Genet Resour Crop Evol (2014) 61:841–851
DOI 10.1007/s10722-014-0079-4

RESEARCH ARTICLE

Assessing rice and wheat germplasm collections using similarity groups

Th. Hazekamp · T. S. Payne · N. R. Sackville Hamilton

Received: 31 August 2013 / Accepted: 7 January 2014 / Published online: 30 January 2014

Abstract Central crop databases or registries are important tools to enhance the use and conservation of plant genetic resources. In 2008–2009 a group of Centers from the Consultative Group on International Agricultural Research worked together on the development of central crop registries for eight different crops. The International Rice Research Institute and the International Maize and Wheat Improvement Center led the development of the crop registries for rice and wheat, respectively. The registries were to compile data across collections and add value to these datasets by assessing the similarity of the accessions from those collections. We describe in detail the methodology developed for the rice and wheat registries. This methodology mainly followed an algorithmic approach to assess the correspondence between pairs of accessions. Accessions which shared a common origin were placed together in similarity groups. Using these groups the similarity of accessions in and among collections was further analysed.

Keywords Central crop databases · Rice · Similarity assessment · Similarity groups · Wheat

Th. Hazekamp (✉)
c/o International Rice Research Institute (IRRI), DAPO Box 7777, 1301 Metro Manila, Philippines
e-mail: tom.hazekamp@gmail.com

T. S. Payne
Wheat Genetic Resources, Wellhausen Anderson Plant Genetic Resource Building, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, DF, Mexico

N. R. Sackville Hamilton
T.T. Chang Genetic Resources Center, International Rice Research Institute (IRRI), DAPO Box 7777, 1301 Metro Manila, Philippines

Introduction

Central crop databases are important tools to enhance the use and conservation of plant genetic resources at the crop level (Bommer 1991; Van Hintum 1997). Central crop databases have been developed for a large number of crops. Knüpffer (1995) listed 64 such databases. During 2008–2009 a number of Centers from the Consultative Group on International Agricultural Research (CGIAR) worked together on the development of eight central crop registries focussed on crops they hold in common. These include barley, cassava, chickpea, forages, *Musa*, potato, rice and wheat. This development was part of the World Bank funded Global Public Goods 2 project (GPG2) which aimed to enhance the CGIAR Centers' capacity to conserve and provide plant genetic resources and associated knowledge to users worldwide as Global Public Goods. The International Rice Research Institute (IRRI) and the International Maize and Wheat Improvement Center (CIMMYT) led the development of the global registries for rice and wheat, respectively,

building on their current systems known as the International Rice Information System (IRIS) and the International Wheat Information System (IWIS). Both are implementations of the International Crop Information System (ICIS). The registries not only compiled and published the passport data, but they also added value by standardizing common data fields, assess the similarity of accessions in and among collections, and fed data back to data providers in order to improve data quality. In this paper we describe the methodology that was used to assess similarity within the rice and wheat collections.

Methodology

Comprehensive passport data were available from seven collections of rice and were included in the rice registry representing a total of 223,397 accessions (Table 1). These included the five CGIAR-held rice collections plus the USDA National Plant Germplasm System (USDA/NPGS) and Chinese open rice collection. Together they represented 29 % of the estimated global holdings of *ex situ* conserved rice accessions (FAO 2010).

Three collections were selected for inclusion in the wheat registry: the CIMMYT, ICARDA and USDA/NPGS wheat collections. Together they totalled 193,635 accessions. This corresponded to 23 % of the estimated global holdings of *ex situ* conserved wheat accessions (Table 2).

For both rice and wheat all CGIAR collections were included. The USDA/NPGS collections were added as they are major collections by themselves and extensive germplasm exchange has occurred with the CGIAR collections. A comprehensive dataset on the Chinese open rice collection had been provided to IRRI by the Chinese Academy of Agricultural Sciences and was included as a more typical model of a national genebank. Also data on historical (i.e. accessions no longer present in the collections but still with records in the databases) and inactive accessions were included in addition to active accessions, as this helped establish linkages among accessions. In particular, the entire rice collection of IITA no longer exists as a collection at IITA. In the late 1980s, IITA gradually transferred its mandate for rice research and

Table 1 Rice collections included in the rice registry 2008–2009

Dataset name	Rice collections	Issue date of dataset used	Number of accessions
AfricaRice	Africa Rice Center (AfricaRice), Cotonou, Benin	12 Dec 2007	19,058
Chinese open	A subset of data open to the public on 13,944 accessions, was provided by the Informatics Center of the Institute of Crop Sciences of Chinese Academy of Agricultural Sciences (CAAS), Beijing, China, from the Chinese Crop Germplasm Information System. This represents approximately 25 % of the rice collection held in the Chinese national genebank at the Institute of Crop Germplasm Resources (CAAS)	8 Jun 2007	13,944
CIAT	International Center for Tropical Agriculture (CIAT), Cali, Colombia	4 Oct 2008	1,635
IITA	International Institute for Tropical Agriculture (IITA), Ibadan, Nigeria	28 April 1998	12,321
IRRI GRC	International Rice Research Institute (IRRI), Genetic Resources Center (GRC), Los Baños, Philippines	14 Mar 2008	117,272
IRRI INGER	IRRI International Network for Genetic Evaluation of Rice (INGER), Los Baños, Philippines	13 Mar 2008	24,716
USDA/NPGS	National Plant Germplasm System (USDA/NPGS), including the National Small Grains Collection at Aberdeen, Idaho, USA and the Rice Genetic Stocks Collection at Stuttgart, Arkansas, USA	26 Jan 2008	34,451
	Total accessions		223,397
	Estimated % of global holdings		(29 %)

breeding in Africa to WARDA (IITA 1991). As part of this transfer, the rice collection at IITA was handed over to WARDA (now AfricaRice). Including IITA data in the analysis was essential to establish possible linkages among accessions held at IRRI and AfricaRice.

Considering the large number of accessions that were assessed for possible common origins in a relatively short period of time, a methodology was developed that allowed us to assess the correspondence among pairs of accessions mainly algorithmically based on combinations of similarity scores. Table 3 lists the passport descriptors used and the type of comparisons made. The similarity scores were a combination of basic string comparisons and comparisons based on the Levenshtein or edit distance (Black 1999). The edit distance counts the number of character additions, deletions or substitutions needed to transform string A to string B. It can be used as a metric to describe the difference between two string values. Manual assessment was limited to the instances where the computed similarity scores alone did not provide clear enough evidence upon which to base a decision.

The comparisons were made by selecting accessions from two collections at the time. Figure 1 outlines the workflow that was followed for each assessment among collections.

Table 2 Wheat collections included in the wheat registry 2008-2009

Dataset name	Wheat collections	Issue date of dataset used	Number of accessions
CIMMYT	International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico	17 Nov 2008	97,641
ICARDA	International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria	26 May 2008	34,612
USDA/NPGS	USDA National Plant Germplasm System (USDA/NPGS) Small Grains Collection, Aberdeen, Idaho, USA	26 Jan 2008	61,382
	Total accessions		193,635
	Estimated % of global holdings		(23 %)

Table 3 Passport descriptors used for comparisons

Descriptor	Type of comparison	Similarity score
Species[a]	Basic string comparison	0 = no match, 1 = match
Collecting number	Basic string comparison	0 = no match, 1 = match
Collecting date of sample	Matching of year, month and day	Value between 0 and 1. For each matching part 0.33 is added to the score
Biological Status of accession	Basic string comparison	0 = no match, 1 = match
Country of origin code	Basic string comparison	0 = no match, 1 = match
Location of collection site	Relative Levenshtein distance[b]	Value between 0 and 1. Calculated as 1-(Levenshtein distance/Max Levenshtein distance)
Latitude and longitude of collecting site	Max difference between lat or Longitude	Value between 0 and 1 Calculated as 1-(max(Latitude difference (abs.) in degrees, Longitude difference (abs.) in degrees)/180)
Accession names	Relative Levenshtein distance[b]	Value between 0 and 1. Calculated as 1-(Levenshtein distance/Max Levenshtein distance)
Pedigree	Relative Levenshtein distance[b]	Value between 0 and 1. Calculated as 1-(Levenshtein distance/Max Levenshtein distance)

[a] The genus was not compared. All rice accessions were from genus *Oryza* while the wheat accessions were from genus *Triticum* only

[b] A Visual Basic routine from http://www.merriampark.com/ld.htm#VB was used to calculate the Levenshtein distance

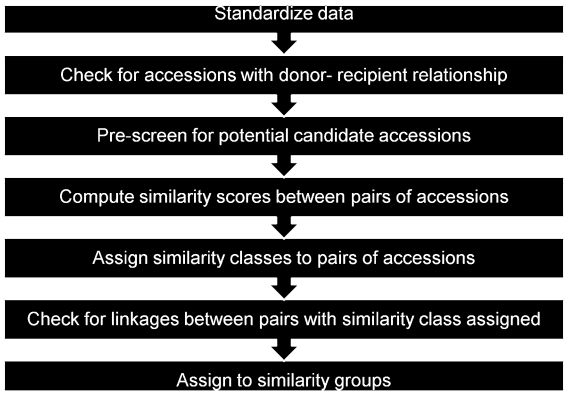

Fig. 1 Workflow similarity assessment

The data to be compared were standardized using the FAO/IPGRI Multi-crop Passport Descriptors (FAO/IPGRI 2001), and the name formatting rules used by the ICIS. This entailed checking that the country codes conformed with the ISO-3166 standard (ISO 2013a), biological status codes were valid, dates were converted to ISO 8601 format (ISO 2013b), latitude and longitude were converted to decimal values. The specific epithet of each accession was checked for spelling mistakes only. Name and identifier data were formatted using the ICIS "standard name routine" to remove leading zeros, extra spaces, etc. (IRRI 2010).

Whenever multiple formats were used to refer to accession numbers, they were reformatted to a single format (e.g. references to "IRRI nnnn", "IRRI-IRGC-nnnn" or "IRGC-nnnn" were all reformatted to "IRGC nnnn"). These modifications were made in separate columns, keeping a copy of the original values as a reference.

The standardized data were first searched for unequivocal evidence of direct donor-recipient relationships, i.e. the recipient's donor accession ID matched an accession ID in the specified donor's data. If indeed such a relationship was found, these pairs were marked as similar.

To reduce the overall number of pair-wise combinations among accessions that would have to be analysed in detail for similarity, the remaining accessions were submitted to a pre-screening protocol. The pre-screening assessed whether any of the standardized names or identifiers given to an accession suggested a resemblance with any other accession. This was considered a minimum requirement as, without any such evidence, accessions would never be classed as similar. The actual assessment score was calculated as 1-(Levenshtein distance/Maximum Levenshtein distance) whereby the Maximum Levenshtein distance equalled the length of the longest of the two strings which were compared. A threshold score was determined by visual inspection of the resulting scores. The scores were sorted high-to-low and by means of a visually inspection a threshold value was determined under which it was judged that no meaningful similarity would be found anymore. The pairs of accessions for which this score exceeded the threshold value were selected for further analysis. For these we calculated the similarity scores for the passport descriptors listed in Table 3 in as far these data were available. No similarity score was assigned when one of the descriptor values was missing.

The results were tabulated, with each row representing one pair of accessions. The table included columns for the two accession identifiers and, for each descriptor selected, the two descriptor values plus the calculated similarity score. Additional columns contained a number of derived statistics such as mean, standard deviation and number of similarity scores calculated per pair. These statistics were used to give added possibilities to filter and order the data. Based on the results of the similarity comparisons, putative relationships between the paired accessions were determined. Depending on the nature of the similarity, pairs of accessions were classified using the similarity classes as listed in Table 4. In making these decisions a very conservative approach was taken whereby if there was doubt about accessions' similarity, they were identified for manual re-inspection. If manual inspection was not conclusive, no similarity class was assigned.

The tentative results of the comparisons were sent for feedback to the organizations which originally provided the data.

All similar pairs, i.e. pairs which had been assigned a similarity class of SC1-3 or SMAN, were processed and placed in similarity groups. Part of this processing was to check for linkages between accession pairs and to merge groups if warranted. Pairs were merged if the accessions in a hypothetical pair A-B were similar AND the accessions B–C were similar, A and C were thus considered to also be similar resulting in A, B and C belonging to the same similarity group.

Table 4 Similarity classes for pairs of accessions

Class	Description
SC1	Similarity Class 1: Strong evidence for a direct provider-recipient linkage between the germplasm accessions based on the same donor accession ID and donor institution
SC2	Similarity Class 2: Medium strength evidence for a direct provider-recipient linkage among the accessions, where an institute indicates it had received an accession from the other institute, but the exact donor accession identifier was missing, and the two accessions had the same values for other key passport data descriptors (e.g. accession name, collecting number)
SC3	Similarity Class 3: No evidence for a direct provider-recipient linkage, but other evidence suggests the accessions may have a common origin based on similarity measures for other key descriptors (e.g. collector's sample ID, accession names, country of origin and pedigree)
SMAN	Pairs which were identified during a manual inspection of the data as similar yet were not assigned in any of the other similarity classes

For the cluster analysis of similarity between the rice and wheat collections the Cluster and Treeview software developed by the Eisen Lab (Lawrence Berkeley National Lab 2013) was used.

Results

Rice collections

For rice, 223,397 accessions were analysed. There were 64,057 (=28.7 %) accessions classified as similar to one or more other accessions in the collections. For approximately one third of these accessions, a direct donor-recipient relationship was established. The remainder were classified as similar using other passport data. The accessions were grouped in 25,235 similarity groups. The size of these similarity groups ranged from two to 28 accessions (Fig. 2). Two-thirds of the similarity groups consisted of two accessions only. Perhaps unsurprisingly, the most duplicated accessions are the most successful old IRRI-bred varieties IR 8, IR 36, IR 20 and IR 24, represented in all the collections.

To find out more about the similarity among accessions within a specific collection, we looked at the similarity groups which contained two or more

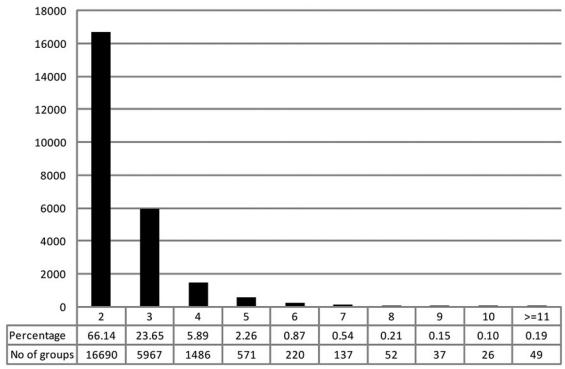

Fig. 2 Frequency distribution of similarity group size for the rice collections reviewed

accessions from the same collection. Table 5 lists the number of similarity groups and accessions involved. Striking differences were observed among the collections. The IITA collection had a relative high number of accessions sharing the same similarity groups. Inspection of scanned copies of IITA's original collecting forms revealed handwritten notes listing the components of mixed samples and the accession ID assigned to each. This indicated that the practice at IITA was to split mixed or heterogeneous collected samples into their components, creating a different accession for each component. Thus, although the passport data showed they had a common origin from a single collected sample and were thus placed in the same similarity group, they are genetically distinct.

The USDA/NPGS collection also contained a relatively high percentage of similar accessions. However, this was related to the fact that the USDA/NPGS data set also contained data on non-active, working, core and quarantine collections. For example, the USDA/NPGS genetic stocks of *Oryza* (the GSOR accessions) are pure-line selections from the original, heterogeneous introductions which were assigned PI numbers. Being pure-line selections, the GSOR accessions are not genetic duplicates of their corresponding PI accessions.

To assess the similarity of accessions among rice collections, we selected the similarity groups that contained accessions from more than one collection. Table 6 lists the number of accessions from a particular collection (row value) which shared similarity groups with another collection (column value). The AfricaRice collection had 526 accessions in similarity groups which also contained one or more accessions

Table 5 Similarity groups containing multiple accessions from the same collection

Collection name	Number of similarity groups with 2 or more accessions from same collection	Number of accessions	Collection size	Percentage of accessions relative to collection size
AfricaRice	252	553	19,058	2.90
Chinese open	63	126	13,944	0.90
CIAT	3	6	1,635	0.37
IITA	1,435	3,738	12,321	30.34
IRRI GRC	852	1,833	117,272	1.56
IRRI INGER	70	144	24,716	0.58
USDA/NPGS	2,145	5,014	34,451	14.55
Total		11,414	223,397	5.10

Table 6 Number of accessions from a collection (row) sharing similarity groups with accessions from another collection (column)

Accessions from	Share group with						
	Africa Rice	Chinese open	CIAT	IITA	IRRI GRC	IRRI INGER	USDA/NPGS
AfricaRice		526	152	7,875	3,844	1,779	722
Chinese open	527		112	366	1,925	1,559	744
CIAT	146	104		60	226	313	1,346
IITA	9,790	399	67		5,144	354	588
IRRI GRC	4,037	1,996	257	4,744		3,298	10,664
IRRI INGER	1,777	1,559	328	329	3,223		1,830
USDA/NPGS	1,105	1,270	1,850	934	12,649	2,741	

from the Chinese open collection. The Chinese open collection had 527 accessions in similarity groups with accessions from the AfricaRice collection. Possibly, there were also accessions from other collections in the similarity groups these two collections shared.

AfricaRice had its highest number of accessions (7,875) in similarity groups with IITA. IITA also had its highest number of accessions (9,790) in groups with AfricaRice. This reciprocal relationship indicated that, within the whole set of collections reviewed, these two collections were the closest related to each other, and reflected the transfer of the IITA collection to AfricaRice. The same type of reciprocal relationship was observed between the USDA/NPGS and IRRI GRC collections. However, the CIAT collection had the highest number of accessions (1,346) in similarity groups together with the USDA/NPGS collection. The USDA/NPGS collection, however, has the highest number of accessions (12,649) in groups with IRRI GRC. So although the CIAT collection shared most with the USDA/NPGS collection, this relationship is not reciprocal.

For each pair of collections the total number of accessions both institutes had in those groups was divided by the size of both collections to obtain a relative proportion (Table 7).

The IITA and AfricaRice rice collections had the highest relative proportion of accessions sharing similarity groups (56.30 %) (Table 7), reflecting the fact that AfricaRice took over the IITA collection. The second highest are the USDA/NPGS and IRRI GRC rice collections (15.37 %). We visualised the degree of (dis)similarity between the collections by means of a cluster analysis (Fig. 3). The relative proportion of accessions in joint similarity groups (Table 7) provided a measure of "similarity" between collections. For the cluster analysis, these were converted to distance scores (=1-"similarity"). Figure 3 shows the resulting dendrogram which clearly demonstrates the clustering of the AfricaRice and IITA collections, and

Table 7 Relative proportion of accessions in joint similarity groups (%)

	Africa Rice	Chinese open	CIAT	IITA	IRRI GRC	IRRI INGER	USDA/NPGS
Africa Rice	2.90	3.19	1.44	56.30	5.78	8.12	3.41
Chinese open		0.90	1.39	2.91	2.99	8.07	4.16
CIAT			0.37	0.91	0.41	2.43	8.86
IITA				30.34	7.63	1.84	3.25
IRRI GRC					1.56	4.59	15.37
IRRI INGER						0.58	7.73
USDA/NPGS							14.55

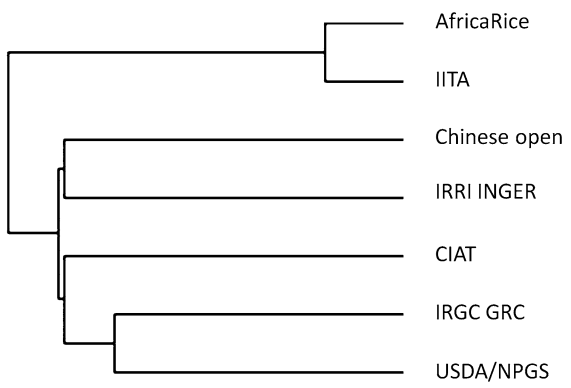

Fig. 3 Cluster analysis of similarity among seven rice collections based on passport data

to a lesser degree, the IRRI GRC and USDA/NPGS collections.

Table 8 lists the total number of accessions found in similarity groups for each collection. This table also shows the percentage of accessions in similarity groups in relation to the collection size. These figures show that the IRRI GRC collection had the largest (absolute) number of accessions in similarity groups. However, taking into consideration the size of this collection it accounts for the smallest relative fraction of the seven collections. A large part of its collection does not seem to be present in any of the six other collections. The IITA collection had the largest relative fraction of accessions in similarity groups. As previously noted, this not only reflected the similarity with many AfricaRice accessions (Table 7), but also includes a fair number cases where multiple IITA accessions shared the same groups (Table 5). The figure for the USDA/NPGS collection is also somewhat inflated due to the fact that the dataset contained data on a number of separate collections which partly overlap. If we consider the high score for the CIAT collection, we see that it shares more than 80 % (1,346 out of 1,635 accessions) of its collection with the USDA/NPGS collection (Table 6).

The data on which this study was based have been incorporated into the IRIS (IRRI 2013).

Wheat collections

Out of the 193,635 accessions analysed, 63,219 (=32.6 %) accessions were classified as being similar to one or more other accessions in the collections reviewed. For just over half of these accessions, a direct donor-recipient relationship could be established. The remaining accessions were classified as similar using other passport data. Similar accessions were grouped in 26,396 similarity groups. The size of these similarity groups ranged from two to 65 accessions. Figure 4 provides an overview of the frequency distribution of the size (i.e. number of accessions) of the similarity groups. Nearly 76 percent of the similarity groups consisted of two accessions only.

To ascertain similarity among accessions within a specific collection, we determined how often two or more accessions from the same collection were classified in the same similarity group (Table 9).

The ICARDA wheat collection listed the largest number of accessions clustered in the same similarity groups (Table 9). Inspection of the ICARDA data showed that accessions sharing the same passport data are often found in batches together (e.g. IG 89396 and IG 90010). ICARDA indeed confirmed that in particular cases single plant selections are made and stored as separate accessions.

To find out more about the similarity of accessions among collections, we looked at the similarity groups

that contained accessions from more than one collection (Table 10). The CIMMYT wheat collection had 6,569 accessions sharing similarity groups with 8,159 accessions of the ICARDA wheat collection. The CIMMYT and USDA/NPGS wheat collections shared more accessions in similarity groups than with the ICARDA wheat collection.

Table 8 Accessions in similarity groups

Collection	Accessions in similarity groups	Collection size	Percentage
AfricaRice	9,619	19,058	51
Chinese open	2,981	13,944	22
CIAT	1,384	1,635	85
IITA	10,781	12,321	88
IRRI GRC	18,035	117,272	16
IRRI INGER	6,204	24,716	26
USDA/NPGS	15,053	34,451	44
Total	64,057	223,397	29

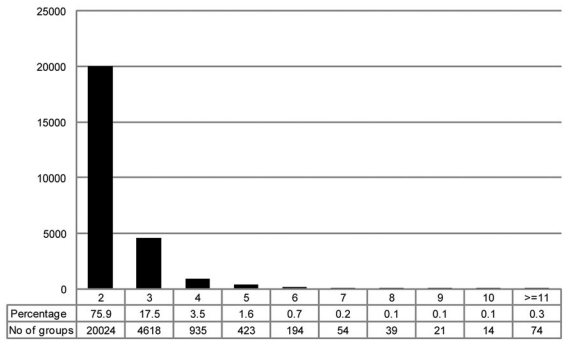

Fig. 4 Frequency distribution of similarity group size for the wheat collections reviewed

For each pair of collections the total number of accessions both institutes had in those groups (Table 10) was divided by the size of both collections (Table 2) to obtain a relative proportion (Table 11).

We visualised the degree of (dis)similarity between the collections by means of a cluster analysis (Fig. 5).The relative proportion of accessions in joint similarity groups (Table 11) provided a measure of "similarity" between collections. For the cluster analysis these were converted to distance scores (=1-"similarity").

The CIMMYT and USDA/NPGS wheat collections shared the largest overlap in terms of similarity (Table 11, Fig. 5). While both CIMMYT and ICARDA appeared to have a similarity with USDA/NPGS wheat collection, the similarity between the two CGIAR collections was the smallest calculated among the wheat collections compared. This indicated that the two CGIAR collections complement each other well.

Table 12 lists the total number of accessions found in similarity groups for each collection reviewed. The table also lists the percentage of accessions in similarity groups in relation to the collection size. The USDA/NPGS wheat collection had the most accessions, absolute and relative, in similarity groups.

Discussion and conclusions

Central crop databases such as the IRIS and the IWIS are important tools to facilitate the use and conservation of germplasm at the crop level. The assessment of similarity among the accessions contained in such systems adds value for the users, civil society, donors and managers of the collections.

Ex situ collections' similarity analysis is important for diversity assessment and collection management. For diversity assessment, similarity analysis can affect

Table 9 Similarity groups containing multiple accessions from the same wheat collection

Collection name	Number of similarity groups with multiple accessions from same collection	Number of accessions involved	Collection size	Percentage
CIMMYT	664	1,531	97,641	1.56
ICARDA	1,847	4,663	34,612	13.47
USDA/NPGS	865	2,511	61,382	4.09

Table 10 Number of accessions from a wheat collection (row) sharing similarity groups with another collection (column)

Accessions from:	Share group with		
	CIMMYT	ICARDA	USDA/NPGS
CIMMYT		6,569	23,376
ICARDA	8,159		10,890
USDA/NPGS	24,046	8,972	

Table 11 Relative proportion of wheat accessions in joint similarity groups (%)

	CIMMYT	ICARDA	USDA/NPGS
CIMMYT	1.56	11.14	29.82
ICARDA		13.47	20.69
USDA/NPGS			4.09

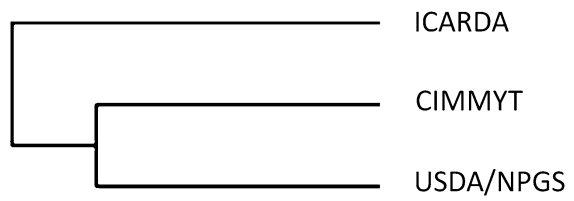

Fig. 5 Cluster analysis of similarity between three wheat collections based on passport data

Table 12 Wheat accessions in similarity groups

Collection	Accessions in similarity groups	Collection size	Percentage
CIMMYT	24,266	97,641	24.8
ICARDA	11,800	34,612	34.1
USDA/NPGS	27,153	61,382	44.2
Total	63,219	193,635	32.6

a global assessment of conservation gaps and overlaps, and can be used to help select unique germplasm to focus limited resources on genotyping or phenotyping. A collection's management can be assisted by similarity analysis leading to a better understanding of under-utilized (unique) materials within the collection, the verification of the accessions' integrity over time and space, and the rationalization of germplasm distribution by distributing from the genebank nearest to a user or by identifying alternative suppliers when material is temporarily not available from the nearest genebank.

This study describes the methodology to perform a similarity assessment for a large number of accessions. In total, 417,032 accessions were reviewed; 223,397 for rice and 193,635 for wheat. The methodology uses similarity scores to classify accessions and only resorts to the manual screening of accession pairs when the computed scores alone cannot provide enough support to reach a decision on whether the accessions should be classified as similar or not. The methodology worked quite well. For wheat, 97 % of the matched pairs were classified as similar based on the computed similarity scores and only 3 % were classified as similar as the result of a manual assessment. For rice the percentages were 91 and 9 %, respectively.

With the use of only passport data we were able to assign 29 % of the rice accessions and 32 % of the wheat accessions into similarity groups. The similarity group sizes ranged from two to 28 accessions for rice and from two to 65 accessions for wheat. This indicates that there were sometimes many similar accessions within the same collection. Multiple introductions of the same germplasm (duplicate introductions) may account for some of these results, but more often this seems to be related to the introduction of selections from a single accession. Using passport data alone, it often was not possible to differentiate between these two possibilities. This is an important limitation to this type of analysis. The complementary use of characterization, evaluation or molecular data could help to differentiate better between accessions as shown by van Hintum and Visser (1995).

We used passport data to establish which accessions had a common origin, but this does not necessarily imply that they are biological duplicates. They may be unintentionally different (through drift or selection, or contamination with the wrong pollen or seed, or mislabelling). More significantly, they may be intentionally different when one accession is derived from another by selecting one component out of an original heterogeneous sample.

In fact, a large proportion of the similarities identified here represent pairs of accessions that have identical passport data but are genetically distinct by design. IITA and sometimes ICARDA split heterogeneous samples into their components, creating a different accession for each component, with all

components sharing identical passport data. USDA purified accessions for the USDA core collection by selecting a single plant from each original heterogeneous accession, and storing the purified sample as a separate genetic stock.

This highlights a significant omission from the Multi-Crop Passport Descriptors, namely that they do not provide a standard for documenting how one accession is derived from another. This is critically important information for interpreting similarities in passport data and would merit a less cursory and separate treatment from the "Ancestral data" descriptor as is currently the case. It should consist of a standardized description, following a controlled vocabulary, of the type of manipulation that took place in deriving the new accession from its source material(s). The controlled vocabulary should include the main alternatives that affect the genetic composition of the new accession such as "single seed descent", "single plant selection", "selected one component of a mixture", "random bulk". A broader controlled vocabulary, including for example hybridization and mutants, is included in the "method of germplasm creation" of the ICIS (IRRI 2010).

This study also highlights the importance of accurately maintaining the donor's accession number (DONORNUMB) of the MCPD. While a receiving genebank must assign its own new accession number to refer to the copy of the accession under its management, to distinguish its sample from the donor's sample, tracing accessions requires the donor's accession ID to be recorded alongside the receiving genebank's accession. Virtually all the uncertainties in this analysis have arisen from the cases where the receiving curator has not followed this standard. It also stresses the importance of persistent, globally unique accession identifiers. Current accession identifiers are assigned to be unique within a certain collection only. Many genebanks just use a number, or prefix a number with an uninformative code like "ACC". The same identifier might also be used in another collection to represent a different accession. This ambiguity could be eliminated through the use of globally unique accession identifiers. Another aspect to consider is the persistency of accession identifiers. It is bad practice to ever change accession identifiers. However, this does occur when germplasm collections are re-organized physically or administratively. As a result, the link to the previously used identifier(s) is obscured or sometimes lost. A persistent identifier eliminates this problem. To avoid any temptation to ever change the identifier it should be semantically coded to avoid falsifiable meaning. An example of deprecated usage is the "TOG" prefix to indicate *Oryza glaberrima*: if the sample is found to have been incorrectly identified, the prefix has to be changed. The current implementation of accession identifiers not only complicates the creation and maintenance of crop, regional or global registries where data comes together from different sources. The implementation is far from ideal to support the ability to unambiguously refer to an accession. The discussions about globally unique and persistent accession identifiers are not new (Knüpffer et al. 2007), but have not (yet) resulted in a widespread implementation.

The efforts needed to establish and maintain a central crop registry are substantial. Once established, such a resource has a large payback potential to enable more effective and efficient use and management of the available germplasm. For example, it is theoretically possible to improve efficiency through sharing responsibilities genebanks, but this is possible only if the correspondence between accessions in the collaborating genebanks has been established—i.e. if the crop registry has been established. In addition, there is increasing pressure on genebanks to save money by rationalising their collections, eliminating duplicates. It is obvious, and has been clearly demonstrated, that demonstrating a common ancestor of two accessions does not demonstrate that they are duplicates; but this information can be used to efficiently prioritise pairs of possible duplicates for more detailed molecular assessment of their genetic similarity. Thirdly, some of the new data were already used by the contributing partners to improve their own data quality: Once two accessions are determined to have originated in the same collected sample, the remaining fields in the two sets of passport data can be compared, and omissions or errors in one set can be corrected using the other set.

We conclude that analyses like that presented here represent only a first step in rationalizing *ex situ* conservation and use. The similarity groups must not be interpreted as duplicate accessions, but as a group of related accessions which share a common origin.

Acknowledgments This study was based on work performed by the International Rice Research Institute (IRRI) and the International Maize and Wheat Improvement Center (CIMMYT) in 2008 and 2009 and funded by the Global Public Goods 2 project.

References

Black PE (ed) (1999) Algorithms and theory of computation handbook, CRC Press LLC, "Levenshtein distance", in dictionary of algorithms and data structures [online], U.S. National Institute of Standards and Technology. 14 August 2008. http://www.nist.gov/dads/HTML/Levenshtein.html. Accessed 20 June 2013

Bommer DFR (1991) The historical development of international collaboration in plant genetic resources. In: Van Hintum ThJL, Frese L, Perret PM (eds) Crop Networks. Searching for New Concepts for Collaborative Genetic Resources Management. Papers of the EUCARPIA/IBPGR symposium held in Wageningen, The Netherlands, 3-6 December 1990. International Crop Network Series No. 4. International Board for Plant Genetic Resources, Rome, p 3–12

FAO (2010) The second report on the state of the world's plant genetic resources for food and agriculture. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy

FAO/IPGRI (2001) FAO/IPGRI multi-crop passport descriptors, December 2001. Alercia A; Diulgheroff S; Metz T (eds). Food and Agriculture Organization of the United Nations (FAO) and International Plant Genetic Resources Institute (IPGRI), Rome, Italy

IITA (1991) Annual Report (1989/1990). International Institute of Tropical Agriculture, Oyo Road, PMB 5320 Ibadan, Nigeria. ISSN 0331-4340

IRRI (2010) ICIS TDM Genealogy Management System. http://cropwiki.irri.org/icis/index.php/TDM_Genealogy_Management_System#Name_Standardization. Accessed 10 June 2013

IRRI (2013) International Rice Information System (IRIS). http://iris.irri.org/germplasm/. Accessed 10 June 2013

ISO (2013a) Country Codes—ISO 3166. http://www.iso.org/iso/country_codes. Accessed 2 Dec 2013

ISO (2013b) Date and Time Format- ISO 8601. http://www.iso.org/iso/home/standards/iso8601.htm. Accessed 2 Dec 2013

Knüpffer H (1995) Central Crop Databases. Standardization in Plant Genetic Resources Documentation. In: van Hintum ThJL, Jongen MWM, Hazekamp Th (eds) Report of the Second Technical Meeting of Focal Points for Documentation in East European Genebanks. CGN, Wageningen, The Netherlands, p 51–62

Knüpffer H, Endresen DTF, Faberová I, Gaiji S (2007) Integrating genebanks into biodiversity information networks. http://www.academia.edu/2693449/Integrating_Genebanks_Into_Biodiversity_Information_Networks. Accessed 2 Dec 2013

Lawrence Berkeley National Lab (2013) Eisen Software. http://rana.lbl.gov/EisenSoftware.htm. Accessed 31 May 2013

Van Hintum ThJL (1997) Central Crop Databases—an overview. In: Lipman E; Jongen MWM; van Hintum ThJL, Gass T, Maggioni L (eds).Central Crop Databases: tools for plant genetic resources management. International Plant Genetic Resources Institute, Rome, Italy/CGN, Wageningen, The Netherlands, p 17–19

Van Hintum ThJL, Visser DL (1995) Duplication within and between germplasm collections. II Duplication in four European barley collections. Genet Resour Crop Evol 42:135–145