Genetic Resources and Crop Evolution **42:** 127–133, 1995.

Duplication within and between germplasm collections

I. Identifying duplication on the basis of passport data

Theo J.L. van Hintum[1] & Helmut Knüpffer[2]

[1] *Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Centre for Genetic Resources, The Netherlands, P.O. Box 16, 6700 AA Wageningen, The Netherlands;* [2] *Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Genebank, Corrensstraße 3, D-06466 Gatersleben, Germany*

Received 3 February 1994; accepted 18 April 1994

Key words: genetic resources management, duplication, databases

Summary

Principles of duplication within and between genebank collections have been explored, terminology is proposed and the difficulties in identifying probable duplication are discussed.

Identical duplication concerns genetically identical accessions, whereas common duplication refers to accessions derived from the same original population that are mixtures of lines with differing genotype frequencies, or random mating populations with the same alleles but differing allele frequencies. Partial and compound duplication are types of incomplete duplication. An additional type of duplication is the relation between the parents in a cross and their offspring, i.e. parental duplication.

Identifying probable duplication on the basis of passport data is often hindered by their incompleteness or poor quality. The genetic identity of accessions is also subject to changes during maintenance in genebanks. Therefore, probable duplicates will often not be true duplicates.

Examples from the European Barley Database illustrate the problems.

Introduction

Most collections in genebanks are still rather haphazard. There is an urgent need to rationalize such collections to improve the efficiency of plant genetic resources conservation. This is a rather difficult task and requires the development of appropriate methodology.

An obvious step in the rationalization process is identifying and minimizing unnecessary duplication within and between collections. Plucknett et al. (1987) estimated the number of crop accessions held in germplasm collections throughout the world to be over 2.5 million, including over 1.2 million accessions of cereals, 369,000 accessions of food legumes, 215,000 accessions of forage legumes and grasses, 137,000 accessions of vegetables and 74,000 clones of root crops. More recent data indicate that the total number of accessions has increased to over 3.8 million (FAO, 1993). For most of the crops, Lyman (1984) estimated that at least 50% of the combined germplasm collections in the world would consist of replicated accessions. Also the percentage of duplication has probably increased considerably since these estimates were made.

Evidently, not all duplication is waste of capacity. For safety reasons duplication of base collections is a standard procedure in most genebanks, and for practical reasons there is usually a high degree of duplication of useful material between active collections. However, duplication can generally be considered as a waste of capacity, especially in obvious cases such as duplication within a collection.

A consistent terminology of duplication has not yet been proposed. Do two samples have to be genetically identical to be duplicates, or is it sufficient if they have originated from the same population?

Many problems have to be faced while identifying probable duplicates, as observed and illustrated by Frese & van Hintum (1989) and Knüpffer (1988). In this paper the principles of duplication will be explored, terminology will be proposed and the problems in identifying probable duplicates will be discussed and illustrated with examples from the European Barley Database (Knüpffer, 1988, 1989).

Types of duplicates

In a narrow sense duplicates in germplasm collections can be defined as genetically identical accessions, i.e. identical duplication. If a large, well mixed seed lot is divided in two halves, these two halves will be identical duplicates. This narrow sense definition is of limited use. It can be applied only for:

1. original populations i.e. material that has not been rejuvenated,
2. material that is completely homogeneous, and
3. vegetatively propagated material.

If safety duplication is accomplished after each rejuvenation by taking a part of the rejuvenated seeds and send it to another genebank to be stored as safety duplicate, this can also be called identical duplication. Generally, apart from this type of safety duplication, identical duplication only occurs in genebanks in the case of duplication of homogeneous, homozygous lines.

In a broader sense, genebank duplicates can be defined as accessions derived from a common original population, having all alleles in common. These common duplicates can be mixtures of lines, with differing genotype frequencies, or random mating populations with the same alleles, but differing allele frequencies (Fig. 1). Common duplication as defined here, is the most frequently occurring type of duplication. Each rejuvenation of an accession of a cross pollinating species or a non-homogeneous accession of a self pollinating species will cause slight changes in the genetic identity of the accessions, i.e. allele frequencies will change. The rejuvenated accession will no longer be identical to the original accession or to accessions in other genebanks derived from the same original accession. Homogeneous accessions are rare in genebanks; even modern cultivars of a self pollinating crop can often be shown to contain "cryptic" variation when new e.g. biochemical traits are studied (e.g. Nielsen & Bay Johansen, 1986).

A different type of duplication is partial duplication, i.e. accessions derived from the same original population, having only a part of the alleles or genotypes in common (Fig. 1). A special type of partial duplication is compound duplication; an accession is duplicated as a compound of another accession if all its alleles are included in the other accession, i.e. if it is a selection of the other (Fig. 1). These cases often occur if landraces or populations change due to drift, selection or contamination, or if variable populations are split into morphologically distinct lines maintained as separate genebank accessions.

An additional type of duplication that should be distinguished is the relation between the parents in a cross and their offspring; each of the alleles in the offspring will be present in at least one of the parents. If both parents and offspring are part of a germplasm collection this can be considered a kind of duplication, that will be called parental duplication.

Problems in identifying probable duplicates

Identifying probable duplication has to rely on whatever information is available on the material. However, two samples with identical passport data obviously are not necessarily duplicates. This was clearly illustrated by Sahu (1989), who screened rice genebank accessions with identical cultivar names for disease resistances and found highly variable reactions. This is why duplicates identified on their passport data only, will be called probable duplicates.

The first step for the identification of probable duplicates has to be made on the basis of passport data. For this purpose reliable documentation is necessary, but not always available. Searching documentation requires appropriate interfaces, as will be shown later. Simple perfect matches between passport data of two accessions are rare, and cannot be relied upon.

Once probable duplicates are identified, it is not certain that they actually are duplicates. The genetic identity of a genebank accession does not always correspond to the passport information. The reliability of the passport data is sometimes low, and the genetic identity of the accessions sometimes changes in time.

Genetic identity

The problems concerning the genetic identity of genebank material are fundamental to genetic resources activities. If the genetic identity of the material no longer corresponds to the passport data, most of the genetic resources efforts become useless.

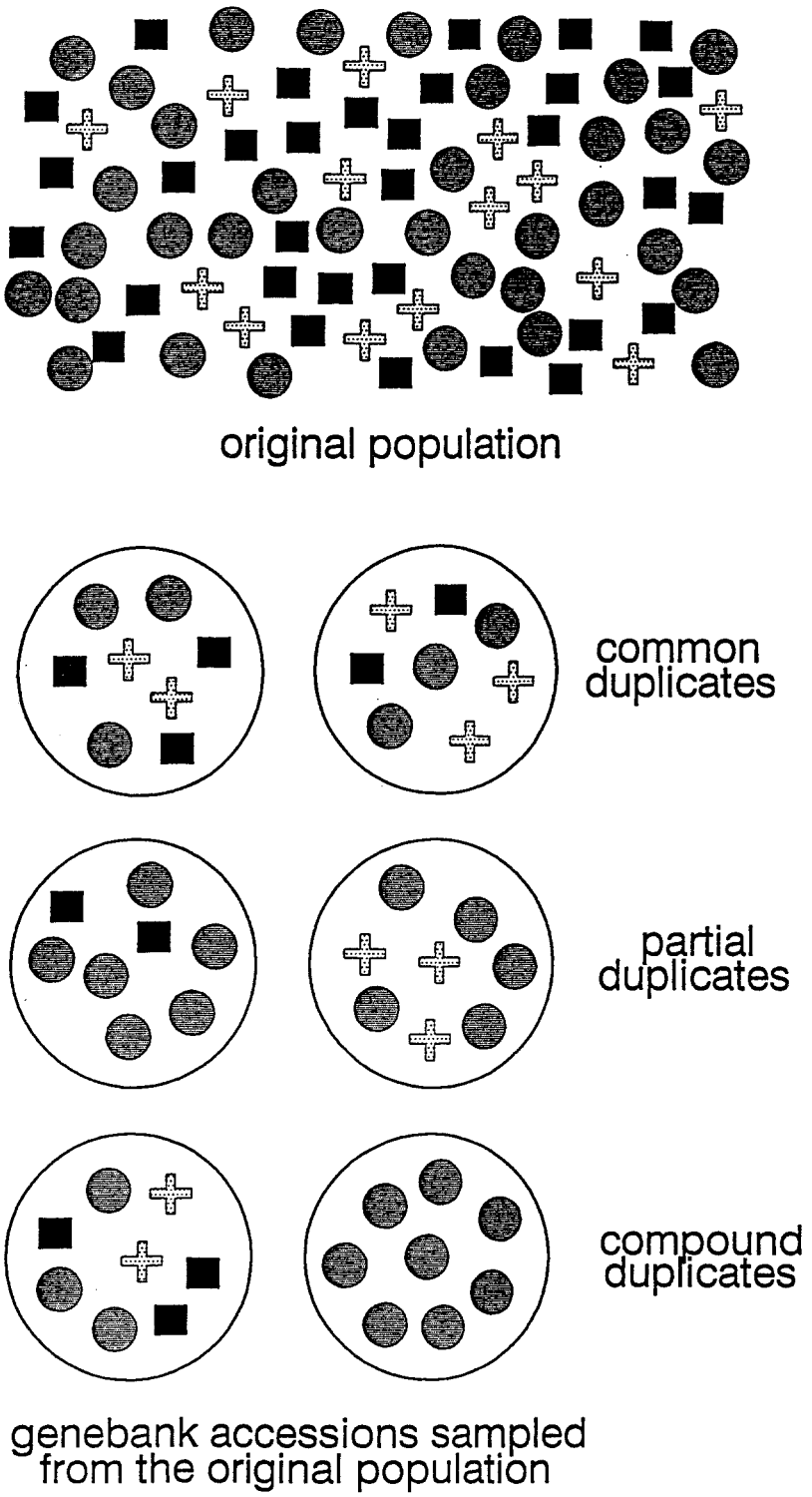

Fig. 1. Types of duplication.

The following causes of changes in the genetic identity of an accession can be distinguished:

- Intentional splitting of a sample into morphologically distinct parts. This is usually applied to prevent natural and unintentional selection during *ex situ* maintenance or to avoid problems in evaluation and documentation.
- Random genetic drift due to small population size during rejuvenation. Loss of alleles can be avoided by increasing the effective population size. This can be achieved by making sure each plant contributes more or less equally to the progeny, or by involving more plants in the rejuvenation. However, the latter is not always possible due to low germinability or a low number of available seeds.
- Natural selection during rejuvenation. This can be avoided by rejuvenating the accession in an environment similar to where it originated, and making sure each plant contributes more or less equally to the progeny by providing optimal conditions. The effects of the differences in fitness of the genotypes within an accession can never be completely avoided.
- Unintentional selection during rejuvenation and seed handling. Many of the mechanical processes during sowing, harvesting and handling of the seed cause selection towards uniformity of seed size, earliness and other characters.
- Contamination of seed lots during rejuvenation and seed handling. Contamination through seed or pollen has to be avoided by properly isolating cross pollinating crops, and generally by handling the material very carefully. Approaches like alternating plots of different self pollinating crops can be very helpful.
- Switching of seed lots during rejuvenation and seed handling.
- Mutation during storage. The rate of mutation is low in the case of seed propagated material, and will have little effect. In the *in vitro* propagated material, this problem can be much larger.

The first factor, the division of a sample in morphologically or taxonomically distinct parts, is an important factor, resulting in partial or compound duplication. It complicates the search for duplicates, but does not disturb the correspondence between sample and passport data. The next three factors, i.e. random genetic drift, natural and unintentional selection, all cause changes in allele frequencies. If the disappearance of alleles can be avoided, the material remains common duplicate, as defined previously. Contamination and switching of seed lots are more serious. Contamination reduces duplication to partial duplication, while switching eliminates identifiable duplication entirely.

Reducing the number of rejuvenations, and organizing the rejuvenations and seed handling properly, will reduce changes of genetic identity (cf. Breese, 1989). The material in most germplasm collections has a long history, and changes might unfortunately have occurred in the past.

Passport information

The other group of complicating factors in identifying duplicates is related to passport information. Errors and omissions do occur. Possible causes of problems are:

- Description of the wrong accession, i.e. lacking correspondence between data and accession.
- Omission of important information such as parents or collection number.
- Errors in interpretation, e.g. donor interpreted and stored as collector.
- Typing errors.
- Homonyms, i.e. the same name was given to different cultivars (cf. Arias et al., 1983).
- Synonyms, i.e. the same cultivar was registered under different names in different countries.
- Translation of cultivar names, or different/inconsistent transcription/transliteration rules applied for names in non-Latin alphabets.
- Difficulties due to taxonomic (re-)classification.

Most of these problems can be avoided by handling information carefully. The last two causes of problems, i.e. translation/transcription/transliteration and (re-)classification, depend of the system that is used and can partly be avoided by standardization. Here, as with the changes in the genetic identity of the accessions, many mistakes have occurred in the past.

While comparing collections, incompatibility of structure, format and coding of information is another complicating factor that can be overcome by standardization. The standard descriptor lists as they have been published for many crops by IBPGR can be helpful for this purpose (e.g. IBPGR, 1992).

Identifying probable duplicates in the European Barley Database

One of the main objectives of establishing an international database of plant genetic resources collec-

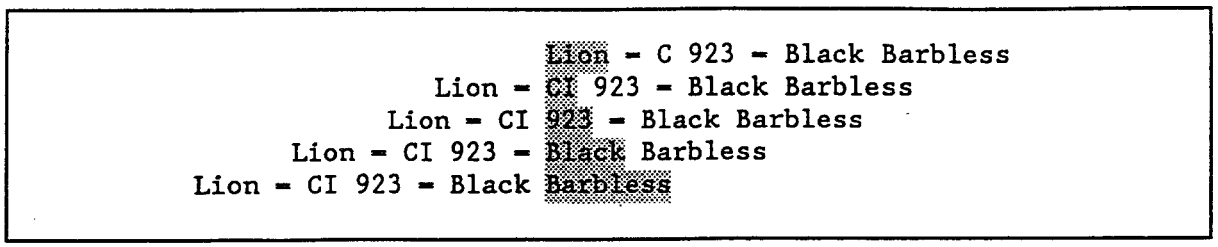

Fig. 2. List of KWIC index entries of the accession 'Lion = CI 923 = Black Barbless'.

tions of a certain crop is the identification of probable duplicates. A method commonly proposed for this purpose is the so-called Soundex procedure based on pair wise comparisons of the "phonetic similarity" of two accession names. This procedure is available in many database management systems. However, if this procedure is applied to a large set of names from several languages, the results will not always be reliable, since the same combination of letters may denote quite different sounds in different languages. It can therefore only be recommended as one method, to be supported by others.

The Soundex function of the database system FoxPro 2.0 would group the following names: 'Bavaria', 'Bibior', 'Bobro', and 'Buhobori'; another group would consist of 'Baeza', 'Bagemia', 'Bakemi', 'Beacon', 'BGM 1', etc., or 'Celinnyj', 'Challenge', 'Challenger', 'Chelmski 16', 'Chilenische Braugerste', 'Chlumecky', 'Clancy', and 'Clansman'. However, in the case of the following group, probable duplicates can be identified that would not be identified using only the KWIC index (see below): 'Closess IV', 'Colcess', 'Colcess IV', 'Colchicum', 'Colses', and 'Colsess'.

To assist in the identification of probable duplicates in the European Barley Database (EBDB) having data on more than 55,000 accessions, the KWIC (key word in context) index, commonly known from bibliographic databases, was used (Knüpffer, 1988, 1989).

A KWIC index provides a keyword surrounded by its context. In Figure 2 an example is presented showing the five index entries of the accession name 'Lion = CI 923 = Black Barbless'. This method makes it possible to detect accessions with matching or similar elements of information, even if these elements are not stored in the database in a unique way. For example, the cultivar 'Britta' may also appear as 'Weibulls Britta' or 'Britta Weibull'. Moreover, the same piece of information may be stored under different descriptors, e.g. a CI number may be found as accession number, accession name or its part, donor number, etc. Collection sites are often confused with accession names. Alphabetically sorted lists of particular descriptors, such as accession name, would reveal only a part of the probable duplicates.

Examples

Nearly 100 accession names containing the word 'Archer' are listed in Knüpffer (1988), among them 'Archer' (7 times), 'Abed Archer' (2), 'Golden Archer' (6), 'Spratt Archer' (9), etc., including also names consisting of several parts, such as 'Spratt Archer 14709' or 'Hansen 378 nutans Archer-type'. These additions to the name of this old British landrace indicate either the distributor of the seed, the type or number of the selection made from the landrace or even the parent with which 'Archer' was crossed, like in the case of 'Spratt Archer' which is a cultivar resulting from the cross between 'Spratt' and 'Irish Archer'.

Another example is based on the number '1102' that occurs as part of the descriptors accession name, collection number or donor number. One group of these accessions consists of names or numbers such as 'Ab. 1102' or 'Abyssinian 1102' and their variants, another group belongs to the collection number '1102' of the Balkan expedition in 1942. Two accessions with the name 'Abyssinian 1102 = L 94' suggest that accessions with the name 'L 94' belong to the same group of duplicates. The complete group of potential duplicates is shown in Table 1. Thus, the relatedness between the accessions 'Abyssinian 1102' and 'L 94' could be revealed only via a third accession linking both designations.

Using the number '1104' as keyword, a group of accession names was found that probably derived from the Estonian name 'Jõgeva 1104'. The Russian transcription 'Йыгева 1104' was obviously transcribed and/or misinterpreted later in very different ways, resulting in the following accession names: 'Iigeva

Table 1. Duplicate group 'Abyssinian 1102 = L 94' (see text).

Genebank: Acc. No.	Donor: Donor Number	Cnt	Accession Name
DDRGAT: HOR 2551	DDRHOHEIPZ: Ab 1102/47	ETH	
POLPOWSIN: 21540	DDRGAT: 2551/84	ETH	
NLDGBN: 1507		ETH	A. Hor. 2551
HUNRCA: 1241	NLD	ETH	ABESSINIA 1192 L 94
DEUBGRC: 10842	DEUBBABRAU: 1618	ETH	ABESSINISCHE 1102
GBRPBI: 8806	NLDSVP	ETH	Abyssinian 1102
FRAINRACLF: CFH 3013		ETH	ABYSSINIAN 1102 = L 94
GRCCERI: 7111		NLD	Abyssinian 1102 = L 94
POLIHAR: 2628	POLBAKOW: 12	ETH	ETHIOPIA AB. 1102
POLIHAR: 3161	DDRGAT: HOR 2551	ETH	ETHIOPIA AB. 1102/47
POLIHAR: 482	DDRGAT: HOR 3036	ETH	ETHIOPIA L 94
DDRGAT: HOR 3036	DEUBBABRAU: BBA 1465	ETH	L. 94

The first column gives the genebank holding the accession, e.g. DDRGAT is the former East German genebank in Gatersleben, and the corresponding accession number, in the case of the first line 'HOR 2551'. The second column gives the donor institute and the donor number. The last two columns give the country of origin (ETH is Ethiopia and NLD is The Netherlands) and accession name (after Knüpffer, 1988).

1104', 'Ingav 1104', 'Ingoc 1104', 'Iygeva 1104', 'Iygewa 1104', 'Jigeva 1104', Jygeva 1104', and 'Nytscheva 1104' (Knüpffer, 1988).

The problem of homonyms and synonyms is illustrated in Knüpffer (1989). Out of 25 accessions given in a table, 15 accessions with the name 'Askania' are listed, among them four with the synonym 'Belfor' indicated, and two with the synonym 'Dominator', the former being a Dutch spring barley, and the latter a German winter barley.

Donor numbers and numbers in other collections, i.e. the so-called 'parallel numbers' (Frese & Hintum, 1989; Hintum, 1989), can help in detecting spelling variants (errors). For example, the Canadian accession number 'CAN 1126' was found in six accessions in association with the following accession names: 'Galore', 'Calore', 'Gabore', 'Galover' and 'Vantage', the latter one probably due to an error in the number. The USDA number 'CI 1024' was associated with 'Quinn' (5 accessions) and 'Liunn' (1 accession), the latter probably being a misspelling (Knüpffer, 1989).

Other examples can also be found in the second paper of this study (Hintum & Visser, 1994).

Conclusions

In practical genebank work duplication within and between collections can take many forms. Only in special cases two accessions will be identical. More commonly the occurrence of alleles and their frequencies will differ in accessions derived from the same original accession. Duplication of the parents in a cross by their offspring should also be taken into account when discussing duplication.

Passport data can sometimes be used to identify probable duplication. Due to low reliability of passport data and low stability of the genetic identity of genebank accessions this can only be used as an indication of probable duplication.

Identification of probable duplication on the basis of passport data can only to a small extent be done automatically. Manual screening must be supported by appropriate interfaces, allowing multi field searches, KWIC indexing, searches on "phonetic similarity", etc.

Central databases compiled from several germplasm collections, such as the EBDB, are an indispensable tool in identifying probable duplication between collections.

References

Arias, G., L. Reiner, A. Penger & A. Mangstl, 1983. Directory of Barley Cultivars and Lines. E. Ulmer, Stuttgart.

Breese, E.L., 1989. Regeneration and multiplication of germplasm resources in seed genebanks: the scientific background. International Board for Plant Genetic Resources, Rome, Italy, 69p.

FAO, 1993. Data from the FAO world Information and Early Warning System on Plant Genetic Resources.

Frese, L. & Th. J. L. van Hintum, 1989. The International Data Base for Beta. In: International Crop Networks Series. 3. Report of an International Workshop on *Beta* Genetic Resources, IBPGR, Rome. pp. 17–45.

Hintum, Th. J. L. van, 1989. GENIS: A fourth generation information system for the database management of genebanks. Plant Genetic Resources Newsletter 75/76: 13–15.

Hintum, Th. J. L. van & D. L. Visser, 1994. Duplication within and between germplasm collections. II Duplication in four European barley collections. Genetic Resources and Crop Evolution 42: 135–145.

IBPGR, 1992. Descriptors for white clover (*Trifolium repens* L.). International Board for Plant Genetic Resources, Rome, Italy.

Knüpffer, H., 1988. The European Barley Database of the ECP/GR: an introduction. Kulturpflanze 36: 135–162.

Knüpffer, H., 1989. Identification of duplicates in the European Barley Database. In: Report of a Working Group on Barley (Third Meeting). IBPGR, Rome. pp. 22–43.

Lyman, J. M., 1984. Progress and planning for germplasm conservation of major food crops. FAO/IBPGR Plant Genetic Resources Newsletter 60: 3–21.

Nielsen, G. & H. Bay Johansen, 1986. Proposal for the identification of barley varieties based on the genotypes for 2 hordein and 39 isoenzyme loci of 47 reference varieties. Euphytica 35: 717–728.

Plucknett, D.L., N. J. H. Smith, J. T. Williams & N. Murthi Anishetty, 1987. Genebanks and the world's food. Princeton University Press, Princeton, New Jersey.

Sahu, R. K., 1989. Screening for duplicates in the germplasm collection. International Rice Research Newsletter 14: 4.