Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS

R.J. Hijmans[1]✉, L. Guarino[2], M. Cruz[1] and E. Rojas[1]

[1] International Potato Center, Lima, Peru. Email: r.hijmans@cgiar.org
[2] International Plant Genetic Resources Institute, Americas Office, Cali, Colombia

Summary

Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS

The DIVA-GIS software allows analysis of genebank and herbarium databases to elucidate genetic, ecological and geographic patterns in the distribution of crops and wild species. It useful for scientists who cannot afford generic commercial GIS software, or do not have the time to learn how to use it, and for others who require a GIS that is specifically designed for genetic resources work. Coordinate data are often absent from genebank databases, or if present are sometimes inaccurate. DIVA-GIS helps improve data quality by assigning coordinates, using a large digital gazetteer. DIVA-GIS can also be used to check existing coordinates using overlays of the collection-site and administrative boundary databases. Maps can then be made of the collection sites. Analytical functions implemented in DIVA include mapping of richness and diversity, distribution of useful traits and location of areas with complementary diversity. DIVA can also extract climate data for all terrestrial locations, which can be used to describe the environment of collection sites.

Key words: Diversity, documentation, GIS, geographic distribution, spatial analysis

Résumé

Outils informatiques pour l’analyse des données de ressources génétiques vegetales: 1. DIVA-GIS

Le logiciel DIVA-GIS permet l’analyse de bases de données des banques de gènes et des herbiers pour étudier les patrons génétiques, écologiques et géographiques dans la distribution des cultures et espèces sauvages. Cet outil est dirigé a des chercheurs qui ne peuvent faire l’acquisition d’un Système de Information Géographique (SIG) générique commercial ou qui n’ont pas le temps d’apprendre á utiliser ces logiciels, et pour ceux qui souhaitent disposer d’un SIG adapté á la recherche de ressources génétiques. Les coordonnées géographiques des bases de données des banques de gènes sont souvent absentes, et parfois imprécises. DIVA-GIS aide à améliorer la qualité des données par l’assignation avec un dictionnaire géographique digital. On peut aussi utiliser DIVA-GIS pour vérifier les coordonnées avec des juxtapositions des lieux de collecte avec des cartes digitales des limites administratives. Après cette correction de données on peut réaliser des cartes des localités de collection et analyser ces données. Les fonctions analytiques de DIVA contiennent des cartes de numéro d’espèces et d’autres indices de diversité, de la distribution de caractères importants, et des zones avec diversité complémentaire. DIVA peut aussi extraire des données climatiques pour n’importe quel endroit sur terre; ces données peuvent être utilisées pour décrire l’environnement des sites de collection.

Resumen

Sistemas de Información Geográficas para investigación en Recursos Fitogenéticos: 1. DIVA-GIS

El programa DIVA-GIS apoya al análisis de bases de datos de bancos de genes y de herbario para hallar patrones genéticos, ecológicos y geográficos en la distribución de especies silvestres y cultivadas. Está dirigido a científicos que no tengan recursos para un Sistema de Información Geográfica (SIG) comercial genérico o no tengan tiempo para aprender a usar éstos y para todos los que quieren un SIG hecho específicamente para trabajar en recursos genéticos. Para muchas accesiones de bancos de genes, faltan las coordenadas geográficas, y a veces son inexactos. DIVA ayuda a mejorar la calidad de los datos por la asignación automática de coordenadas, usando un diccionario geográfico digital. DIVA también puede ser utilizado para verificar coordenadas existentes haciendo sobreposiciones de sitios de colecta y bases de datos de límites administrativos. Después de depurar los datos, se pueden hacer mapas de los sitios donde las accesiones fueron colectadas y analizar los mismos. Las funciones de análisis implementadas en DIVA incluyen el mapeo de número de especies y otros índices de diversidad; de la distribución de caracteres útiles; y de áreas con diversidad complementaria. DIVA también puede extraer datos del clima para cualquier localidad en la tierra; estos datos pueden ser usados para describir el medio ambiente de los sitios de colección.

Introduction

N.I. Vavilov developed his theory of centers of diversity and origin of crops on the basis of numerous expeditions to collect crop genetic resources and subsequent evaluation and characterization work in the Soviet Union. Vavilov’s work is a classic example of the dual role of collecting expeditions: to make genetic variation available for use and also help elucidate genetic, ecological and geographic patterns in the distribution of species (Bennett 1970). Analysis of such eco-geographic patterns can make considerable contributions to several plant genetic resources research activities, including planning collecting programs, targeting genetic resources for breeding programs, developing core collections, selecting and designing sites for *in situ* conservation and assessing the potential impact of the products arising from the use of plant genetic resources. For detailed overviews, see Guarino *et al.* (1999, 2001).

Mapping and spatial analysis of genebank data can be carried out with off-the-shelf geographic information system (GIS) software. However, some of these software packages are too expensive for small programs or institutes and they do not provide specific options that enable rapid and uncomplicated analysis of biological diversity data. This paper is the first in a series describing GIS software specifically designed to be used for spatial analysis of data associated with genetic resources collections. The computer program described here, DIVA-GIS version 1.4, hereinafter called DIVA, was developed at the International Potato Center (CIP) in collaboration with the International Plant Genetic Resources Institute (IPGRI), and with support from the System-wide Genetic Resources Program (SGRP). It is available free of charge from the CIP website (http://gis.cip.cgiar.org). Base-map data (e.g. administrative boundaries, altitude) for use with DIVA are also provided for use with DIVA via the Internet. Other software that can be used specifically with genetic resources (or biodiversity) data include Floramap (Jones *et al.* 2001) and Worldmap (Williams 1994).

DIVA can import genebank databases containing passport, characterization and evaluation data, using the latitude and longitude fields. If latitude and longitude are not known, but locality information is available (such as department, province, and place name), DIVA can help in assigning the most likely coordinates. DIVA can also automatically check the accuracy of coordinate data.

When the data have been imported, completed, and checked for errors, DIVA can map the locations where genebank samples were collected. More interestingly, DIVA can also create analytical maps for use in developing plans and strategies for future collecting and *in situ* conservation activities. These include maps indicating the number of observations, the number of distinct classes of observations, and the value of diversity indices for an array of grid cells. DIVA can also provide estimates of the climate in the locations where germplasm was collected (or of any other terrestrial location).

The DIVA desktop

Most of the DIVA screen is made up of a map and its associated legend (Fig. 1). A map is drawn using geo-referenced databases called themes. For example, a map of the world can be made up with the following themes: altitude, national boundaries, main rivers and capital cities. Each theme on the map is also listed in the legend.

To manage the content of the map and they way themes are displayed, three menus are available: *File*, *Theme* and *View*. The *File* menu has functions for file and project management, exporting data and maps and printing. The *Theme* menu has functions for inspecting and managing individual themes. The *View* menu has functions that allow managing the map (e.g. zooming in and out). These menus are described in more detail in the DIVA manual (Hijmans *et al.* 2001).

There are three additional menus: *Analysis*, *Tools* and *Help*. The functions in the *Analysis* and *Tools* menus are discussed in more detail below.

Using DIVA: tools

Coordinate data in genebank databases are frequently scarce and occasionally inaccurate. This seriously complicates spatial analysis of genebank data and makes the results unreliable. However, there is much that can be done to improve the quality of the data and DIVA can make this task easier. DIVA can assign coordinates to accessions that have a locality description but no coordinates, and can help verify the accuracy of accessions that do have coordinates. These processes are described below.

Coordinate notation

The best system for the digital notation of geographic coordinates is decimal degrees. The commonly used sexagemal system contains numbers, symbols and letters (e.g., 12°34’12’’S) and needs to be stored as text (which is prone to error), or as six separate numerical variables. The decimal system only has a number, and no letters, with the sign indicating the hemisphere (+ = N or W, – = S or E) (e.g. –12.5700), hence only two variables are needed. Decimal degrees should be stored with 4 or 5 decimals. At the equator, one unit of the fourth decimal (0.0001 degrees) equals about 10 metres (less at other latitudes). That should be sufficiently precise unless a differential GPS (Global Positioning System) is used during the fieldwork and precision at the meter level is available. In those cases five digits would be better. To allow the user to assess the accuracy of the coordinate data it would be good practice to document how these were obtained (e.g. whether with a GPS or read from maps).

Assign coordinates

Coordinate data are often absent from genebank databases, particularly in older collections. For example, only 9% the accessions of six major genebanks of the US Department of Agriculture have coordinates (Steiner and Greene 1996). However, 50% of the accessions have a locality description. That means that there is scope for assigning coordinates to at least another 41% of the accessions (most accessions with coordinate data will also have a locality description). This can be done by searching for the locality names on maps or in gazetteers. A gazetteer is a list of names of geographic features and the coordinates of their locations. Fortunately, digital gazetteers are available.

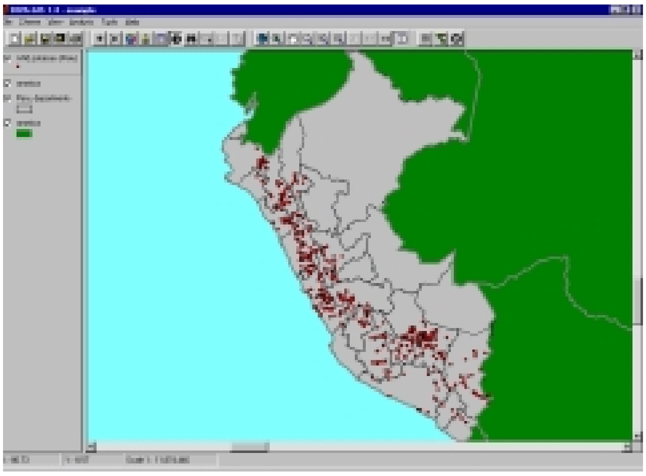


Fig. 1. The DIVA main window with a map of four themes.

DIVA can search for the coordinates of localities. The user specifies an input file that should ideally have the following fields indicating where the accession was found: country, first and second administrative subdivision and up to two locality names. For both locality names the distance and direction from the collecting site to that locality can be indicated. A digital gazetteer (the database of foreign geographic feature names from the US National Imagery and Mapping Agency[1]) is then used to search for the locality and assign its coordinates to the accession.

Create shapefile

Having assigned coordinates to all records, the data can be

[1] http://164.214.2.59/gns/html/index.html

imported to create a 'shapefile'—a file format that is used to show data on a map. The user has only to specify a DBF file and the two numerical fields in the file that contain the latitude and longitude data. A DBF file is a database format that can be made by many programs including Microsoft Excel® and Access®.

The advantage of using the shapefile format is that it is a commonly used format. It is the native format of the popular GIS software ArcView, and most other GIS programs can import shapefiles or export their data towards this format.

Check coordinates

The first time a shapefile is made from a genebank database it is likely to have some gross errors, such as points falling in the sea or on the wrong continent. Such unlikely locations are easy to identify, and often also easy to correct. However, there may also be other errors than cannot be so easily recognized.

DIVA's *Check Coordinates* tool helps to identify such errors, using a method described by Hijmans *et al.* (1999). By simultaneously querying the accessions database and an administrative boundaries database, a new (temporary) database is created. For each accession, this new database contains the location names according to the genebank database and according to the administrative boundaries database. These names should be the same, and any mismatches probably reflect errors (or changes in names or boundaries).

Extract

The *Extract* tool assigns environmental data to points. This allows for so-called retro-classification: environmental characterization of collecting site after, rather than during, collecting (Greene *et al.* 1999). This can be useful because traits can be related to ecological conditions at the places where the collections were made. Currently, monthly mean data for minimum and maximum temperature and precipitation are included. The data are extracted from a global interpolated climate database at a 10-minute resolution.

Using DIVA: Data analysis with grids

Once the accession database has been mapped, DIVA can carry out various analyses. Most of these analyses are based on grids. A grid divides the world into equal-sized cells, the size of which can be changed by the user. A calculation is then performed on each of the cells. For instance, the number of observations (points) in each cell can be calculated. The advantage of using grids rather than areas such as countries or administrative regions is that equal-sized grid cells can be compared more objectively. DIVA's analytical facilities are described below.

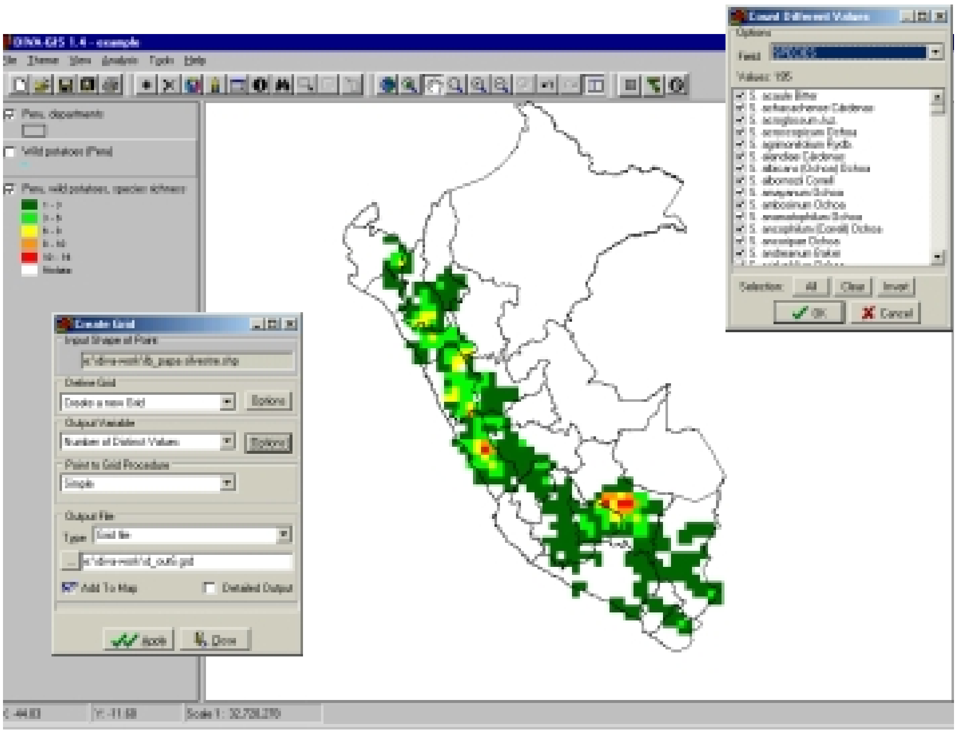
Number of Observations

The *Number of Observations* function allows the user to determine the number of observations in each grid cell. Points in the shapefile that are not relevant for a given analysis may be excluded by (de)selecting them on the basis of their value for a specific field in the database that describes the points. The number of observations in each grid cell can be determined by three methods: the 'Simple', 'Inverse Distance-Weighted' and 'Circular Neighborhood' methods.

In the 'Simple' method, points are simply assigned to the grid cell they fall in. Shortcomings of this method are that: a point that is on a border between grid cells is arbitrarily assigned to one grid cell; the value of a point that falls within a grid cell is assigned to that grid cell only, irrespective of the nearness of the point to other grid cells; the results are sensitive to the arbitrary origin of the grid; and, uncertainty about the location of the point is not taken into account. These shortcomings can largely be overcome with the circular neighbourhood and inverse distance-weighted (IDW) techniques, as implemented in DIVA. When the circular neighborhood option is chosen, calculations are made for a circle with its center in the middle of a grid cell and a specified radius. In the IDW method, inverse distance-weighted values are assigned to the four nearest grid cells. For more details see Hijmans *et al.* (2001).

Number of Different Values

The *Number of Different Values* function calculates the number of distinct classes of a certain variable that occurs in each grid cell. For example, if the input database consists of the locations where different wild species were observed, the database field that indicates the species names can be selected and the number of different species per grid cell will be produced. Figure 2 shows an example of such an analysis.



Create Grid

Input Shape of Points: s:\diva-work\rb_papa silvestre.shp

Define Grid: Create a new Grid [Options]

Output Variable: Number of Distinct Values [Options]

Point to Grid Procedure: Simple

Output File — Type: Grid file; s:\diva-work\rl_out6.grd

☑ Add To Map ☐ Detailed Output

[Apply] [Done]

Count Different Values

Options — Field: SPECIES

Values: 195

- ✓ S. acaule Bitter
- ✓ S. achacachense Cárdenas
- ✓ S. acroglossum Juz.
- ✓ S. acroscopicum Ochoa
- ✓ S. agrimonifolium Rydb.
- ✓ S. alandiae Cárdenas
- ✓ S. albicans (Ochoa) Ochoa
- ✓ S. albornozii Correll
- ✓ S. amayanum Ochoa
- ✓ S. ambosinum Ochoa
- ✓ S. anamatophilum Ochoa
- ✓ S. ancophilum (Correll) Ochoa
- ✓ S. ancoripae Ochoa
- ✓ S. andreanum Baker

Selection: [All] [Clear] [Invert]

[OK] [Cancel]

Fig. 2. The Create Grid and Output Options windows, together with a main map window showing wild potato species richness in Peru.

Diversity Indices

A number of diversity indices can also be calculated for each grid cell. A variable (field) from the input database is selected (e.g., species) for which a diversity index is calculated. The formulas for all indices were taken from Magurran (1988), who provided a detailed description of their properties. The mathematical descriptions of the different diversity indices are given in Table 1.

Table 1. Diversity indices

Index	Formula
Margalef	$D_{Mg}=(S–1)/\ln(N)$
Menhinick	$D_{Mn}=S/\sqrt{N}$
Shannon	$H´=–\Sigma p_i \ln p_i$
Simpson	$D=\Sigma(n_i(n_i–1)/N/(N–1))$
Brillouin	$HB=(\ln N!–\Sigma \ln n_i!)/N$

Where S is the number of unique classes per cell; N is the number of observations per cell; n_i is the number of individuals in the ith class; and p_i is the proportional abundance of the i th class ($=n_i/N$).

Complementary Site Selection

The *Complementarity Site Selection* (abbreviated to *Complementarity*) procedure aims at identifying sets of grid cells that are complementary to each other, i.e. that capture a maximum amount of diversity in as few cells as possible. Instead of using richness, an adjustment can be made in which rare observations are given more weight.

The procedure is based on the algorithm described by Rebelo (1994). The discussion below covers species, but any multi-state variable could be used for the analysis. The procedure used is less straightforward than it might seem. Whereas the selection of the first cell is easy – the cell with highest species richness (or a random choice between ties if there are any) – the choice of the next cell(s) depends on the previously selected cells. This is because the species in the cell with the second highest number of species may also be present in the first cell. This is a non-linear optimization problem. Rebelo (1994) developed a procedure that calculates an approximate optimal solution, and this has been implemented in DIVA.

An iterative procedure is used. In each iteration the 'value' of each grid cell is calculated, based on the observations in that cell, and in relation to the observations in the cells already selected. If there are two or more cells with the same value, one is selected at random. Hence, this procedure can lead to slightly different results each time it is run.

Statistics

If a numerical variable is selected from the accession database, statistics can be calculated for that variable, for each grid cell. The statistics included are listed in Table 2.

Using DIVA: point-based data analysis

An alternative to the use of grids are 'point-based' approaches, such as used by the Spatial Intra-specific Diversity (SID) software described by Nelson *et al.* (1997) and now also implemented in DIVA. Diversity indices are calculated based on all observations lying within a user-defined circle about each point. The results are assigned to the location of the central observation and output to a database. The results can then be mapped again in DIVA.

With the *Distribution Statistics* function, statistics for each unique value (class) of a multi-state variable can be calculated, for example, for each species in a database of wild relatives in a given genepool. Currently there are two statistics (others are being added), number of observations and MaxD.

Freely available GIS databases

Country-level GIS databases can be downloaded from http://gis.cip.cgiar.org. These databases can be used together with the genetic resources data that are being mapped and analyzed. There are shapefiles with data on administrative boundaries, country boundaries and first and second level administrative subdivisions for most countries. For all countries there are grids available for altitude, land cover and population density. These databases are all taken from existing public domain databases. In most cases, however, the data in these databases are difficult to obtain, being available in huge (global) files stored in difficult-to-use formats and are therefore not available to the non-specialist. Data from other sources can also be used in DIVA.

Conclusions

DIVA is easy to use stand-alone software and is available cost-free. It is thus a good starting point for people who work on plant genetic resources but who do not have access to commercial GIS software. A particularly useful feature of the DIVA project is that we also provide many country-level GIS databases. The lack of access to base map data is often perceived as another constraint to GIS adoption.

However, people who do have access to commercial GIS software programs may still want to use DIVA as it has functions specific to plant genetic resources that are not available, or are difficult to carry out, in other programs. This is facilitated by the use of standard GIS data formats (the ESRI shapefile) in DIVA. There is also a function to 'export' the gridfiles to IDRISI format and in forthcoming releases more formats for import and export of data will be included.

DIVA is probably most useful for analysis of distribution data covering larger areas, such as can be typically obtained from genebanks. An example in which DIVA was used extensively is a study by Hijmans and Spooner (2001), who describe the geographic distribution of wild potato species in North, Central and South America.

Table 2. Statistics

Min	Minimum value
Max	Maximum value
Mean	Mean value
STD	Standard deviation
CV	Coefficient of variation
Range	Difference between Max and Min
Range/Mean	Range divided by the mean
Median	Median value
Mode	Mode value

The problems that may occur with the quality of the coordinate data of genebanks and how one can deal with these problems have been discussed. However, it is equally important that attention is given to the information quality of genebank databases. Genebank databases do not necessarily provide an unbiased sample of existing diversity due to the way collections are made (Hijmans *et al.*, 2000). The extent to which this influences the results is partly dependent on the size of the grid cells chosen. For example, an area (grid cell) with high species richness may be associated with low species richness due to a small number of observations in that area. However, this problem may diminish if the size of the grid cells is increased.

The next release of DIVA (Version 2), which is planned for late 2001, will have more analytical functionality, particularly for geographic analysis of molecular data, but also improved data handling functions.

Acknowledgements

We thank the Environmental Systems Research Institute, Inc. (ESRI), the International Potato Center (CIP), the International Plant Genetic Resources Institute (IPGRI) and the System-wide Plant Genetic Resources Program (SGRP) for support. Andrew Jarvis and Karen Williams provided constructive feedback during the development of DIVA-GIS.

References

Bennett, E. 1970. Tactics of plant exploration. Pp. 157-179 *in* Genetic resources in plants: Their exploration and conservation (O.H. Frankel and E. Bennett, eds.). F.A. Davis Company, Philadelphia.

Greene, S.L., T. Hart and A. Afonin. 1999. Using geographic information to acquire wild crop germplasm: II. Post collection analysis. Crop Sci. 39:843-849.

Guarino, L., N. Maxted and M. Sawkins. 1999. Analysis of georeferenced data and the conservation and use of plant genetic resources. Pp. 1-24*in* Linking genetic resources and geography: emerging strategies for conserving and using crop biodiversity (S.L. Greene and L. Guarino, eds.). American Society for Agronomy Special Publication 27. ASA, CSSA, and SSSA, Madison, Wisconsin.

Guarino, L., A. Jarvis, R.J. Hijmans and N. Maxted. 2001. Geographic information systems (GIS) and the conservation and use of plant genetic resources. Proceedings of the International Conference on Science and Technology for Managing Plant Genetic Diversity in the 21st Century, Kuala Lumpur, Malaysia (in press).

Hijmans, R.J. and D.M. Spooner. (2001).Geography of wild potato species. Am. J. Botany / 88:2101-21120.

Hijmans, R.J., M. Schreuder, J. De la Cruz and L. Guarino. 1999. Using GIS to check coordinates of genebank accessions. Genet. Resour. Crop Evol. 46:291-296.

Hijmans, R.J., K.A. Garrett, Z. Huamán, D.P. Zhang, M. Schreuder and M. Bonierbale. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. Conserv. Biol. 14(6):1755-1765.

Hijmans, R.J., M. Cruz, E. Rojas and L. Guarino. 2001. DIVA-GIS, Version 1.4. A geographic information system for the management and analysis of genetic resources data. Manual. International Potato Center, Lima, Peru.

Jones P.G., L. Guarino and A. Jarvis. 2001. Computer tools for spatial analysis of plant genetic resources data: 2. FloraMap. Plant Genet. Resour. Newsletter (in press).

Magurran, A.E. 1988. Ecological diversity and its measurement. Princeton University Press, USA.

Nelson, A., G. LeClerc and M. Grum. 1997. The development of an integrated Tcl / Tk and C interface to determine, visualise and interrogate infraspecific bio-diversity. Internal document. CIAT, Cali, Colombia.

Rebelo, A.G. 1994. Iterative selection procedures: centres of endemism and optimal placement of reserves. Strelitzia 1:231-257

Steiner, J.J. and S.L. Greene. 1996. Proposed ecological descriptors and their utility for plant germplasm collections. Crop Sci. 36:439-451.

Williams, P.H., 1994. WORLDMAP priority areas for biodiversity. Version 3. Privately distributed, London.