Mapping genebank collections

by **Ma. Cristina Paule, Jane Girly Cuerdo, Mary Anne Reyes, Arnel Rala, Jacob van Etten, Andrew Nelson,** and **Robert J. Hijmans**

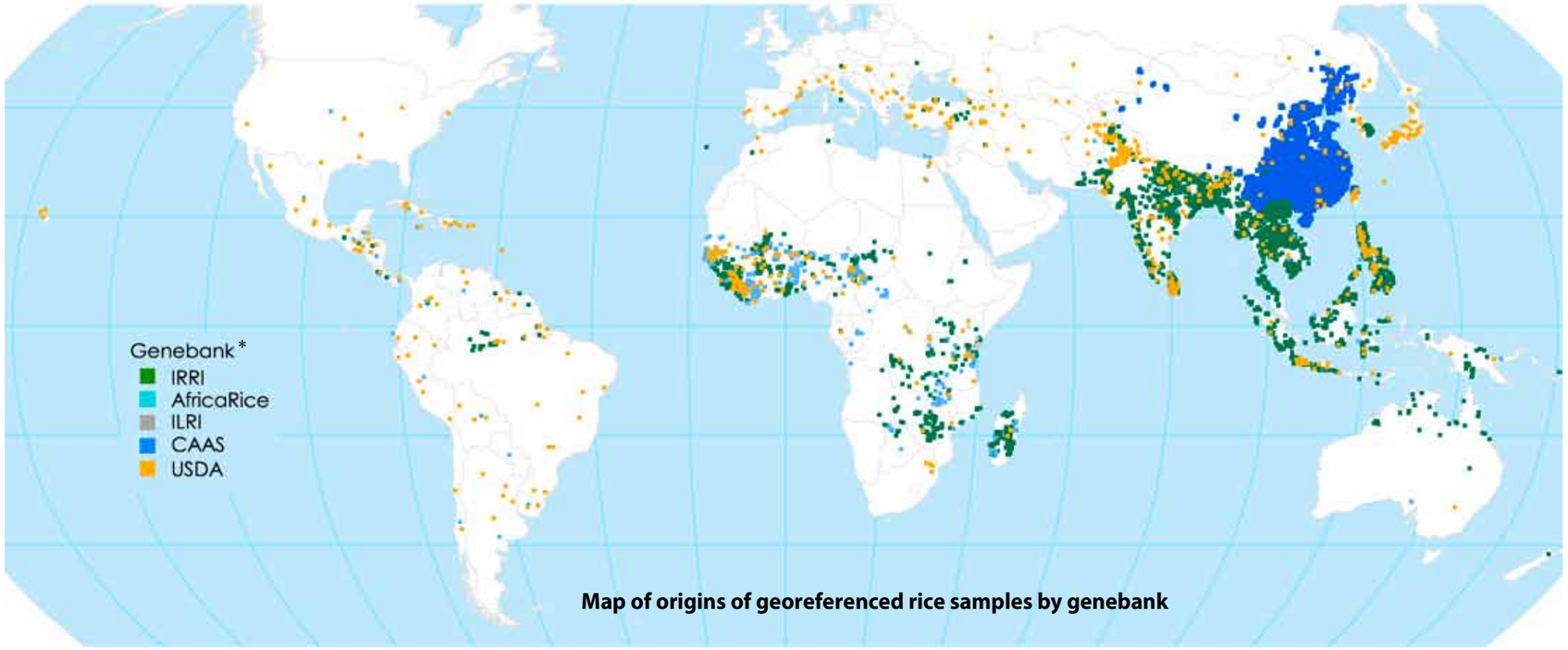
The genetic diversity in traditional and modern rice varieties, and in the wild relatives of rice, is vital to agricultural development, as it provides the basis for rice improvement. Sometimes, a single rare gene found in only a few varieties can greatly contribute to rice production. A good example of this is the short-straw gene or the *SUB1* gene that confers submergence tolerance.

Ironically, the success of rice breeding has rapidly replaced traditional varieties with modern ones, particularly in Asia's irrigated rice fields. Thus, the use of crop diversity bites its own tail, becoming a threat to its persistence. One important approach to avoid the loss of crop biodiversity, or "genetic erosion," is the use of genebanks. A genebank assembles a broad collection of rice varieties and conserves them for breeding. In total, there are 780,551 samples of rice in genebanks. The International Rice Research Institute (IRRI) has 109,136 samples, while genebanks in India and China have 86,119 and 70,104 samples, respectively.[1]

The samples in genebanks, and their associated data, can also be used to further explore and discover geographic patterns in crop diversity. In fact, these patterns are used to reconstruct the prehistorical origins and spread of crops. These data can be used to narrow down the geographical search area in finding samples with important traits, such as a certain type of disease resistance. The data also serve to identify gaps in a genebank collection, for which collecting expeditions are still needed.

Such analyses are possible only if we have a decent database in which each sample is described. Crop descriptors can include morphological characteristics, agronomic traits, and genetic data. For geographic analysis, it is crucial also to know where the sample was taken from.

Fortunately, "passport" information such as name of species, common name, institute code, date of acquisition, and location is generally available in genebank databases. However, for many samples, no geographic coordinates are available. Most samples were collected before the global positioning system became available. So, their locations were simply expressed in text format. At IRRI, we have worked on improving this situation to refine the Institute's genebank and also other genebanks. We have used tools such as biogeomancer[2] to assign coordinates to genebank samples, using their available text description as references. This is called "georeferencing" (see box). We also cross-checked the coordinates and corrected those that were found to be wrong.



Map of origins of georeferenced rice samples by genebank

***IRRI** = International Rice Research Institute, **AfricaRice** = Africa Rice Center (formerly known as WARDA), **CAAS** = Chinese Academy of Agricultural Sciences, **USDA** = United States Department of Agriculture, **ILRI** = International Livestock Research Institute.

Georeferencing. This refers to the process of converting text descriptions of locations to geographic coordinates that can be used in an analysis. This involves breaking down a description such as "14 kilometers north and 2 kilometers west of La Paz, Tarlac, Luzon, Philippines" into its geographic components, which can then be used to assign a coordinate to the record—in this case, 15.578N, 120.704E. However, processing such information is not always as straightforward as this example. Records often struggle with variations in spellings, obsolete names, and sometimes misspelled entries.

Dealing with inaccurate or wrong location data. Records that do have location data are also checked for errors. Some common errors are mistaking longitude for latitude (a mistake probably made from time to time); swapping east with west or north with south; missing a latitude or longitude coordinate; truncating the decimal points in a coordinate, resulting in lower precision; or simply putting in the wrong data. Not all errors can easily be corrected, but we have corrected those coordinates that fell in an ocean, or in the wrong country. Needless to say, determining the most likely location for an accession is a time-consuming task.

We have now almost completed georeferencing the IRRI collection. This map illustrates the progress we have made so far. We have also included data for some other rice collections.[3]

The coverage of the rice collections reflects the global distribution of rice cultivation and diversity, which were found to be highest in East, Southeast, and South Asia, and in West Africa (see map), as well as in the regional focuses of the institutes with rice genebanks (most samples in the genebank of AfricaRice are from West Africa).

An important question is whether or not there are still gaps in the results of this impressive collecting effort. That is hard to say because the data presented were taken from just a limited number of genebanks, not all, and some records did not have coordinates. Further analysis must also be done to incorporate genetic and other data to include the relative importance (contribution of additional diversity) of new sites. Nevertheless, some gaps stand out. In the collections mapped here, very few samples came from Japan and Korea (but these countries have genebanks of their own). Collections in parts of Myanmar and Cambodia are also sparse.

The process of georeferencing these key rice accession databases continues and we aim to produce a complete picture of global rice diversity soon, as it will certainly play a crucial role in the conservation and management of rice genetic resources for future generations.

Ms. Paule, Ms. Cuerdo, and Ms. Reyes are researchers; Mr. Rala is an associate scientist; and Dr. Nelson is a geographer in IRRI's Geographic Information Systems, Social Sciences Division. Dr. van Etten is an assistant professor, School of Biology, IE University in Spain, and Dr. Hijmans is an assistant professor, Department of Environmental Science and Policy, University of California-Davis.

[1] According to "The State of the World's Plant Genetic Resources for Food and Agriculture," FAO, 2009. Available at www.fao.org/ag/AGP/agps/PGRFA/wrlmap_e.htm.

[2] Guarlnick et al. 2006. PLoS Biology. www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0040381; http://bg.berkeley.edu/latest/.

[3] This work was supported by the GPG2 project (World Bank/Bioversity) and by the United States Department of Agriculture.