

J. Jansen · H. Verbakel · J. Peleman  
Th. J. L. van Hintum

## A note on the measurement of genetic diversity within genebank accessions of lettuce (*Lactuca sativa* L.) using AFLP markers

Received: 19 April 2005 / Accepted: 13 November 2005 / Published online: 3 December 2005  
© Springer-Verlag 2005

**Abstract** This paper discusses a statistical approach for measuring genetic diversity within genebank accessions of a self-fertilising species. This approach is applied to lettuce (*Lactuca sativa* L.), using AFLP marker data on a set of 1,390 accessions, representing six different lettuce types. Knowledge of the within-accession genetic diversity is important for decisions about the way accessions have to be maintained by genebanks. It is argued that if the within-accession diversity is small, as can be expected for a self-fertilising species like *L. sativa*, the best approach is to sample as many accessions as possible with only two plants per accession and estimate the within-accession diversity by the proportion of accessions of which the individuals are different.

### Introduction

Genebanks conserve the genetic diversity of crop species, which forms the raw material of plant breeding. If possible, genetic diversity is conserved in the form of accessions: batches of seed sampled from wild populations, traditional landraces, modern cultivars, genetic stock or other research material.

In genebank collections of self-fertilising species the genetic diversity is largely distributed between accessions. The exception may be landraces of a self-fertilising

species which may be composed of mixtures of pure lines. Between accessions, genetic diversity has been studied for many crop species. A survey of methods for the analysis of this form of genetic diversity is given by Mohammadi and Prasanna (2003). Depending on the species, considerable diversity can be found within accessions. If it concerns a cross-fertilising species propagated using open pollination, the within-accession diversity will generally be large. If it concerns a strictly self-fertilising species the within-accession diversity, especially of modern varieties, will be small or even absent. Within-accession diversity has great implications for both the conservation and the use of accessions.

Homogeneous accessions of self-fertilising species may be considered to consist of one genotype. All plants of the accession are equal and, in principle, one plant could suffice to regenerate an accession. However, in the genebank practice, always more plants are used, e.g. in order to guarantee enough seed. If within-accession diversity is present, always a greater number of plants are required to regenerate an accession since the variants occurring in the accession must be represented in the sample that is used for regeneration. In addition, if redundancy is to be reduced in a collection, within-accession diversity can cause many complications. For example, accessions will no longer be either completely identical or entirely different, but may partly overlap (van Treuren et al. 2004). It is generally assumed that within-accession diversity occurs mainly in cross-fertilising species and, to a lesser extent, in wild populations and landraces of a self-fertilising species.

When molecular markers came available it appeared that in modern varieties of a self-fertilising species, some residual diversity could be found. For example, van Treuren and van Hintum (2001) showed, using AFLP markers, that even accessions of modern barley varieties contain diversity, although this crop species is usually considered as completely self-fertilising. This diversity, less than 2% of the polymorphisms, can be expected to be either point mutations or remnants of the original diversity in the cross that yielded the variety.

Communicated by A. E. Melchinger

J. Jansen (✉)  
Biometris, P.O. Box 100, 6700 AC  
Wageningen, The Netherlands  
E-mail: johannes.jansen@wur.nl

H. Verbakel · J. Peleman  
KeyGene N.V., P.O. Box 216, 6700 AE  
Wageningen, The Netherlands

Th. J. L. van Hintum  
Centre for Genetic Resources, P.O. Box 16,  
6700 AA Wageningen,  
The Netherlands

The next logical question is how to determine the amount of diversity present within accessions. Morphological characterization can only show very few characteristics and variants, and often suffers from environmental influences. Therefore, biochemical or molecular markers are more appropriate. For genotyping plant material, various marker systems have been used: isozymes, RAPDs, RFLPs and currently also AFLPs and micro-satellites. The use of molecular markers for genebank management has been discussed in many papers such as Bretting and Widrechner (1995) and van Hintum and van Treuren (2002). Marker systems yield data sets with varying number of markers. In addition, the reliability and reproducibility varies between markers systems. After the choice of the marker system, the question remaining is how to sample accessions and how to quantify diversity within accessions.

In this paper, six crop types of *Lactuca sativa* L. (lettuce) will be investigated and compared with regard to genetic diversity within accessions. In the set of data that could be used for this study, each accession was represented by two individuals. The data used in this study were generated in a much larger project aimed at characterising the entire Centre for Genetic Resource (CGN)'s lettuce collection with molecular markers (van Hintum 2003). In this paper a measure of within-accession diversity is proposed, which is closely related to the gene diversity measure of Nei (1987). Statistical properties of these measures are investigated. Comparison of crop types requires determination of the inaccuracy of the estimate of within-accession diversity. The measure is used to compare six crop types of lettuce.

## Materials and methods

### Plant material

The plant material consisted of the complete CGN collection of *L. sativa* (as present at the time of sampling). Table 1 lists the numbers of accessions per lettuce crop type that were analysed in the current study. Two plants of each accession were analysed.

**Table 1** Numbers of accessions of different types of *L. sativa*

Type	Number of accessions
Butterhead (Bu)	716
Cos (Co)	213
Crisp (Cr)	220
Cutting (Cu)	151
Latin (La)	58
Stalk (St)	32
Total	1,390

### DNA extraction

All accessions were grown in the greenhouse, one accession per pot. When the plants had two mature leaves, 1.5 cm<sup>2</sup> of leaf tissue was sampled from the third immature leaf. Most accessions reached this stage 2 weeks after sowing. DNA was extracted from the sample leaf tissue. For DNA extraction, a protocol based on van der Beek et al. (1992) was used.

### AFLP analysis

The AFLP protocol followed the procedures described by Vos et al. (1995). An AFLP pre-screening was carried out in which 4 *EcoRI/MseI* +3/+3 and 20 *EcoRI/MseI* +3/+4 AFLP primer combinations were screened on a DNA pool containing 10 randomly chosen accessions. Based on the number of AFLP markers per lane and the distribution of the AFLP markers in a lane, the combinations of primer *EcoRI* + ACA with primers *MseI* + CTAT, *MseI* + CTGT and *MseI* + CTTG were selected. All individuals were genotyped on the basis of presence/absence of AFLP bands for these three primer combinations.

### Measuring within-accession diversity

#### Diversity measure

The diversity measure that will be considered is based on the probability  $p(a,m)$  that two random individuals from an accession  $a$  are different with regard to a marker  $m$ . This probability is directly related to the band frequency  $f(a,m)$  of marker  $m$  in accession  $a$ . The relationship takes the form:

$$p(a,m) = 1 - f(a,m)^2 - (1 - f(a,m))^2 \\ = 2f(a,m)(1 - f(a,m)). \quad (1)$$

According to Eq. 1, the probability  $p(a,m)$  as a function of the band frequency  $f(a,m)$  has a parabolic form. It is equal to zero if the band frequency  $f(a,m)$  is either zero or unity, and 0.5 if the band frequency  $f(a,m)$  is equal to 0.5. This formula is identical to the one given by Nei (1987) for heterozygosity. However, it should be noticed that in Nei's formula  $f$  refers to allele frequencies rather than band frequencies.

The band frequency may vary between accessions and, perhaps more importantly, between lettuce types. The average probability  $\bar{p}(m)$  that two random individuals from one accession are different with regard to marker  $m$  is given by

$$\bar{p}(m) = 2\bar{f}(m)(1 - \bar{f}(m)) - 2\text{var}_a(f(a,m)), \quad (2)$$

in which  $\bar{f}(m)$  is the average band frequency of marker  $m$  with regard to variation between accessions and  $\text{var}_a(f(a,m))$  the variance of the band frequency of

marker  $m$  with regard to variation between accessions. As a consequence, variation between accessions with regard to band frequency will deflate the average probability that two random individuals from one accession are different.

The type of statistical distribution, which can be used to describe variation in band frequency between accessions or markers in a general way, is the beta distribution (Mood and Graybill 1963). According to this distribution the between-accession variance of the band frequency of marker  $m$  may be written as

$$\text{var}_a(f(a, m)) = \phi_a(m)\bar{f}(m)(1 - \bar{f}(m)), \quad (3)$$

in which  $\phi_a(m)$  is a positive constant. The variance of the band frequency between accessions is proportional to the variance as expected from the average band frequency. As a consequence, by combining Eqs. 2 and 3, the average probability  $\bar{p}(m)$  that two random individuals from one accession are different with regard to marker  $m$  may be written as

$$\bar{p}(m) = 2(1 - \phi_a(m))\bar{f}(m)(1 - \bar{f}(m)). \quad (4)$$

The average probability that two random individuals from one accession are different as a function of the band frequency will follow a more or less quadratic shape. The parameter  $\phi$  is closely related to the correlation of genes within individuals or inbreeding coefficient as defined by Weir and Cockerham (1984). It should be noticed again that the parameter  $\phi$  refers to bands rather than alleles.

#### *Statistical analysis (1): estimation of $\phi_a(m)$*

For marker  $m$  the total variance of the marker data, i.e. the variance ignoring the influence of classifying factors such as accession, is equal to,

$$V_T(m) = \hat{f}(m)(1 - \hat{f}(m)),$$

in which  $\hat{f}(m)$  represents the observed band frequency of marker  $m$ . This variance can be split into two components, one component representing the variation between accessions  $V_B(m)$ , the other component representing the variation within accessions  $V_W(m)$ . The parameter  $\phi_a(m)$  can be estimated by

$$\hat{\phi}_a(m) = \frac{V_B(m)}{V_T(m)} = \frac{V_B(m)}{V_B(m) + V_W(m)}. \quad (5)$$

For each marker  $m$ , the variance components  $V_B(m)$  and  $V_W(m)$  can be estimated using residual maximum likelihood (REML) (Paterson and Thompson 1971). The calculations can be carried using the REML facilities of Genstat (Genstat Committee 2003).

It should be noticed from the estimate  $\hat{p}(m)$  of the average probability  $\bar{p}(m)$  that two random individuals from one accession are different with regard to marker  $m$  and may be written as twice the estimate of the within-

accession variance component  $\hat{V}_W$  for marker  $m$  with regard to variation with accessions, i.e.

$$\hat{p}(m) = 2\hat{V}_W.$$

#### *Statistical analysis (2): comparison of lettuce types*

In order to be able to interpret the difference between lettuce types, two different analyses were carried out. First, for each combination of lettuce type and AFLP marker the average band frequency was calculated. Using these average band frequencies a principal component analysis (Chatfield and Collins 1980) was carried out to show diversification between lettuce types with regard to band frequencies of AFLP markers.

Secondly, for each accession the number of AFLP markers for which the two plants of the accession were different was determined. At the same time, for each accession the number of AFLP markers for which the two plants had marker observations was determined. Using these numbers, a comparison of lettuce types with regard to average within-accession diversity was made using a generalised linear model for overdispersed binomial data (McCullagh and Nelder 1989; Williams 1982).

All calculations were carried out using Genstat (Genstat Committee 2003).

#### *Effects of errors in the data*

Estimates of diversity measures may be affected by errors in the data. Suppose,  $p_0$  denotes the true probability that two individuals within an accession are different with regard to a marker. Furthermore, suppose that errors occur at random at a rate equal to  $r$ . Then, the 'realised' probability  $p_r$  that individuals within an accession are different is equal to

$$\begin{aligned} p_r &= 2r(1 - r) + p_0(1 - 4r(1 - r)) \\ &= 2r(1 - r) + 2f(1 - f)(1 - 4r(1 - r)), \end{aligned}$$

in which  $f$  denotes the band frequency. The relative error expressed as a percentage,

$$100(p_r - p_0)/p_0,$$

can be used to gauge the effect of random errors on the gene diversity. It should be noticed that if  $f=0.5$ , random errors have no effect on the probability that two individuals within one accession are identical.

## **Results**

### **Band frequencies**

In total, 151 polymorphic AFLP markers were scored: 54, 32 and 65 for the three combinations of primer

*EcoRI* + ACA with primers *MseI* + CTAT, *MseI* + CTGT and *MseI* + CTTG, respectively. Figure 1a shows a histogram of the band frequencies of all 151 polymorphic AFLP markers. Most AFLP markers have a band frequency either close to zero or close to unity (39 AFLP markers have a band frequency smaller than 0.01 and 28 AFLP markers have a band frequency larger than 0.99). Figure 1b shows that variation in band frequency between accessions is much smaller compared to that between markers. This may be the result of the fact that the majority of AFLP markers have a band frequency close to zero or unity and, as a consequence, do not vary across accessions.

#### Diversity within accessions

For each combination of accession and AFLP marker it was determined whether the two individuals were different. For each AFLP marker the proportion of accessions, for which the two plants were different, was calculated. A histogram of these proportions is shown in Fig. 2a. For each accession the proportion of AFLP markers for which the two plants were different was calculated. A histogram of these proportions is shown in Fig. 2b. For 14 AFLP markers, plants were identical within accessions for all 1,390 accessions. For 938 accessions, no difference between the two plants was found with regard to all AFLP markers.

#### Relationship between within-accession diversity and band frequency

For each marker the value of the between-accession variance with regard to band frequency has been plotted against the total variance in Fig. 3a. This figure shows that the parameter  $\phi$ , i.e. the ratio of the between-accession variance and the total variance, is more or less constant across markers. The estimate  $\hat{\phi}$  of  $\phi$  [obtained by linear regression with a fixed intercept at (0,0)] is 0.955 (SE=0.0014).

For each marker the value of the proportion of accessions, for which the two plants are different, has been plotted against the band frequency in Fig. 3b. This figure shows that in general the larger proportions of accessions for which the two plants are different are found if the band frequency is intermediate. The quadratic shape of the relationship shown in Fig. 3b is in agreement with previously derived properties. The solid curve is a graphical representation of expression 4 using  $\hat{\phi} = 0.955$ .

#### Principal component analysis of lettuce types with regard to band frequency

A principal component analysis was carried out on the band frequencies obtained for combinations of *L. sativa* types and AFLP markers. Figure 4 shows the results for the first two axes. These account for 55.3 and 23.1% of the variance, respectively.

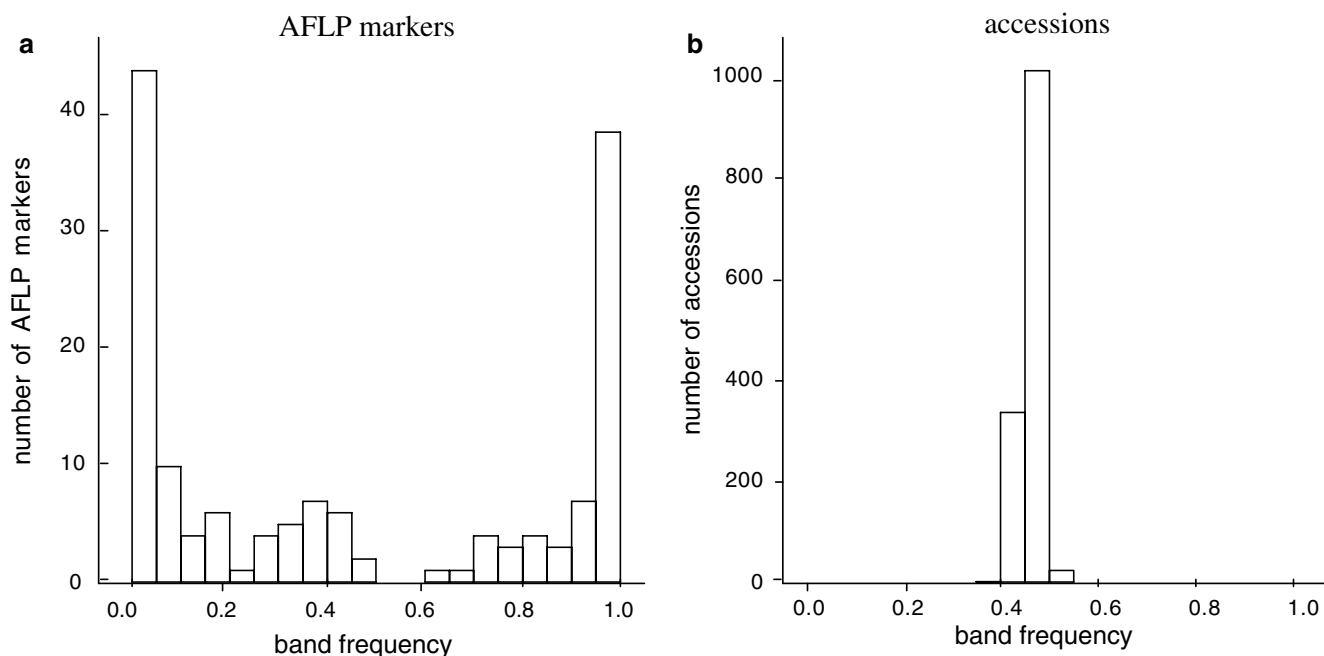
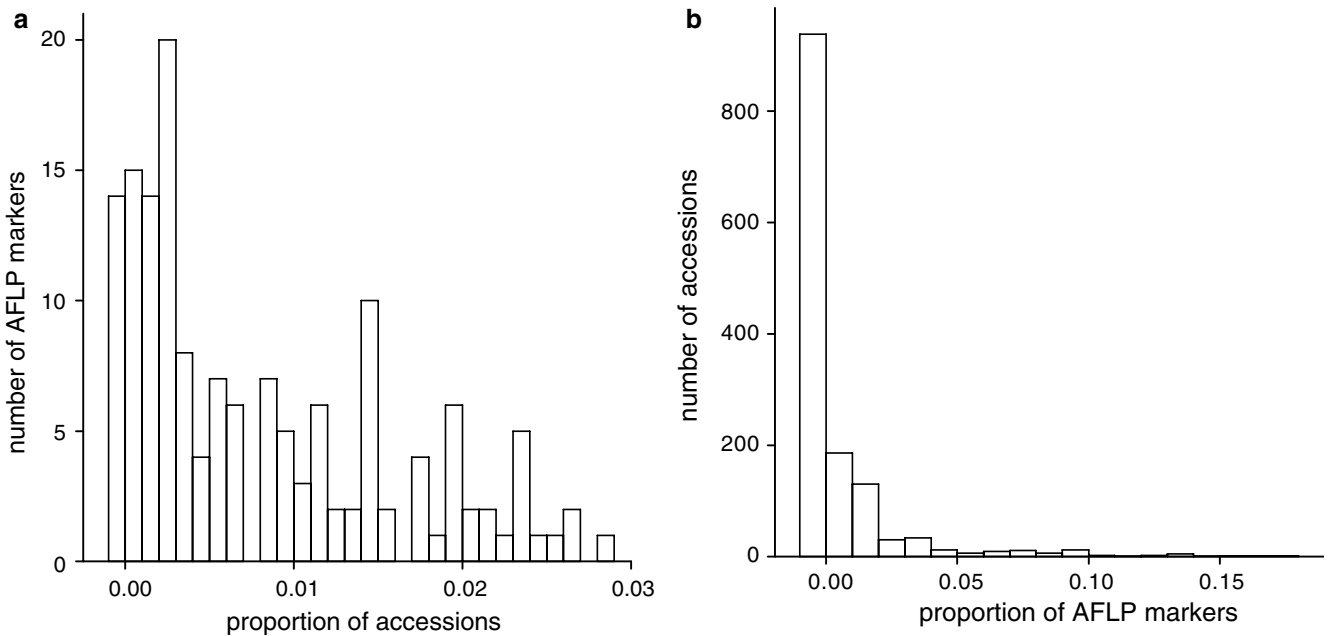
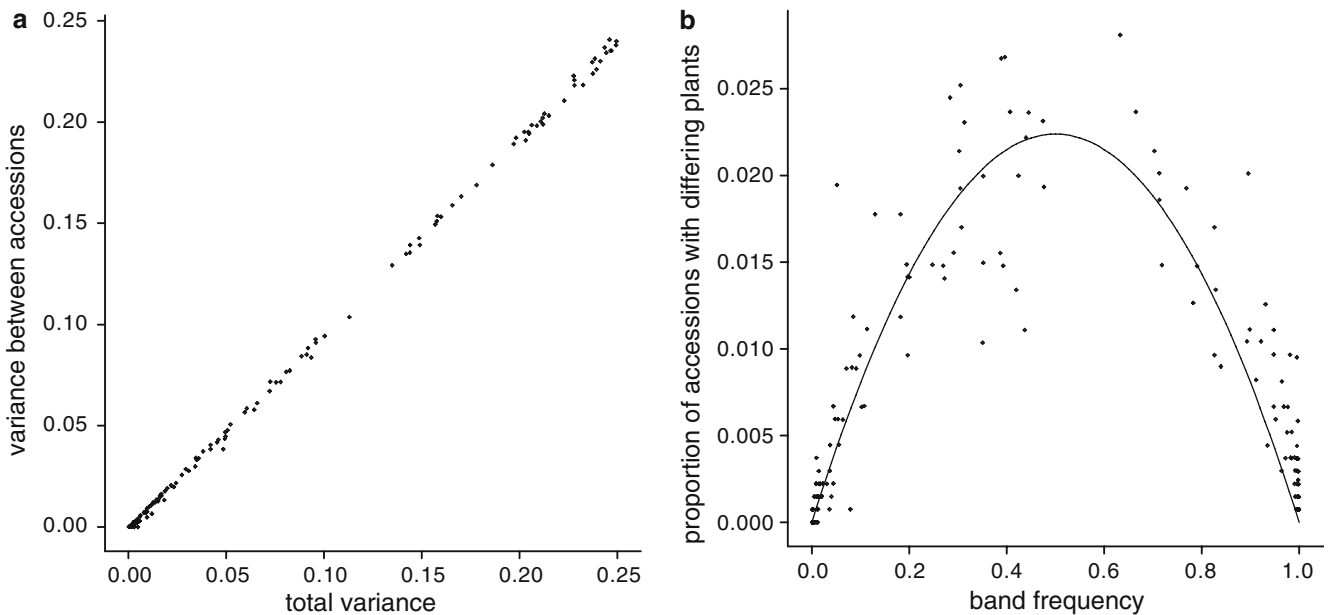


Fig. 1 Histogram of band frequencies: **a** AFLP markers, **b** accessions



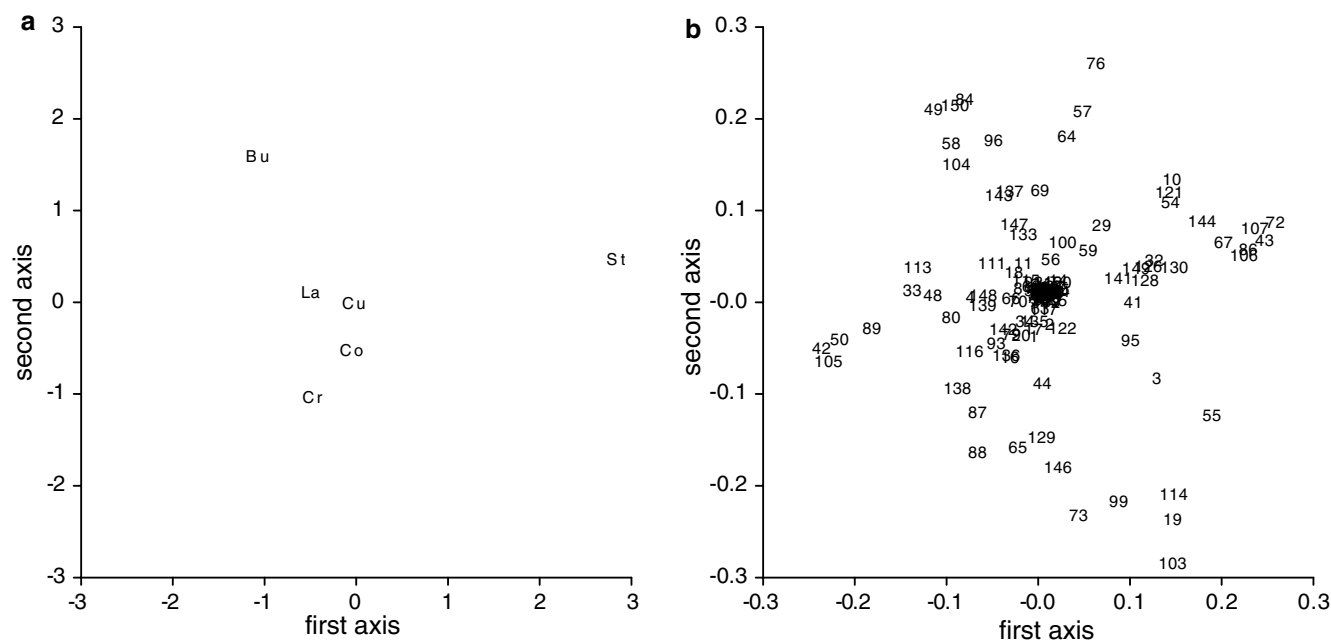
**Fig. 2** **a** Histogram of the proportion of accessions for which the two plants were different. **b** Histogram of the proportion of AFLP markers for which the two plants were different



**Fig. 3** **a** The variance in band frequency between accessions versus the total variance. **b** The proportion of accessions for which the two plants are different versus the band frequency

Figure 4a shows that the first axis mainly represents a difference between type Stalk lettuce and the others, while the second axis represents a difference between Butterhead lettuce and the other types. A major conclusion is that AFLP markers that have very different band frequencies in different lettuce types are present. As a consequence these markers will contribute differently in different lettuce types to the within-accession diversity measures.

Figure 4b shows the loadings of the AFLP markers for the first two principal axes. The loadings are the contributions of the individual AFLP markers to the first and second principal axes. Figure 4b shows that a large number of AFLP markers do not contribute to differences between the lettuce types; the loadings of these markers are found close to the origin (0,0). Some AFLP markers show large differences in band frequency between Stalk lettuce and the other lettuce types (e.g.



**Fig. 4** Results of a principal component analysis on average band frequencies: **a** scores of the lettuce types (indicated by *abbreviated names*), **b** loadings of the markers (indicated by *marker numbers*)

AFLP markers with numbers 72 and 42 and AFLP markers close to these markers). Other markers show large differences in band frequency between Butterhead lettuce and the other lettuce types (e.g. AFLP markers with numbers 49 and 103 and AFLP markers close to these markers).

#### Comparison of lettuce types

For each lettuce type the value of the average proportion of accessions with a difference within accession is shown in Table 2. Each value is accompanied by the corresponding standard error. The type Stalk lettuce has the highest within-accession diversity, whereas type Crisp lettuce has the smallest within-accession diversity.

**Table 2** Estimates of the average percentage of accessions with a difference within accession for the different lettuce types accompanied by standard errors

Type	Average percentage of accessions with a difference within accession	
	Estimate	Standard error
Butterhead	0.40	0.032
Cos	0.63	0.074
Crisp	0.33	0.053
Cutting	1.02	0.111
Latin	0.49	0.124
Stalk	1.36	0.277

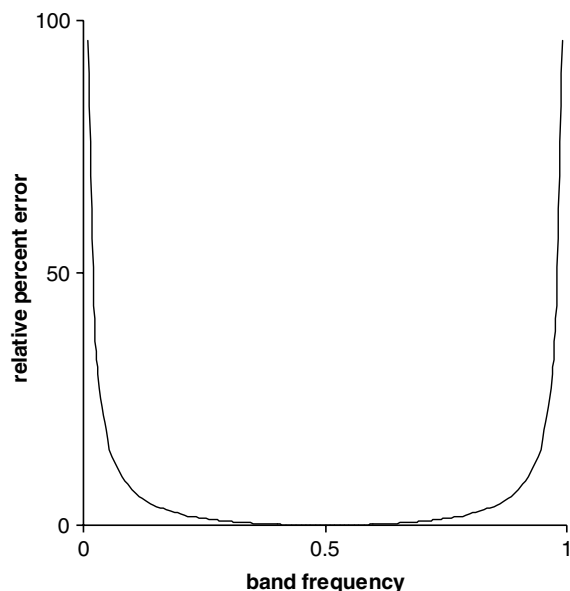
The effect of errors on within-accession diversity

Figure 5 shows the relative percent error as a function of the band frequency if the error rate is 1%. It shows that errors have far more effect if band frequencies are close to zero and unity. They have very little or no effect if the band frequency is close to 0.5.

#### Discussion

The measure of within-accession diversity considered in this paper is very closely related to the gene diversity measure of Nei (1980). Nei's measure refers to allele frequencies rather than band frequencies. Band frequencies could be transformed into allele frequencies by assuming the accessions are completely homozygous in the case of a self-fertilising species or in Hardy–Weinberg equilibrium in the case of an out-breeding species. However, this is unnecessary for the purpose of this analysis or, in general, for maintaining genebank material.

The probability that two individuals of one accession are different with respect to the set of AFLP markers obtained in this study is on average approximately 1%. Results (see Table 2) were obtained using a generalised linear model for overdispersed binomial data. The differences between the lettuce types can be traced back to the different origins of the accessions. The accessions of Crisp and Butterhead lettuce are mainly modern breeder's varieties; in order to obtain breeder's right protection the varieties have to be homogenous. The accessions of Stalk and Cutting lettuce are mainly landraces. They are the opposite of Crisp and Butterhead lettuce with



**Fig. 5** The relative percent error as a function of the band frequency if the error rate is 1%

regard to domestication and crop improvement (I.W. Boukema, personal communication).

A large proportion of accessions shows no within-accession diversity. In such a case it is not sensible trying to estimate the within-accession diversity for each accession separately with great accuracy. In most cases the within-accession diversity will be estimated as zero even if the number of individuals is increased. A sensible strategy would be to estimate the within-accession diversity by the proportion accessions of which the individuals show differences, on the basis of small number of individuals. A value of two would be a good suggestion.

Cross-fertilising species naturally exhibit a much larger within-accession diversity than found in this study. In such case, estimation of the within-accession diversity may require another approach with more than two plants per accession and based on the proportion of pairs of different individuals per accession. The proportion  $P$  of pairs of individuals that are different with regard to a marker is equal to

$$P = 2F(N - F)/N^2,$$

where  $F$  and  $N$  denote the number of individuals carrying a band and the total number of individuals in an accession, respectively.

In this study, the majority of AFLP markers have band frequencies close to zero or unity. An intrinsic property of these markers is that they contribute very little to the within-accession diversity. Moreover, they are more vulnerable to typing errors than markers with band frequencies around 0.5. The effect of errors can be reduced by selecting markers with a band frequency between 0.3 and 0.7 within the group of accessions considered.

The principal component analysis shows that lettuce types are different with regard to band frequency for a number of AFLP markers, i.e. the markers with high absolute loadings. Since within-accession diversity is directly linked to band frequency, this means that markers may contribute differently to the average within-accession diversity of different lettuce types. It should be noticed that usually markers with loadings close to zero are markers with average band frequencies close to zero or unity in the entire group of accessions.

The methods described in this paper can be applied using standard statistical software. However, general application of the methods would require further testing of the model on data from other self-fertilising species, e.g. barley.

The accessions of lettuce considered in this paper exhibit very little within-accession diversity. An important problem for genebank curators is whether they want to maintain these low levels of within-accession diversity in future generations. Regeneration of accessions would inevitably require large numbers of plants.

**Acknowledgements** This research was part of a project called 'Molecular markers of genebanks: application of marker technology for the improvement of ex situ germplasm conservation methodology', which was supported by a research grant from the European Union in Framework 4 Biotech program (Bio4 CT 9 PL96.2062). The authors are grateful to Ietje W. Boukema for information and discussions on the origin of the lettuce accessions.

## References

- van der Beek JG, Verkerk R, Zabel P, Lindhout P (1992) Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: Cf9 (resistance to *Cladosporium fulvum*) on chromosome 1. *Theor Appl Genet* 84:106–112
- Bretting PK, Widrechner MP (1995) Genetic markers and plant genetic resource management. *Plant Breed Rev* 31:11–86
- Chatfield C, Collins AJ (1980) Introduction to multivariate analysis. Chapman & Hall, London
- Genstat Committee (2003) Genstat® release 7.1 reference manual. VSN International, Hemel Hempstead
- van Hintum ThJL (2003) Molecular characterisation of a lettuce germplasm collection. Eucarpia leafy vegetables 2003. In: Proceedings of the eucarpia meeting on leafy vegetables genetics and breeding, Noordwijkerhout, the Netherlands, 19–21 March 2003. Centre for Genetic Resources, Wageningen, pp 99–104
- van Hintum ThJL, van Treuren R (2002) Molecular markers: tools to improve genebank efficiency. *Cell Mol Biol Lett* 7:737–744
- McCullagh P, Nelder JA (1989) Generalized linear models (2nd ed) Chapman and Hall, London
- Mohammadi SA, Prasanna BM (2003) Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop Sci* 43:1235–1248
- Mood AM, Graybill FA (1963) Introduction to the theory of statistics. McGraw-Hill, New York
- Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York
- Paterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58:545–554
- van Treuren R, van Hintum ThJL (2001) Identification of intra-accession diversity in selfing crops using AFLP markers: implications for collection management. *Genet Resour Crop Evol* 48:287–295

- van Treuren R, Magda A, Hoekstra R, van Hintum ThJL (2004) Genetic and economic aspects of marker-assisted elimination of redundancies from a wild potato germplasm collection. *Genet Resour Crop Evol* 51:277–290
- Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* 23:4407–4414
- Weir BS, Cockerham CC (1984) Estimating  $F$ -statistics for the analysis of population structure. *Evolution* 38:1358–1370
- Williams DA (1982) Extra-binomial variation in logistic linear models. *Appl Stat* 31:144–148