The **International Network for the Improvement of Banana and Plantain** (INIBAP) was created in 1985. Its headquarters are in Montpellier, France. Regional networks have already been established in West and East Africa, Latin America and the Caribbean and the Asia and Pacific region.

The objectives of INIBAP are :

- to initiate, encourage, support and coordinate research aimed at improving the productivity of banana and plantain;
- to strenghthen national and regional programs and facilitate the interchange of improved and disease-free genetic material through assisting in the establishment and analysis of regional and global trials of new and improved cultivars;
- to coordinate and support the collection and exchange of documentation and information related to these crops;
- to coordinate and support training for researchers and technicians from developing countries.

Parc Scientifique Agropolis-Montpellier
Bat. 7, Bd de la Lironde
34 980 Montferrier sur Lez FRANCE

Jarret, R.L. (Ed.). 1990. Identification of Genetic Diversity in the Genus *Musa* : Proceedings of an international workshop held at Los Baños, Philippines, 5-10 September 1988, 211p.

Laid out by IBPGR Publications, Rome, Italy.
Cover: BEPAC, Montpellier, France.

International Network for the Improvement of Banana and Plantain

inibap

Réseau International pour l'Amélioration de la Banane et de la Banane Plantain
Red Internacional para el Mejoramiento del Banano y el Plátano

Identification of genetic diversity in the genus *musa*

Proceedings of an international workshop held at Los Baños, Philippines 5-10 September 1988

Editor: R.L. JARRET

Co sponsors and collaborators:

Directorate General of Science, Research and Development (DG XII) of the European Economic Community.

International Board for Plant Genetic Resources (IBPGR)

Philippines Council for Agriculture, Forestry and Natural Resources, Research and Development (PCARRD)

MUSAID : A COMPUTERIZED DETERMINATION SYSTEM

X. Perrier and H. Tezenas du Montcel

"Research on bananas and plantains was being seriously handicapped by a confusing terminology and the absence of an appropriate methodology for the unequivocal identification of genetic diversity . . . ." This comment by De Langhe in his introduction to the workshop emphasizes the inadequacy of our current knowledge of *Musa* taxonomy as it affects genetic improvement programs. MUSAID attempts to resolve this problem by providing a microcomputerized system to assist the user who wishes to determine the identity or taxonomy of a particular clone. MUSAID, which is an interactive user-friendly programme, identifies an unknown type by comparing it with data on a variety of well-defined clones. A sequence of questions are asked and after each response the level of resemblance to entries in the comparison data base is determined and a score is assigned to the unknown type.

PREVIOUS WORK AND POSSIBLE APPROACHES

The most frequently used method for determining taxonomic relationships in the genus *Musa* is the scoring method based on the series of 15 morphological characters described by Simmonds and Shepherd (1955). Each character is assigned a score between one (typical *M. acuminata*) and five (typical *M. balbisiana*). This method is fairly effective with edible bananas but is unreliable with wild species.

Identification keys, which have been the most valuable tool available to botanists for decades, present some drawbacks: a) if an error occurs at some point, there is no result at all or a false determination, b) if some characters cannot be scored the determination cannot be completed, c) it is very difficult to add new types or additional characters to identify new types, and d) the creation of a key to a particular flora is often the work of a lifetime.

In recent years, computerized methods have facilitated the creation of identification keys. An excellent review was presented by Sneath and Sokal (1973). However, the drawbacks remain and, furthermore, since it is impossible to compute all possible keys, algorithms cannot be optimal. The selected key is therefore, on average, an effective key (very much so in some types and less so in others) but it may not be the most effective one.

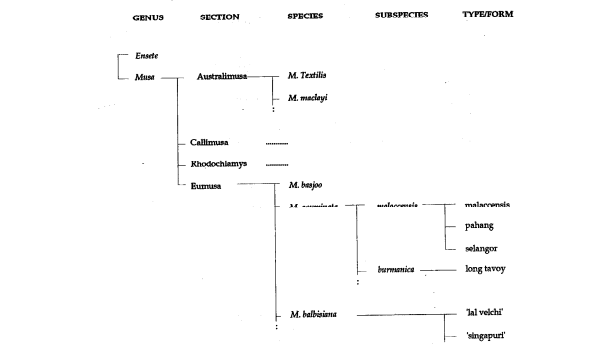
Improved data-processing methods and increased computer capacity has allowed new interactive approaches to be used. Expert systems are certainly the most attractive. These systems were developed for diagnosis of human diseases and the identification of toxic pests. They are very useful when knowledge is expressed in various forms that cannot take the form of an algorithm. This is not quite the case with problems of taxonomy since data can take the very simple form of a matrix crossing taxa and characters which would be very difficult to translate into expert system rules. Furthermore, expert system software is still in an experimental phase and efficient microcomputer software is not yet available. Another point that must be considered is the ease with which new types or new characters may be added. This is of great importance for a genus such as *Musa* since new information is being rapidly acquired.

Another method is a probabilistic approach which provides procedures to select the best descriptor in an interactive program. Willcox *et al.* (1973), Bascomb *et al.* (1973) and Lapage *et al.* (1973) give good descriptions and are useful references. These investigators attempted to identify bacteria using 56 different tests. The PANKEY package developed by Pankhurst (1978) is also an interactive program for the identification of an unknown type. MUSAID follows a similar pattern with some important differences. It offers the functions of an expert system combined with a probabilistic approach.

TAXA REFERENCE FILE

The taxa reference file is the core of MUSAID since all unknown types to be determined will be compared against the data in this file.

Figure 1
Example of a wild banana classification scheme



Taxonomists generally organize botanical families on several levels. Figure 1 is an example of a wild banana classification scheme on five levels: genus, section, species, subspecies and form. Cultivated bananas are usually organized on three levels; group, subgroup, and cultivar (Fig. 2). Two points must be emphasized: between 400 and 500 edible or wild types of *Musa* are known but this list is not yet complete; nor can the position of each type in the classification tree be considered definitive since two taxonomists might place the same types on different branches. Since MUSAID was developed to deal with both wild and cultivated bananas, the classification schemes in Figures 1 and 2 were mixed to give a tree as shown in Figure 3.

A prime consideration in the development of MUSAID was whether or not the cultivated banana groups are at levels similar to the wild banana species. It seems logical to assume that the wild types and their homologous cultivars (for example *M. acuminata* and members of the AA group) are at the same level. The most frequently described units, such as different *malaccensis*

Figure 2
Example of a cultivar classification scheme

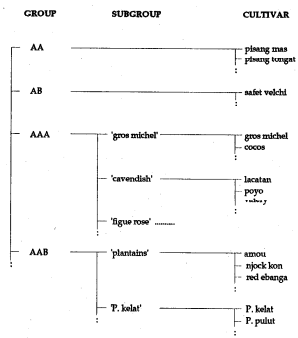
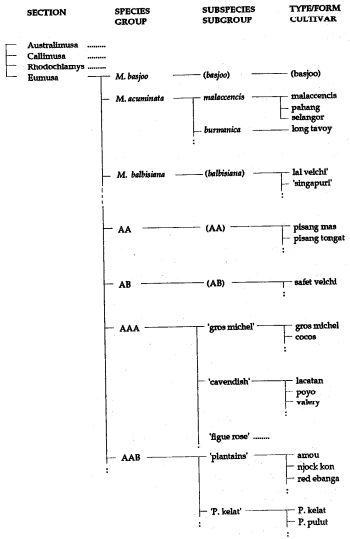


Figure 3
Example of a wild and cultivated banana classification used in MUSAID



forms and different Cavendish cultivars, are at the same level. This classification scheme has four levels:

- level 1: type/form and cultivar,
- level 2: subspecies and subgroup,
- level 3: species and group,
- level 4: section.

If necessary, a fifth level (genus) may be added, which would allow the addition of *Ensete*.

This scheme does not have branches of equal size. *M. basjoo* is classified at level three. Cavendish cultivars are described at level one. Some artificial units were created (units between brackets in Fig. 3) such as *basjoo* at level one or two, or *balbisiana* at level two. Thus, *basjoo* at level one is the only form of the *basjoo* subspecies, which is the only subspecies of the species *M. basjoo*. These artificial units have no influence on determination but make it easier to manage the classification tree.

In this paper the word taxa is used to qualify the described units even if this is not its precise botanical meaning.

TAXONOMIC DESCRIPTORS

These are morphological descriptors with between two and nine states, frequently three or four. They should be as independent as possible of environmental influences such as altitude, weather and nutrition. Therefore, quantitative measures of size such as height and weight are not suitable descriptors unless they are divided into wide classes (dwarf, normal, giant) or combined with other descriptors such as male bract length-to-width ratio. Potential descriptors are numerous, and there has been a general tendency to define an optimal subset to allow for the correct identification of all types. This approach has been used more or less successfully.

It is very difficult to place a value on a descriptor. A descriptor can be of low interest for the majority of the taxa but may be very discriminant for others. It may influence the states of other descriptors while having little value on its own. Others again do not characterize a type but provide greater details on an upper level. The introduction of a new taxa may increase the value of some descriptors. Since it was difficult to select a convenient subset, another approach was chosen which lets the system define the best descriptor in each case from a wide list of possibilities, without prior assumptions. Another problem was defining the states of each descriptor. The states must be easy to identify, be exclusive (a taxon can have only one state), cover all possible variations and concern only one type of information (it is preferable to define two descriptors rather than to create equivocal states).

MUSAID currently works with a list of 120 descriptors established by M. Tezenas du Montcel, based on his own and colleagues' experience (Simmonds and Shepherd, 1955; Champion, 1961; De Langhe, 1961; Tezenas du Montcel *et al.*, 1983; IBPGR, 1984). The descriptors are separated into different sections, each covering a portion of the banana plant. 'Father and son' relationships were defined for some descriptors'. A 'son' descriptor can exist for a taxon only if a particular state of its 'father' descriptor is selected. For example, the shape of the male bud can be considered only if the male bud is present.

The first descriptors on the list are environmental parameters such as geographical area and characteristics of the site. They are not identification criteria, but define conditions during observation of the taxon under investigation and allow the probability of some types to be weighted.

ERROR PROBABILITIES ASSOCIATED WITH EACH DESCRIPTOR

One of the most important differences between MUSAID and other systems such as PANKEY is the probability matrix associated with each descriptor. When the user selects state 2 (medium pedicellate) of descriptor 99 (fruit base insertion) (Fig. 4), MUSAID considers that the true value of the unknown taxon may be state 2 (probability = 0.40), but it may also be state 1 (probability = 0.30) or state 3 (probability = 0.30). This flexibility is helpful in dealing with the ambiguities of descriptors. Some descriptors are very difficult to score accurately. For example, it may be difficult to distinguish between potentially white, yellow or cream flower parts. Furthermore, this color may be modified by external conditions (flower age, lighting, etc.). Even when the expression of descriptors is as independent of environmental conditions as possible, given taxon may not have exactly the same values for all descriptors in different locations.

Probability matrices are very difficult to define. Characters that are easy to evaluate and are without variation will have a very narrow probability distribution. In these instances the answer will usually be exact. Some descriptors, however, such as colors, may be difficult to evaluate, giving a

Figure 4
Three morphological descriptors - their states and the associated probability matrix (x 100)

40	PEDUNCLE HAIRINESS				
	coarsely hairy - long hairs	59	30	10	1
	coarsely hairy - short hairs	30	49	20	1
	finely hairy	10	20	65	5
	glabrous	1	1	5	93
46	MALE INFLORESCENCE AXIS				
	absent	99	1		
	present	1	99		
99	FRUIT BASE INSERTION				
	short pedicellate	50	30	20	
	medium pedicellate	30	40	30	
	long pedicellate	20	30	50	

wider probability distribution and an answer that should be judged cautiously. For example, the presence or absence of the male inflorescence axis (Fig. 4) is very easy to score and the probability of error is very small: 0.99 and 0.01 (the value 0.00 never appears in matrix,). On the other hand, fruit pedicel length is very difficult to score correctly and error probability is higher.

For the character 'peduncle hairiness' (fig. 4), glabrous can never be confused with coarsely hairy. However, on a coarsely hairy peduncle, hair length may be difficult to determine. Therefore the probability matrix takes these possibilities of error into account. As opposed to other systems where correct determination is impossible once an error has occurred, it is always possible to complete an identification with MUSAID. Probabilities must, however, be valued by a highly experienced taxonomist and will be valid when used within the geographical area where the taxonomist works. The values currently used were developed by Tezenas du Montcel and are valid for the Caribbean area. For a truly comprehensive system, it is essential that taxonomists from different geographic areas agree on their values.

TAXA SCORE COMPUTATION

The system calculates the likelihood of an unknown taxon being identical to one of the data-file taxa. It questions the user and assigns a score to each taxon prior to receiving an answer(prior probability). The answer to each question adds new information which prompts the system to compute a new score (posterior probability). Score computation is based on Baye's conditional probabilities which for a taxon (i) are:

$$Q_i = [P_i \times a(k, m_r, r)] \; / \; \sum_{i=1}^{n} [P_i \times a(k, m_r, r)]$$

where:

Q_i : posterior probability
P_i : prior probability
k : descriptor number
m_r : descriptor k value for taxon (i)
r : answered state
$a(k,m_r,r)$: value of line m_r and row r in the error probability matrix associated with descriptor k

At step 0, all taxa have the same 1/n score (n : number of taxa in data file) since there is no information and the taxa cannot be distinguished from one another.

Scores are calculated as probabilities. They are smaller than one and sum to one for all taxa. This means that it is not possible to find high scores for several taxa at the same time. If a given type has a score of 0.90, the sum of all the others will be 0.10. If the same taxon appears in the data file twice, with two different names, even a perfect determination will not assign a score greater than 0.5 for each. Score values must be judged by comparison with other scores. These probabilities are computed only for level one. Upper-level scores are estimated from the levels below: level 2 from level 1, level 3 from level 2, etc. This estimation is a function of the scores and the number of lower-level units. The purpose of upper-level scores is not to define the point of determination, which is done through taxa score investigation, but rather to estimate the likelihood that an incompletely identified taxon might belong to one group or another. The use of upper-level scores is only valuable during the determination process to examine the evolution of identification --specifically to obtain information on upper levels when an identification is unsuccessful.

MUSAID also has an advantage over other systems in that it can identify the species, subspecies etc. when a complete determination has failed as a result of errors in descriptor answers, insufficient information or because the unknown taxon is not in the data file.

DESCRIPTOR CHOICE AT EACH STEP

There are three means of selecting a descriptor at each step: user choice, forward linkage and backward lingage.

User Choice

This is the most basic function. The experienced user, such as a taxonomist who wishes to use his own method of determination, chooses the descriptor himself.

Forward Linkage

This approach is referred to as forward linkage by analogy with expert system terminology. The system looks for the best descriptor to perform the most rapid determination of an unknown taxon from all data-file taxa without prior assumption of adherence. The choice criterion is based on maximization of posterior probability variation.

All posterior taxa scores are computed for all possible answers of each descriptor. The best descriptor is the one which provides the greatest number of high scores and the greatest number of small scores. The computation is repeated for all states of all descriptors and is a rather long procedure, particularly on a microcomputer. Some calculation methods allow reasonable computation times to be achieved.

Descriptor choice depends on a) dispersion of descriptor probability error (precise, clearly distinguishable characters will be preferred) and, b) information already obtained by other descriptors. As scores are computed from prior probabilities, taxa already showing a high score are favored. The MUSAID search strategy is to try to confirm tendencies resulting from previous steps.

If the user knows the species (group) or subspecies (subgroup) of an

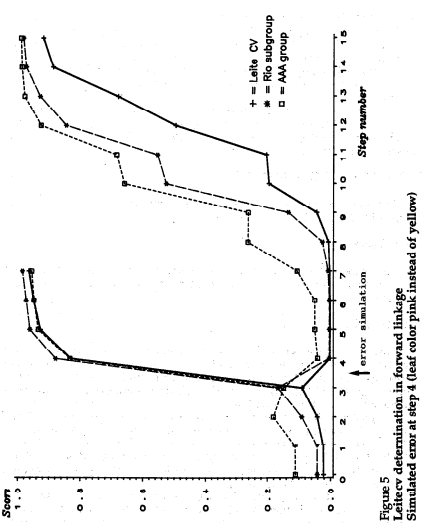


Figure 5
Leitecv determination in forward linkage
Simulated error at step 4 (leaf color pink instead of yellow)

unknown taxa, he can select a taxa subset and perform a determination only among the taxa of this subset. Figure 5 is an example of a forward linkage determination. The unknown taxa was cv. Leite, Rio subgroup, AAA group. After evaluating four descriptors, the 'Leite' score was 0.82. After the fifth descriptor the score increased to 0.90. At this step all other taxa had a very small score. The Rio subgroup and the AAA group also had high scores. The subsequent steps confirmed the determination.

To illustrate the effect of an error, a mistake is simulated at step 4. This is a very important step since a correct answer caused the score to increase from 0.10 to 0.82. In this example, the state yellow was substituted for the correct answer, pink. As a result, the 'Leite' score fell to near zero and the system deviated. After the addition of four new descriptors, the 'Leite' score began to increase and it reached 0.90 after the thirteenth descriptor. It should be emphasized that the Rio or the AAA level scores increased more rapidly. If the determination had been stopped at step 10, the 'Leite' identification would not have been complete (score close to 0.2). However it could have been determined that the unknown taxon belonged to the AAA group (score close to 0.70) and possibly to the Rio subgroup.

Backward Linkage

In backward linkage, it is assumed that the unknown taxon belongs to a definite and selected subpopulation of several taxa (a group or some subspecies, for example) or even to a definite taxon. The aim is to test that assumption and the program looks for a descriptor which gives the highest scores for the selected subpopulation, and the lowest scores for all others.

These three methods of selecting a descriptor can be mixed during a determination process. For example, the user may introduce some reliable characters directly and then select a subpopulation assumed to contain the taxon. This assumption may then be verified by backward linkage. If the assumption is verified, the determination is made by forward linkage on the selected subpopulation. For example, consider that the unknown type is cv. Williams belonging to the Cavendish subgroup and the AAA group (Fig. 6). The user, familiar with this group, expects the unknown taxon to belong to the Cavendish subgroup. Cavendish is selected and backward linkage is used. After four steps, Cavendish reaches a score greater than 0.95 and the assumption of belonging to the Cavendish subgroup is recognized. A choice must then be made between the different Cavendish taxa. Forward linkage is used, but only on this subgroup. The 'Williams' score quickly reaches 0.85. If forward linkage alone had been used, nine steps would have been necessary to reach the same probability level.

Another advantage of MUSAID is that the user may choose more than one state for a descriptor. This flexibility, which is very useful when the user is not sure of its true value or when no information is available on a character, is not found in other systems.

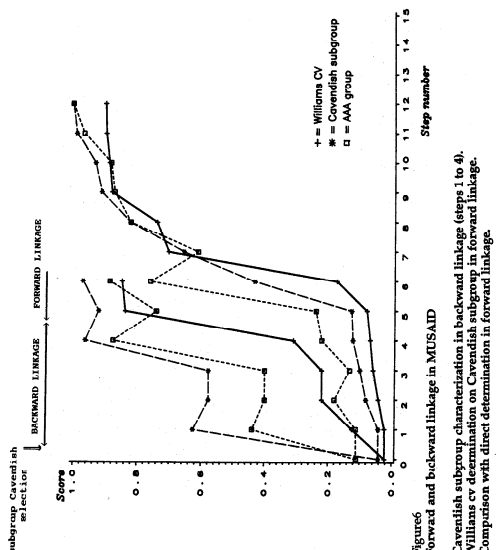
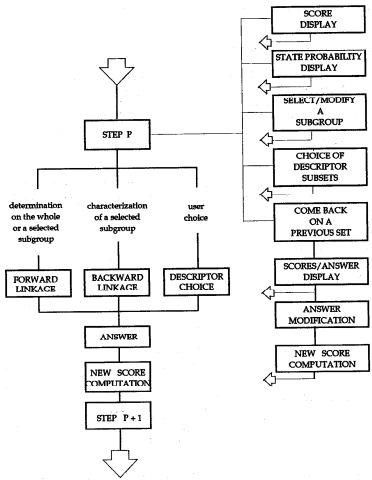


Figure6
Forward and backward linkage in MUSAID

**Cavendish subgroup characterization in backward linkage (steps 1 to 4).
Williams cv determination on Cavendish subgroup in forward linkage.
Comparison with direct determination in forward linkage.**

Figure 7
MUSAID program organization



STOPPING THE DETERMINATION PROCEDURE

There are no mathematical rules to determine when an identification has been achieved. Once a taxon has obtained a good score, a further three or four steps in backward linkage will confirm the identification. By definition, the

selected descriptors are the most discriminant and therefore most able to confirm the identification. It is necessary to specify that the conclusion has a probabilistic nature. It is not an assertion, but a likelihood estimation. When a good score is not achieved, the unknown taxon may not be in the reference file. However, upper-level scores may provide information on the species (group) or subspecies (subgroup). These conclusions may be verified by backward linkage.

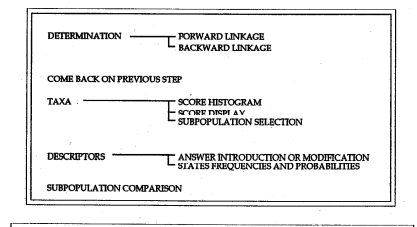
MUSAID PROGRAM ORGANIZATION

The organization of MUSAID is shown in Figure 7. In addition to the previously described procedure, MUSAID also provides different functions at each step. Figure 8 depicts an example of the main menu screen. The program returns to this menu after each step and each function.

Score Display

Scores of each level are displayed in decreasing order. At the lower level, a window gives details on complete identification name and step numbers of highest score increase and decrease. Another window displays a graph of the evolution of the score.

Figure 8
MUSAID main menu



State Probability Display

This is the frequency of each state for a given descriptor at step 0, when each taxon has the same probability. Later in the program this is a probability frequency which takes taxa probabilities into account. This data provides the most probable state of a descriptor, at a given step, and may help in detecting errors.

Selection or Modification of a Taxa Subpopulation

This is the definition of the taxa subpopulation on which the best descriptor choice will be performed. Nevertheless, the scores are always computed on all taxa.

Choice of Descriptor Subsets

This limits the descriptor list in some sections. It is useful during a determination when data on some descriptors cannot be obtained (e.g. ripe fruit descriptors during the flowering phase) or to define the most important descriptors of a given section.

Come Back on a Previous Step

It is possible to return to a chosen step and to display scores at this step. It is also possible to look at the descriptor, the response to that descriptor and to modify the answer. Afterwards all scores are computed again from this step. This may be useful to see the effects of answer modifications.

DATA-FILE MANAGEMENT

Banana taxonomy is a changing area and taxa lists are not yet definitive. Progress in research will result in modifications to the classification tree. Descriptor states or probability matrices may need to be changed. Other descriptor sections may need to be added.

MUSAID operates independently of its data files, which may therefore be modified without change to the program itself. A separate program called MUSDAT manages the data files. It enables the system to:

a) introduce a new taxon and record its values for each descriptor,
b) modify the name, classification position, etc. of descriptor values for taxon already recorded,
c) input a new descriptor and the record of its values for each taxon,
d) modify states, error probability matrices, and values of descriptors.

CONCLUSIONS

MUSAID is a step-by-step procedure for identifying an unknown taxon. Several of its features make it unique. MUSAID:

a) selects the best descriptor in each case according to the current context with no prior assumption about the importance of a descriptor,
b) takes account of observation errors,
c) takes a classification tree into account and provides information at each level,
d) allows data file modifications and addition of new descriptor types e.g. biochemical markers.

MUSAID is intended primarily as a determination tool for non-specialist users. It runs on desk-top and portable PCs under MS-DOS. It is very simple to use and is menu-driven. MUSAID allows the user to step back through procedures and retrace error points. It can provide multiple or null answers.

MUSAID software is still under development and improvements are still being made. The programme was written in a compiled basic, which is not a structured language and does not deal with large arrays. Translation into a more suitable language is foreseen. Some functions still have to be written in MUSAID and MUSDAT, particularly group comparison functions, which will select the best descriptors to discriminate between two chosen subpopulations. In addition, other taxa have yet to be recorded and descriptor error probabilities must be specified.

MUSAID was developed to complement the IRFA genetic improvement program. However, if banana taxonomists are interested in this program, it may be made more widely available through INIBAP. In order to enhance the effectiveness of MUSAID, taxonomists should work towards: a standardized name and numerical codification for each taxon, the compilation of a dictionary of synonyms, the definition of states for each descriptor and the addition of new descriptors.

REFERENCES

Bascomb, S., S.P. Lapage, S.P. Curtis and W.R. Willcox. 1973. Identification of bacteria by computer: Identification of reference strains. J. Gen. Microbiol. 77: 291-315.

Champion, J. 1967. Les bananiers et leur culture. 1. Botanique et Génétique, IFAC, Paris. pp. 212.

De Langhe, E. 1961. La taxonomie du bananier plantain en Afrique équatoriale. J. Agric. Trop. Bot. Appl. 8: 10-11.

IBPGR. 1984. Revised banana descriptors. IBPGR, Rome. pp. 54.

Lapage, S.P., S. Bascomb, W.R. Willcox and M.A. Curtis. 1973. Identification of bacteria by computer: General aspects and perspectives. J. Gen. Microbiol. 77: 273-290.

Pankhurst, R.J. 1978. Biological identification - The principles and practices of identification methods in biology, Edward Arnold, London.

Simmonds, N.W. and K. Shepherd. 1955. The taxonomy and origins of the cultivated bananas. J. Linn. Soc. Bot. (London) 55: 302-312.

Sneath, P.H.A. and R.R. Sokal. 1973. Numerical taxonomy: The principles and practice of numerical classification, W.H. Freeman, San Francisco.

Tezenas du Montcel, H., E. De Langhe and R. Swennen. 1983. Essai de classification des bananiers plantains (AAB). Fruits 38: 461-474.

Willcox, W.R., S.P. Lapage, S. Bascomb and S.P. Curtis. 1973. Identification of bacteria by computer: Theory and programming. J. Gen. Microbiol. 77: 317-330.