Molecular Ecology Resources (2008) **8**, 572–574 doi: 10.1111/j.1471-8286.2007.02011.x

COMPUTER PROGRAMS

COMBI.PL: a computer program to combine data sets with inconsistent microsatellite marker allele size information

HELGE TÄUBERT* and DANIEL G. BRADLEY†
*Insitute of Animal Breeding and Genetics, University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany,
†Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland

Abstract

Combining two data sets with allele information from overlapping microsatellite markers is often desirable, particularly in population genetic studies where a substantial body of published data exists. When genotyping is performed in different laboratories, allele size calling may not be presumed to be consistent. Our approach solves this problem by assigning allele sizes across studies using maximum-likelihood theory. Using data overlaps in samples and markers, allele shifts between two studies are calculated for each overlapping marker and a single file containing allele frequencies of consistent alleles is produced. The program (COMBI.PL) is written in PERL and available at http://data40.uni-tz.gwdg.de/~htaeube.

Keywords: allele size, combination, microsatellite markers

Received 25 July 2007; revision accepted 29 August 2007

Microsatellite markers are the tool of choice for many population genetic studies as a consequence of their multiallelic variability, their ubiquity in vertebrate genomes and availability of typing technologies. One feature of work within many species (for example livestock) is that of multiple parallel investigations with some overlap in breed/population samples and markers. There is a need to combine data collections but this can pose difficulties. Whereas other data types such as resequencing and single nucleotide polymorphism (SNP) alleles are readily portable, meta-analyses of microsatellite data that might address wide phylogeographical questions are very frequently constrained by inconsistent allele-size calling between platforms, between laboratories and even between users.

Problems in combining multilocus microsatellite systems are recently described by Pasqualotto *et al*. (2007). Here, allele sizes from two genotyping laboratories, which both used capillary electrophoresis for sizing, were compared by sequencing. Size differences of up to 6 bp were encountered between the sequence results and those estimated from genotyping. In addition, because the two genotyping laboratories used different machines in running conditions, allele size differences of up to 3 bp were identified between them.

Correspondence: Helge Täubert, Fax: +49551395587; E-mail: htaeube@gwdg.de

Differences in microsatellite allele lengths between laboratories have also been described, for example in San-Cristobal *et al*. (2006). Pig marker diversity data from the PigBioDiv project (Blott *et al*. 2003), using an ABI PRISM 3700 multicapillary sequencer in one laboratory and an ABI PRISM 377 sequencer in the second laboratory were combined with data from the PigMap study (Archibald *et al*. 1995). Inconsistencies in allele-size calling were obvious and these were resolved by genotyping a common four animals from the PigMap reference population in both laboratories in order to set up conversion tables.

In lieu of resequencing, we propose a mathematical solution to combine multilocus microsatellite information based on partial overlaps of sampled breeds between two studies.

Figure 1 shows two typical patterns of allele frequency differences encountered when data is shared. Each graph illustrates the results of parallel analyses of a single microsatellite marker typed in alternate samples with similar breed origins: marker Hel5 in Hereford cattle and marker ETH225 in Holstein cattle. On the left, a constant shift between study 1 and 2 can be seen. The alleles have the same frequency pattern, but the allele lengths are shifted. On the right, study 1 has odd and study 2 has even allele lengths but sequence of alleles in both studies remains constant. Such differences may be corrected by eye but this can be inconsistent and may be a source of error.

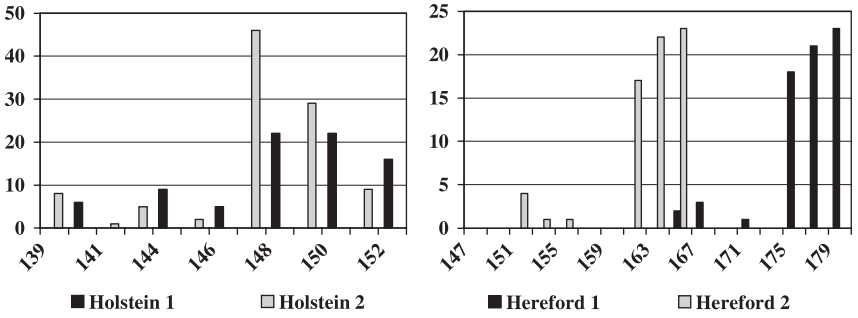

Fig. 1 Allele frequencies (in percentage) of two breeds typed for the same two markers in different studies. Labels on the *x*-axis show the measured allele lengths of the same marker.

Table 1 Example of alleles found of same breed and marker in two studies

Study 1			Study 2		
Allele size	*N*	*P*	Allele size	*N*	*P*
134	18	0.30	153	13	0.325
136	24	0.40	155	18	0.450
138	3	0.05	159	9	0.225
140	15	0.25			
	60	1		40	1

Table 2 All possible allele combinations and their likelihood

Combinations	1	2	3	4
	134 → 153	134 → 153	134 → 153	134 →
	136 → 155	136 → 155	136 →	136 → 153
	138 → 159	138 →	138 → 155	138 → 155
	140 →	140 → 159	140 → 159	140 → 157
Likelihood	–25.6	–19.4	–35.6	–34.0

We describe an algorithm based on maximum-likelihood calculation to combine the allele information. The basis of this algorithm is the following assumption:

The expected frequency of allele *l* from marker *m* in breed *i* in study 2 is equal to the correspondent frequency in study 1.

$$E(p_{iml2}) = p_{iml1}$$

with *i*, breed; *m*, marker; and *l*, allele

For simplification, only alleles within the same breed and marker are used here.

For a specific combination of alleles of one marker in study 1 to be the same as in study 2, the likelihood can be calculated.

$$L = \prod_{1}^{l_1} p(l_1)^{n_{l_2}} \quad (1)$$

l_1, allele in study one; $p(l_1)$, frequency of allele *l* in study one; n_{l2}, number of allele *l* in study 2

expressed as log(*L*)

$$\log(L) = \sum_{1}^{l_1} n_{l_2} \cdot \log(p_{l_1}) \quad (2)$$

Between n_{m1} alleles in study 1 and n_{m2} alleles in study 2 there are

$$n_{m1}Cn_{m2} = \frac{n_{m1}!}{n_{m2}! \cdot (n_{m1} - n_{m2})!} \text{ (with } n_{m1} \geq n_{m2}) \quad (3)$$

possible combinations, because the order of the alleles must not be changed. The best combination of the alleles is the one with the highest likelihood.

To illustrate, we use the following example: assuming only one common breed and marker in two studies, the following alleles were found:

In study 1, four alleles and in study 2, three alleles were found (Table 1). To find out which allele of study 2 corresponds to in study 1, all possible combinations and their likelihood need to be calculated (Table 2).

Number of possible combinations with $n_{m1} = 4$ and $n_{m2} = 3$ is 4C3 = 4.

The likelihood of combination 2 is the highest, flagging the best allele correspondence.

Merging studies with different breeds is more challenging because of fluctuations in allele spectra between breeds that are due to natural genetic diversity. The allele frequencies are not expected to be equal, which is the basic assumption of our algorithm. In these cases, the allelic shift will be calculated from common (overlapping) breeds in the studies, where the genetic diversity is assumed to be zero. The calculated constant shift will afterwards be transferred to the alleles of nonoverlapping breeds.

The algorithm only works correctly, if there is a constant shift of allele length between two studies. This can be expected, if the molecular analysis is conducted properly (Pasqualotto *et al*. 2007). Other patterns, such as gaps in the allele lengths or a shift with a change in variance can be a sign of other problems in the data. In this case, the alleles

should not be combined and a marker should be omitted from the data set.

A program was developed in PERL to perform the computations and to combine different data sets. It can be used on Windows and Unix platforms and is available at http://data40.uni-tz.gwdg.de/~htaeube. The PERL modules Math::BigInt and Algorithm::Combinatorics, which are freely available on the CPAN website (www.cpan.org) need to be installed.

The input files have to be semicolon (;) delimited text files, for example CSV files from Excel. Two input files will be read, one for each data set to be combined. The first line contains the names of the markers (twice for each allele) in the same order they appear in the data. The following lines show the marker alleles for each individuals. The first column in each data line is the name of the breed. It may have an extension number for each individual. Then the alleles will follow, two of each marker.

It is important to name the markers and breeds in both data sets the same (case sensitive).

The program calculates the allelic shift and writes the calculated shift for each marker on the screen. In cases of nonconstant shifts, the user will be asked to revise the assignments. In some cases, it may happen that two allele combinations have the same likelihood. In this case, the user is asked to revise the assignments, too. If a marker has no constant shift at the end of the program, it will be omitted from the data set.

Three output files are written to disk.

INFILE is the combined data set containing the allele frequencies in PHYLIP format (Felsenstein 1989). It can directly be used to calculate genetic distances. If a marker was not genotyped in one breed, all frequencies of this marker show '–99', in order to avoid confusions with zero alleles.

ASSIGN.OUT is the information of assignments calculated for each marker.

ALLELE_COMBI.OUT is a file containing adjusted allele information of all individuals of the two input files in GENEPOP format.

On the website, two sample input data sets are freely available, EX1.TXT and EX2.TXT, containing encoded data from 65 breeds and 31 markers to illustrate the procedure of combining partly overlapping data sets.

References

Archibald AL, Haley CS, Brown J *et al.* (1995) The PiGMaP consortium linkage map of the pig (*Sus scrofa*). *Mammalian Genome*, **6**, 157–175.

Blott S, Andersson L, Groenen M *et al*. (2003) Characterisation of genetic variation in the pig breeds of China and Europe — the PigBioDiv2 project. *Archivos de Zootechnica*, **52**, 207–217.

Felsenstein J (1989) PHYLIP — phylogeny inference package. *Cladistics*, **5**, 164–166.

Pasqualotto AC, Denning DW, Anderson MJ (2007) A cautionary tale: lack of consistency in allele size between two laboratories for a published multilocus microsatellite typing system. *Journal of Clinical Microbiology*, **45** (**2**), 552–528.

SanCristobal M, Chevalat C, Haley CS *et al*. (2006) Genetic diversity within and between European pig breeds using microsatellite markers. *Animal Genetics*, **37**, 189–198.