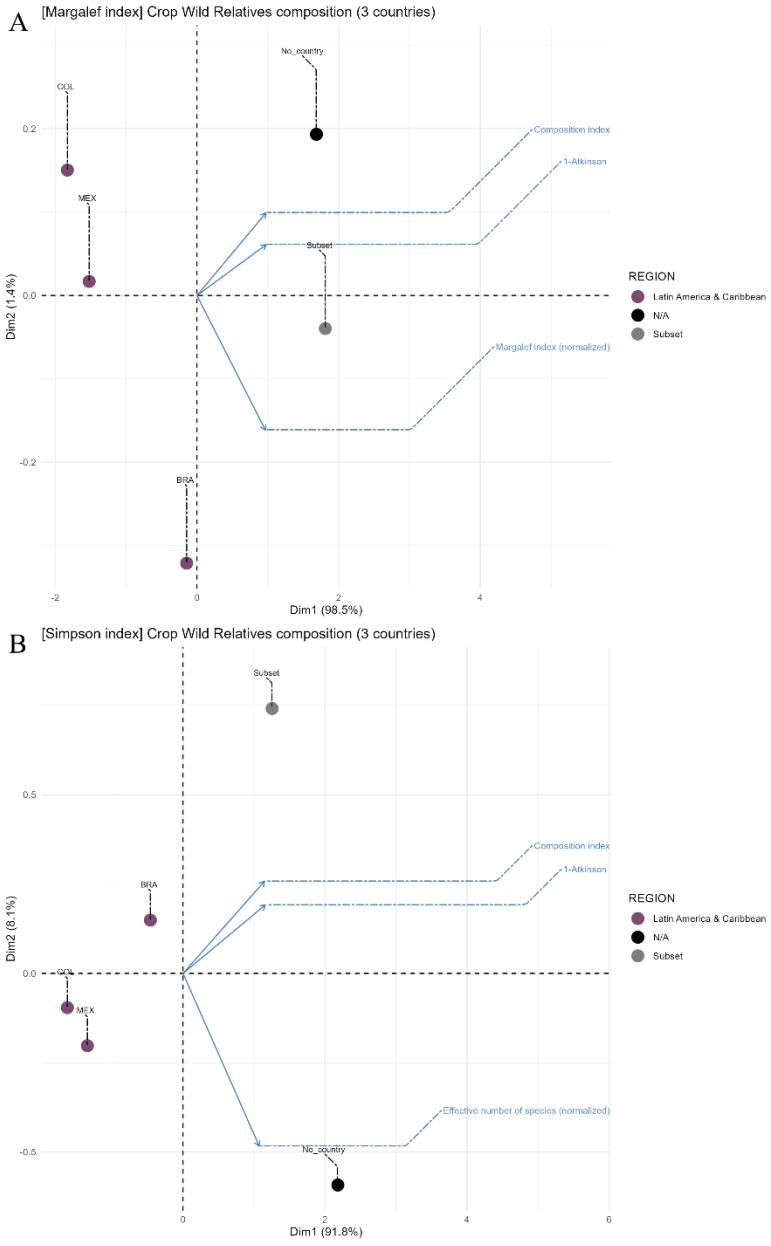
7.2. Supplementary figures



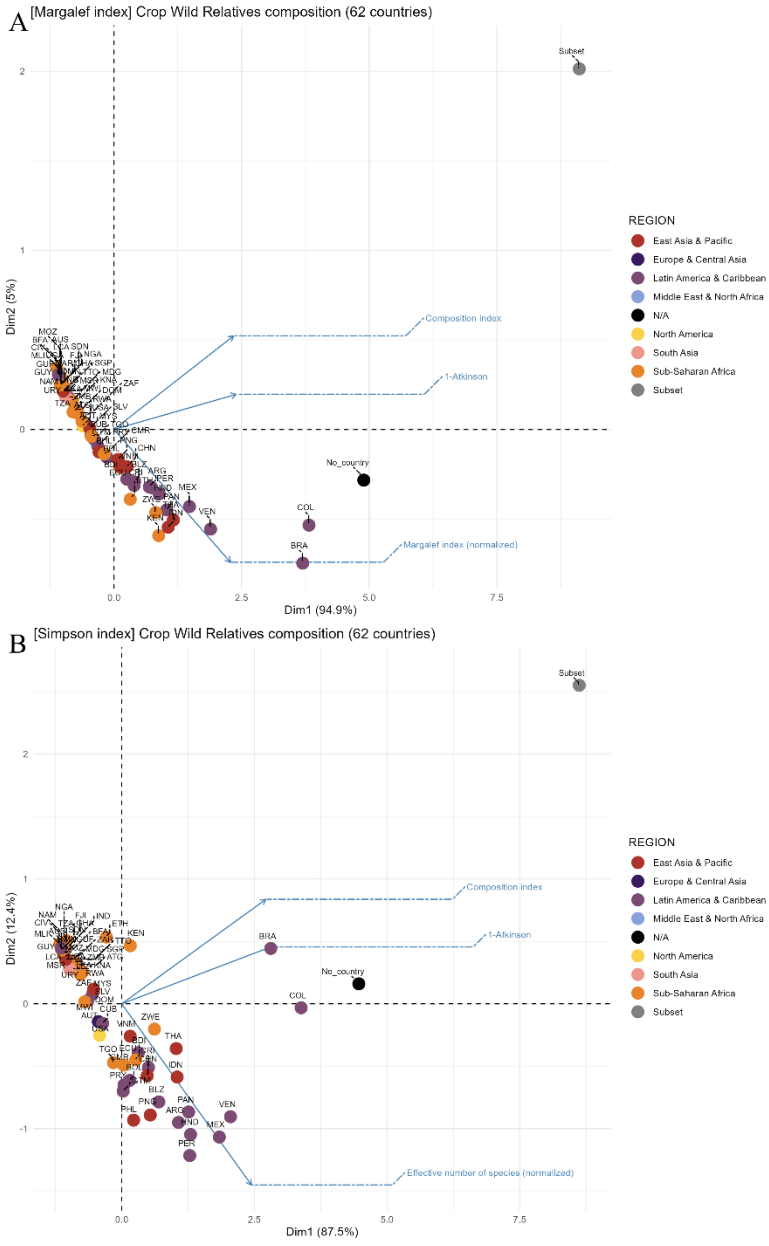
Supplementary Figure 1. NIPALS graphical representation of species diversity, equality, and the composition index for the crop wild relatives taxa in CIAT Beans collection. Each color represents a world region according to World bank criteria. A) NIPALS representation using Margalef index B) NIPALS representation using Hill-Simpson index (Effective number of species).



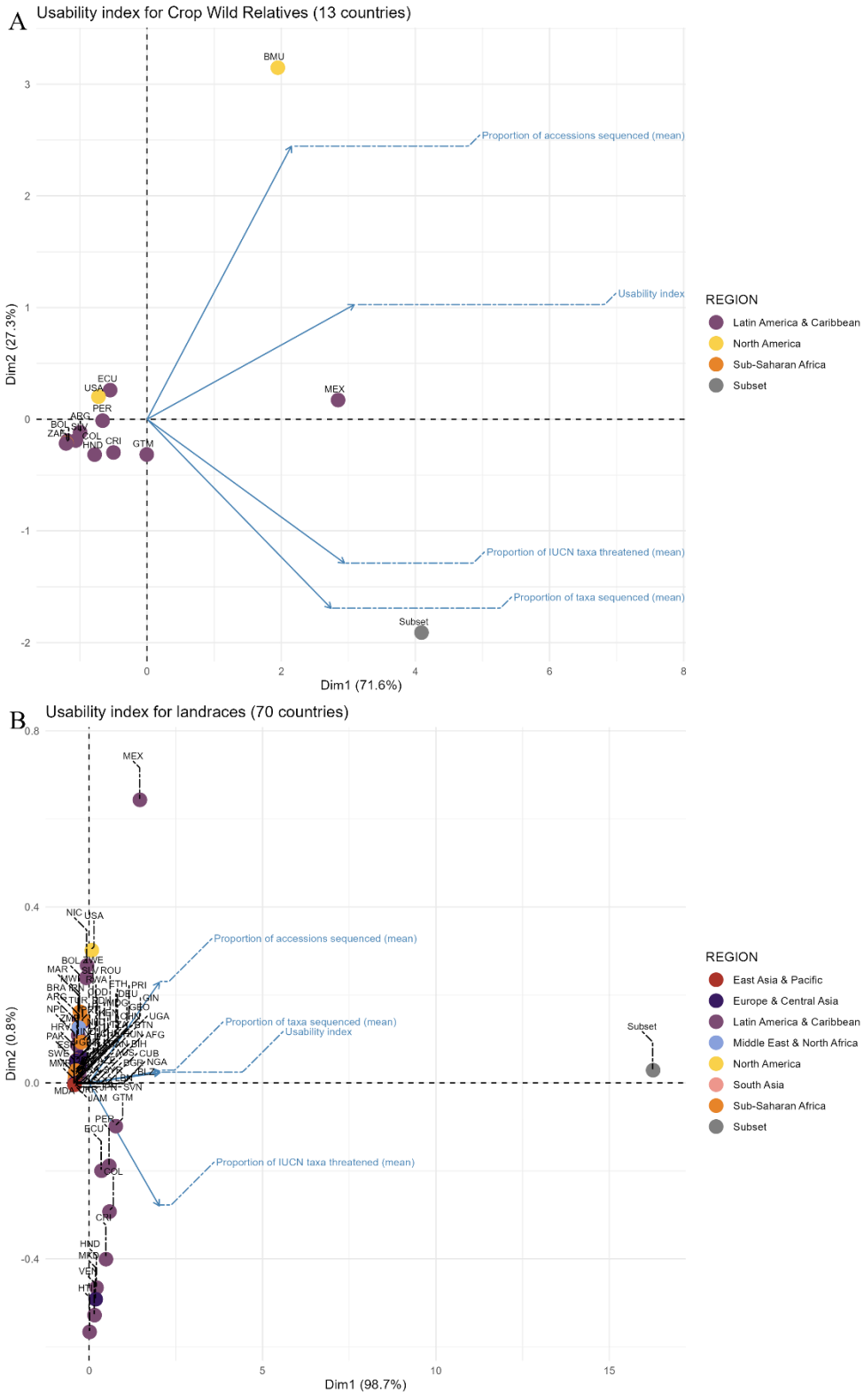
Supplementary Figure 2. NIPALS graphical representation of species diversity, equality, and the composition index for the landraces taxa in CIAT Beans collection. Each color represents a world region according to World bank criteria. A) NIPALS representation using Margalef index B) NIPALS representation using Hill-Simpson index (Effective number of species).



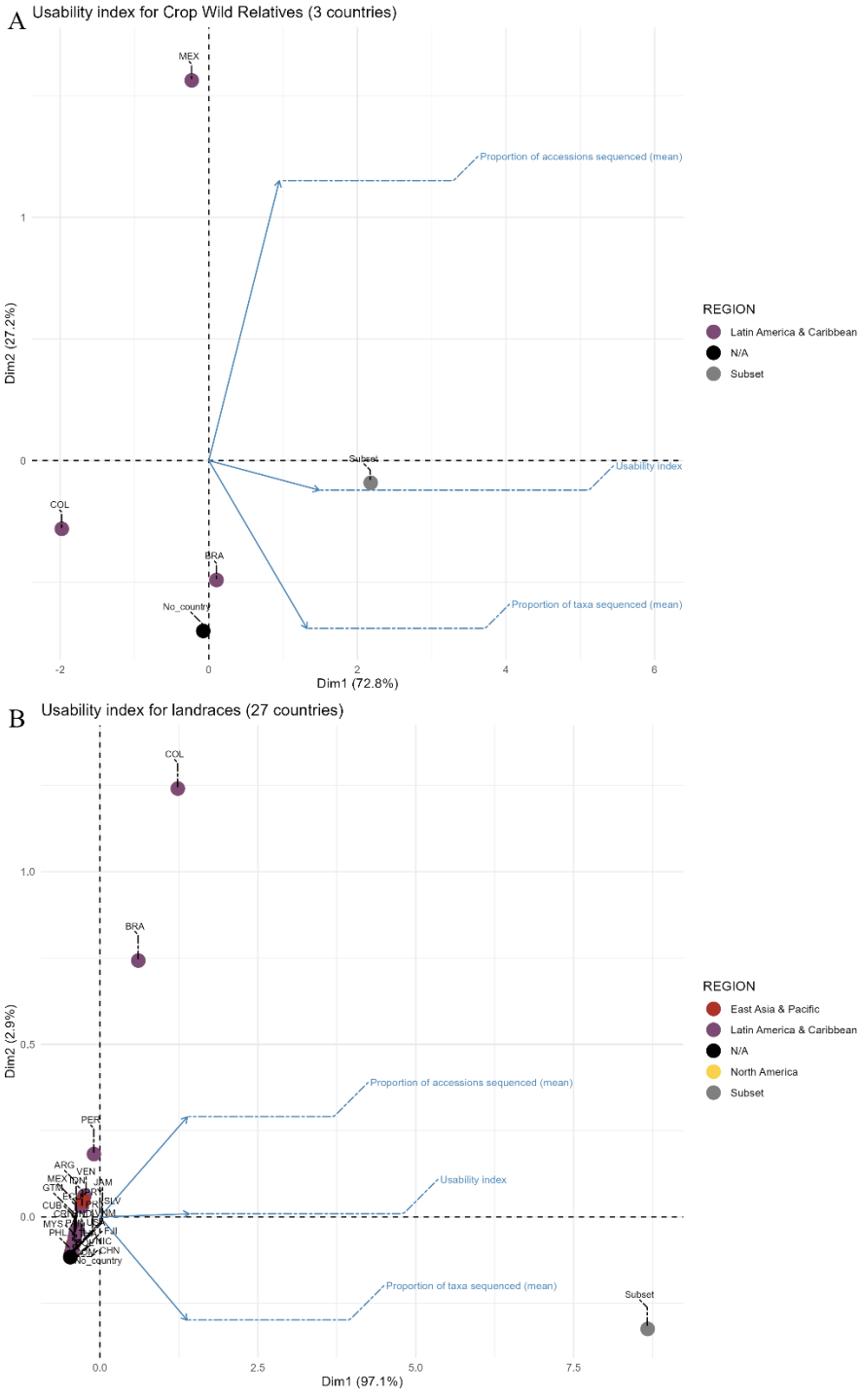
Supplementary Figure 3. NIPALS graphical representation of species diversity, equality, and the composition index for the crop wild relatives taxa in CIAT Cassava collection. Each color represents a world region according to World bank criteria. A) NIPALS representation using Margalef index B) NIPALS representation using Hill-Simpson index (Effective number of species).



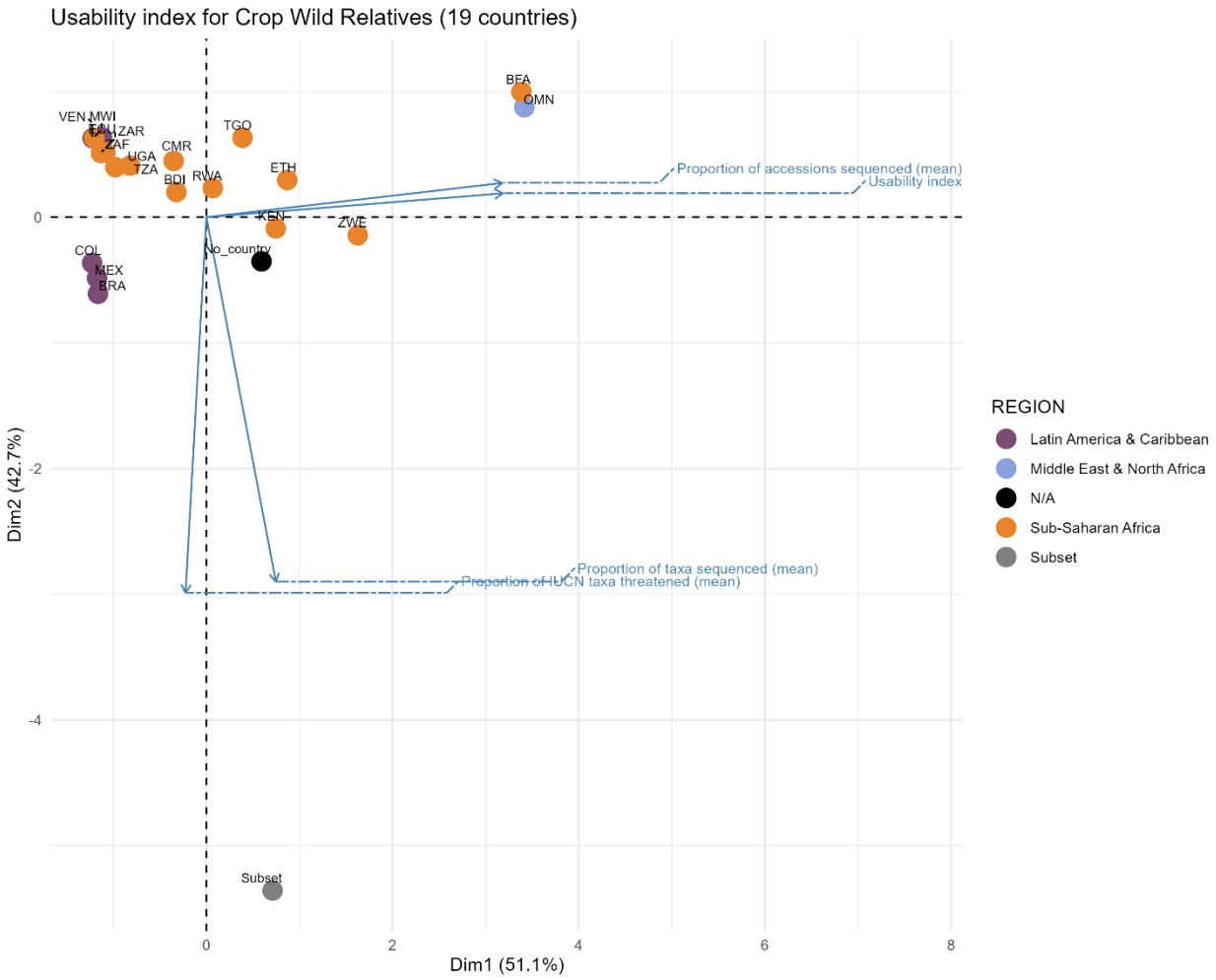
Supplementary Figure 4. NIPALS graphical representation of species diversity, equality, and the composition index for the crop wild relatives taxa in CIAT Forages collection. Each color represents a world region according to World bank criteria. A) NIPALS representation using Margalef index B) NIPALS representation using Hill-Simpson index (Effective number of species).



Supplementary Figure 5. NIPALS graphical representation of usability for the crop wild relatives and landraces taxa in CIAT Beans collection. Each color represents a world region according to World bank criteria. A) NIPALS representation for crop wild relatives B) NIPALS representation for landraces.



Supplementary Figure 6. NIPALS graphical representation of usability for the crop wild relatives and landraces taxa in CIAT Cassava collection. Each color represents a world region according to World bank criteria. A) NIPALS representation for crop wild relatives B) NIPALS representation for landraces.

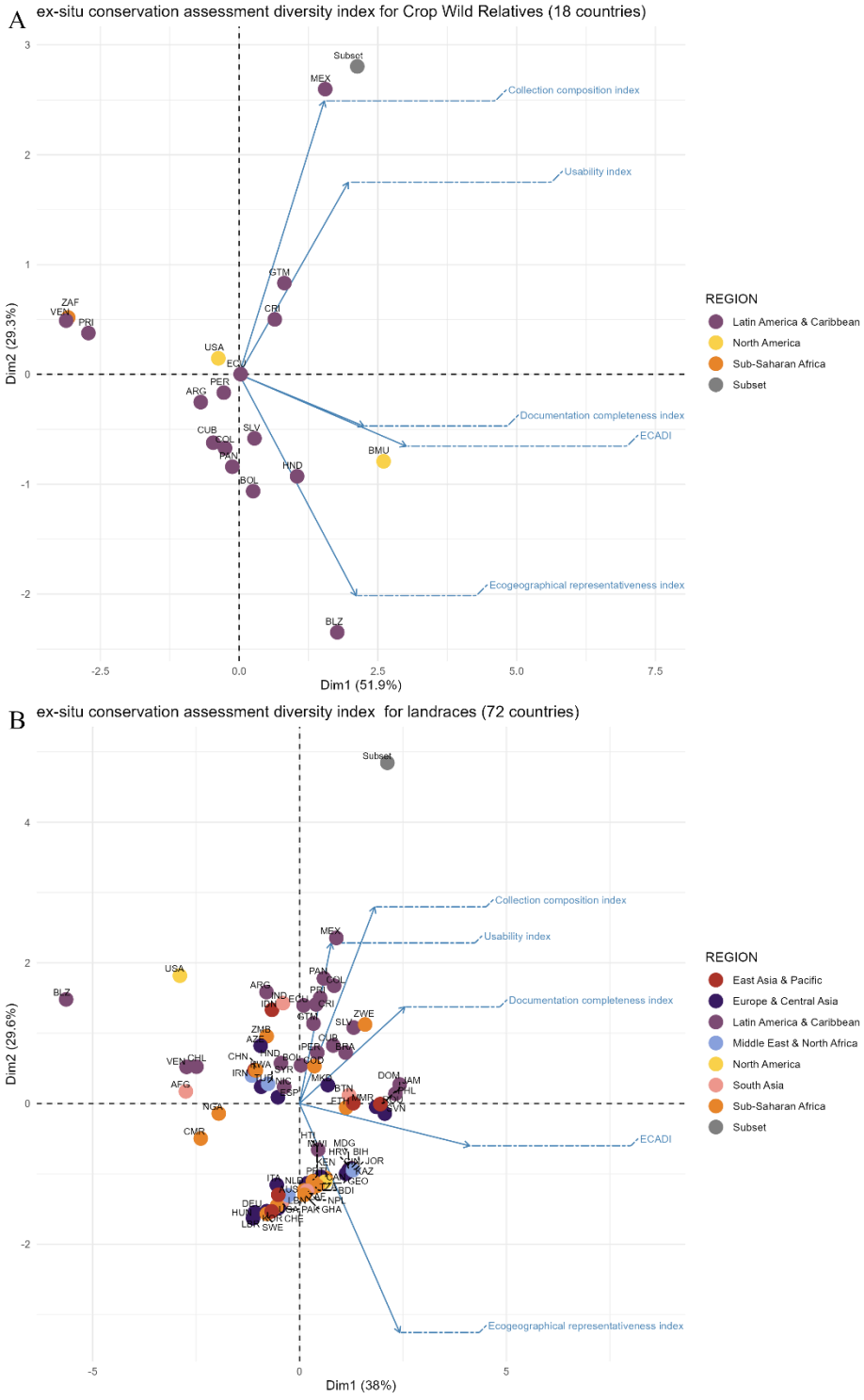


928

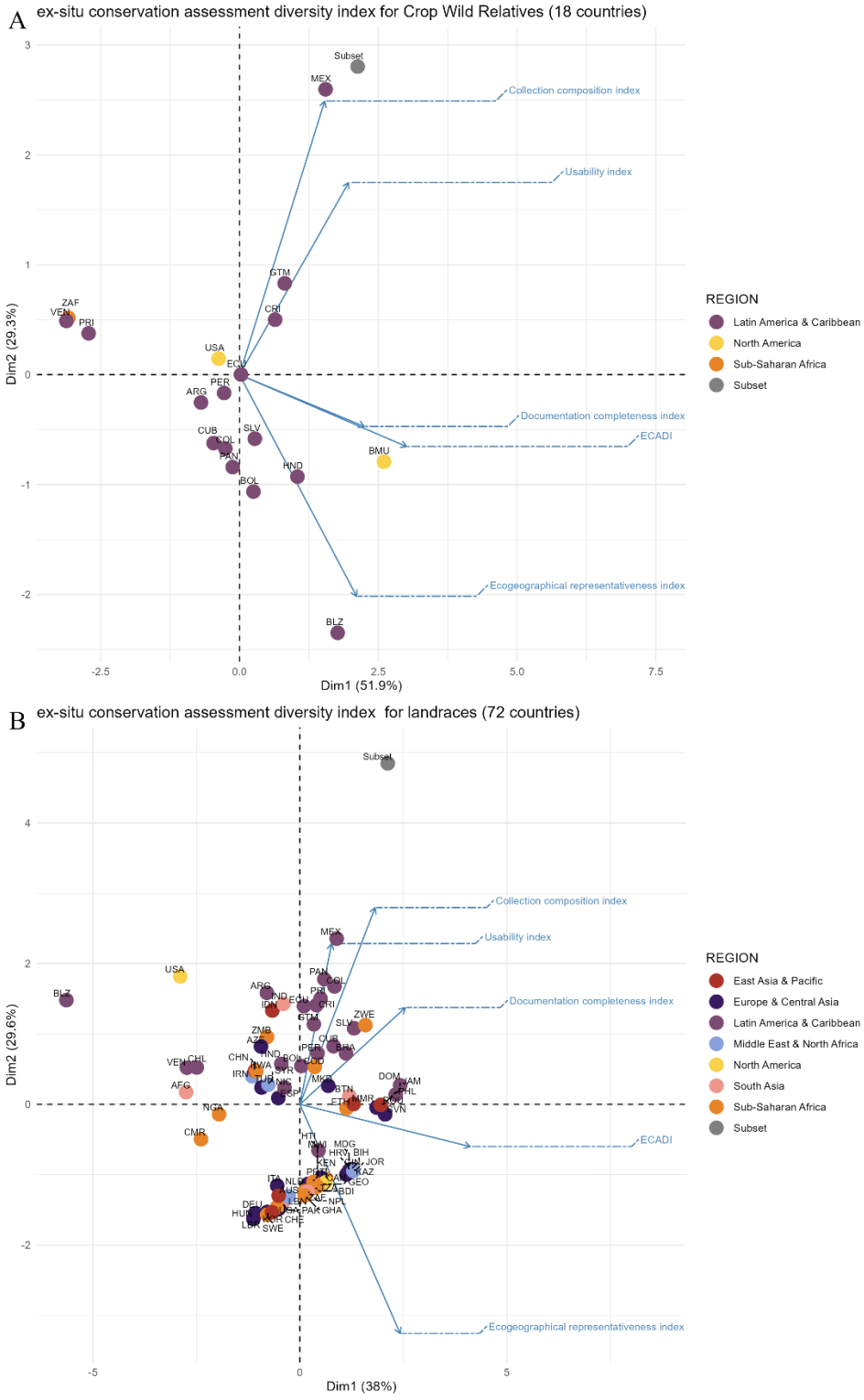
929 **Supplementary Figure 7.** NIPALS graphical representation of usability for the crop wild relatives taxa in CIAT
930 Forages collection. Each color represents a world region according to World bank criteria.
931

932

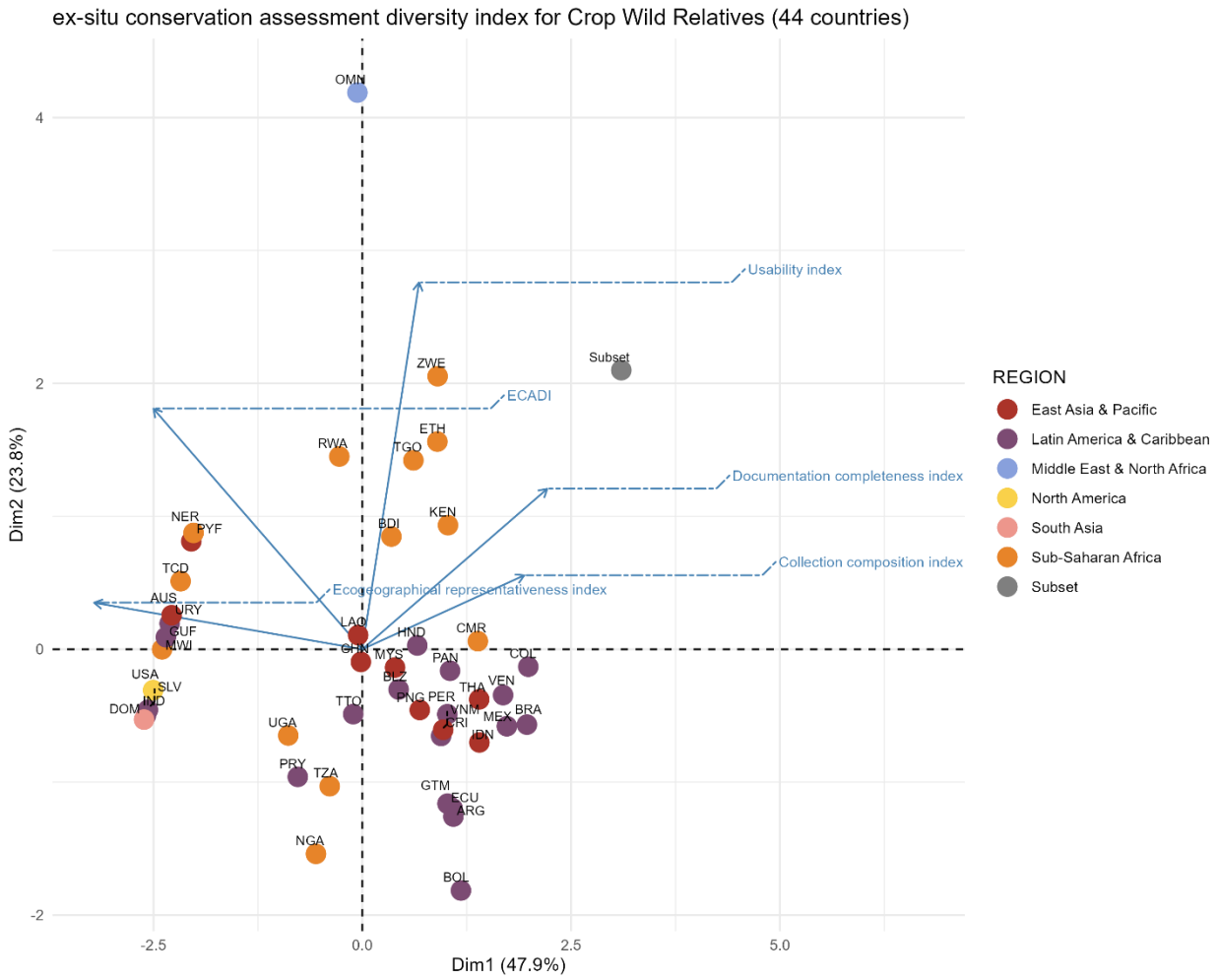
933



Supplementary Figure 8. NIPALS graphical representation of ECADI for the crop wild relatives and landraces taxa in CIAT Beans collection. Each color represents a world region according to World bank criteria. A) NIPALS representation for crop wild relatives B) NIPALS representation for landraces.



Supplementary Figure 9. NIPALS graphical representation of ECADI for the crop wild relatives and landraces taxa in CIAT Beans collection. Each color represents a world region according to World bank criteria. A) NIPALS representation for crop wild relatives B) NIPALS representation for landraces.



945

946 **Supplementary Figure 10.** NIPALS graphical representation of ECADI for the crop wild relatives taxa in CIAT
947 Forages collection. Each color represents a world region according to World bank criteria.
948