Chapter 18

Informativeness of Microsatellite Markers

M. Humberto Reyes-Valdés

Abstract

Simple sequence repeats (SSR) are extensively used as genetic markers for studies of diversity, genetic mapping, and cultivar discrimination. The informativeness of a given SSR locus or a loci group depends on the number of alleles, their frequency distribution, as well as the kind of application. Here I describe several methods for calculating marker informativeness, all of them suitable for SSR polymorphisms, proposed by several authors and synthesized in an Information Theory framework. Additionally, free access software resources are described as well as their application through worked examples.

Key words Marker informativeness, Microsatellites, Information theory, PIC, Coancestry, Cultivar discrimination, QTL mapping, Software

1 Introduction

1.1 Microsatellites

Simple sequence repeats, also called microsatellites, are ubiquitous on eukaryotic genomes. They are usually composed by di- or three-nucleotide sequences, repeated around ten times. Their sequence patterns induce hypervariability in the number of repeats across any given locus, due to phenomena related to DNA replication and recombination. This high variation in length has proven to be highly useful for genetic marking, scored through amplification by the polymerase chain reaction (PCR). As it has been the case with other genetic markers, microsatellite polymorphisms have been successfully applied in areas such as the study of genetic diversity, genetic mapping, and cultivar identification.

The informativeness of SSR markers varies across loci and populations. It depends mainly on the number of alleles and their frequencies. Furthermore, their informativeness depends on the type and strategy of application. Thus, it is important to identify informative SSR markers and quantitatively evaluate their informativeness in order to delineate optimum strategies for their use, in terms of maximum efficiency and minimum cost.

Stella K. Kantartzi (ed.), *Microsatellites: Methods and Protocols*, Methods in Molecular Biology, vol. 1006,
DOI 10.1007/978-1-62703-389-3_18, © Springer Science+Business Media, LLC 2013

1.2 Information Theory

I will base the general approach to informativeness calculation for SSR markers in the framework of information theory, a branch of mathematics dedicated to the storage, transmission, recovering, and measuring of information. The pioneer work in this subject was made by Claude Shannon (1), while he was working for the Bell Laboratories. His theory was based on the so-called information channel, which comprises a source of information, an encoder, a noisy channel, a decoder, and a destination. A key concept in information theory is the Shannon entropy, a measure of uncertainty. For a discrete variable, the Shannon entropy of the variable M is given by the following equation:

$$H(M) = -\sum_{i=1}^{g} p_i \log_2 p_i,$$

where $p_1, p_2, \ldots, p_g$ are probabilities assigned to the possible values of $m_1, m_2, \ldots, m_g$ of a random variable M. For g possible values of a discrete random variable, the maximal value of the Shannon entropy is $\log_2(g)$, occurring when $p_1 = p_2 = \ldots = p_g$, whereas the minimum is 0 for any $p_i = 1$. In the previous equation, the expression $0 \log_2(0)$ equals 0 by definition. Based on the entropy concept, the mutual information between two variables X and M is defined as the average reduction in the uncertainty about X given knowledge of the value of M, in accordance with the following expression:

$$I(X;M) = H(X) - H(X \mid M) = H(M) - H(M \mid X),$$

where $H(X|M)$ is the average entropy or uncertainty in X, given knowledge of the value of the variable M. Information is symmetrically defined in terms of entropies; in fact, the expression for $I(X;M)$ can also be defined as the information conveyed about a variable M by the variable X, and it can be also written as $I(M;X)$. The Shannon entropy has been applied in several situations involving genetic markers, for example, in the measurement of linkage disequilibrium (2), inference of ancestry (3), SNP selection for association studies (4, 5), statistics for association (6), information for QTL mapping (7), and transcriptome analysis (8). The entropy concept can be used as a general, firmly mathematically founded framework for calculating information provided by genetic markers for several applications.

1.3 Informativeness for Genetic Markers

The so-called Polymorphism Information Content or PIC (9) is a statistic defined to one particular type of human pedigree: one parent is affected by a rare dominant disease and is heterozygous at the disease locus, whereas the other parent is unaffected by the disease. This locus is associated with a marker with several codominant alleles. In this context, an offspring is said to be informative if we can infer from his genotype which marker allele is co-inherited with

the disease allele. Thus, PIC is defined as the expected fraction of informative offspring from this type of pedigree (10). The expression for this statistic is assuming Hardy–Weinberg equilibrium:

$$\mathrm{PIC} = 1 - \sum_{i=1}^{a} p_i^2 - \sum_{i=1}^{a-1} \sum_{j=i+1}^{a} 2(p_i p_j)^2,$$

where p_i is the frequency of the i-th marker allele and a is the number of different alleles. Since PIC is the proportion of completely informative offspring, and each informative offspring allows the choice between two possible alleles as the co-inherited one, thus producing a mutual information of 1, it can also be considered as average mutual information in accordance with the Shannon theory. Alternatively, for the same type of application, heterozygosity can be used and it is estimated as follows:

$$\mathrm{HET} = 1 - \sum_{i=1}^{a} p_i^2.$$

The PIC statistic will always be equal or lower than heterozygosity, both measures being strongly correlated.

An informativeness expression, often called PIC too (11), has been used as a part of a strategy for the choice of parents in the construction of linkage maps with RFLP markers, based on the concept of gene diversity (12):

$$\mathrm{GS} = 1 - \sum_{i=1}^{g} p_i^2,$$

where p_i is the frequency of the i-th RFLP pattern or any given marker with g genotypes. Since this expression is identical to the Gini–Simpson index (13, 14), originally applied to diversity analysis in ecology, I denote this index as GS. This expression with marker genotype frequencies has been used in several works for cultivar discrimination. It is useful because it estimates, for a large sample, the probability that two random chosen individuals or lines from a population have different banding patterns (15).

An appealing alternative to calculate information for homogeneous cultivar discrimination with marker data is the direct use of Shannon entropy, with p_i being the frequency of the i-th single locus or multilocus marker genotype. In fact, if cultivars are homogeneous, e.g., lines, hybrids, or clones, the conditional entropy of genetic markers given cultivars, say $H(M|X)$, becomes 0; thus, $I(X;M)$ becomes $H(M)$, i.e., the entropy of the distribution of marker frequencies. The following properties are fulfilled by this application of the Shannon entropy to N cultivars: (i) the minimum value is 0, and it is reached when the frequency of any marker genotype equals 1; (ii) the maximum value is $\log_2(N)$, and it occurs

only when the marker genotypes allow distinction of all cultivars; (iii) for g marker genotypes, the maximum value, $\log_2(g)$, is attained when all of them have the same frequency; and (iv) the simultaneous mutual information provided by a set of independent markers is the sum of the individual marker information contents. Thus, the value of the Shannon entropy gives the information that the same number of independent, fully informative, binary loci would theoretically provide, or shortly, it is the *effective number of binary loci*. From information theory, it turns out that the number of bits required to distinguish each unit among a set of N equiprobable cultivars is $\log_2(N)$.

Marker informativeness for inference of coancestry has been proposed with an information theory basis (3). The methodology was developed mainly for genetic mapping in humans, with the key parameter being informativeness for assignment for a given locus (I_n):

$$I_n(Q;J) = \sum_{j=1}^{a}\left(-p_j \log p_j + \sum_{i=1}^{N}\frac{p_{ij}}{N}\log p_{ij}\right),$$

where p_j is the average frequency of the allele j across N populations and p_{ij} is the frequency of the allele j in population i. This is the mutual information between the population Q and an individual allele J. For a given set of populations, the minimum value of I_n occurs when all alleles have the same frequency across populations, and the maximal value $\log(N)$ occurs when $a \geq N$ and no allele is found in more than one population.

The entropy-based founder informativeness was developed for QTL analysis (7). The goal of this statistic is to measure the amount of information about the putative QTL genotype in a given genome site in a linkage map. Assume that for a given locus in a mapping population, there are f putative QTL genotypes, e.g., QQ, Qq, and qq, with probabilities p_1, p_2, …, p_f. The entropy-based founder informativeness, based on marker information, at the map location m in a given member of a mapping population is

$$\mathrm{EFI}(m) = \mathrm{Max}(H) + \sum_{i=1}^{f} p_i \log_2 p_i,$$

where $\mathrm{Max}(H)$ is the maximum entropy of the ensemble of putative QTL genotypes, calculated without marker information and assuming Mendelian segregation. The same paper (7) provides a table for $\mathrm{Max}(H)$ in several mapping populations, whereas probabilities of putative genotypes are calculated in most QTL analysis approaches. The $\mathrm{EFI}(m)$ values averaged for population members at regular intervals across a linkage map allow drawing an information content map.

The approaches described herein can be applied to several types of genetic markers, and all of them are suitable for SSR polymorphisms.

2 Software

I briefly describe the software that can be used to perform the above calculations. However, this list does not discard other alternatives.

2.1 R

R (16) is free software for statistical computing and graphics. It runs in a wide variety of Unix versions, as well as Windows and MacOSX. It can be downloaded from http://www.r-project.org/ and it has a wide availability of packages for diverse applications.

2.2 R/qtl

R/qtl (17) is an R package for QTL analysis in experimental crosses that allows importing data from different standard formats. It uses several methods for QTL analysis, like maximum likelihood and linear regression. Also, it allows numerical calculation of statistical thresholds through permutation tests. Documentation and several tutorials can be downloaded from the R site http://www.r-project.org/.

2.3 Infocalc

The *infocalc* application (18) is a small Perl script, developed by Noah Rosenberg, for calculating statistics for ancestry information content of genetic markers (3). It can be downloaded at the site http://www.stanford.edu/group/rosenberglab/infocalc.html. The instructions are inside the script.

3 Methods

3.1 Polymorphism Information Content

To calculate PIC (9), the following R function can be used with a vector of allele frequencies as argument:

```
pic<-function(x){1-sum(x^2)-sum(x^2)^2+sum(x^4)}
```

Suppose that we have the following set of allele frequencies for a random mating Mendelian population: 0.1, 0.5, 0.2, 0.2. Paste the function on the R console, and after the > prompt execute the `pic` function with its arguments:

```
> pic(c(0.1,0.5,0.2,0.2))
[1] 0.6102
```

In this way we calculate the PIC value of 0.6102. To calculate the maximum PIC for a alleles, use the following function:

```
mPIC<-function(a){(a-1)^2*(a+1)/a^3}
```

Heterozygosity, often called PIC, can be calculated for a vector x of allele frequencies with the following R function:

```
het<-function(x){1-sum(x^2)}
```

For a sample size of n alleles, i.e., $2N$ diploid individuals, an unbiased estimation of heterozygosity is

```
het.unbiased<-function(x,n){het(x)*n/(n-1)}
```

Example: For a set of estimated allele frequencies: 0.1, 0.5, 0.2, 0.2, from a sample of 50 individuals, proceed as follows:

Paste in the R console both functions het and het.unbiased, and press <Enter>

For the basic heterozygosity estimation, write

```
> het(c(0.1,0.5,0.2,0.2))
[1] 0.66
```

To get an unbiased estimation of heterozygosity, type

```
> het.unbiased(c(0.1,0.5,0.2,0.2),100)
[1] 0.6666667
```

There is a web calculator of the PIC statistic and biased heterozygosity as the one given by the het function, designed by Steve Kemp (19).

3.2 Gini–Simpson Index for Genotypic Frequencies

The Gini–Simpson index, often used for cultivar diversity or informativeness for cultivar discrimination, can be calculated with the het function, applied on genotypic frequencies.

Example: Consider the following set of frequencies of cultivar marker genotypes: 0.4, 0.1, 0.2, 0.15, 0.15.

Paste and execute the het function in the R console, and type

```
> het(c(0.4,0.1,0.2,0.15,0.15))
[1] 0.745
```

3.3 Mutual Information for Cultivar Discrimination

If we have a set of homogeneous cultivars, e.g., lines, hybrids, or clones, we can estimate the mutual information between one or more marker loci and cultivar identity, thus providing a measure of the discrimination ability of the marker set. The raw material for calculation is the set of frequencies of marker genotypes, which is in turn used to calculate the Shannon entropy. The following R functions allow the necessary calculations:

```
MyLog2p<-function(x){if(x==0)    0    else    x*log(x,2)}
#Defining x logx
entropy<-function(x){-sum(sapply(x,MyLog2p))}
```

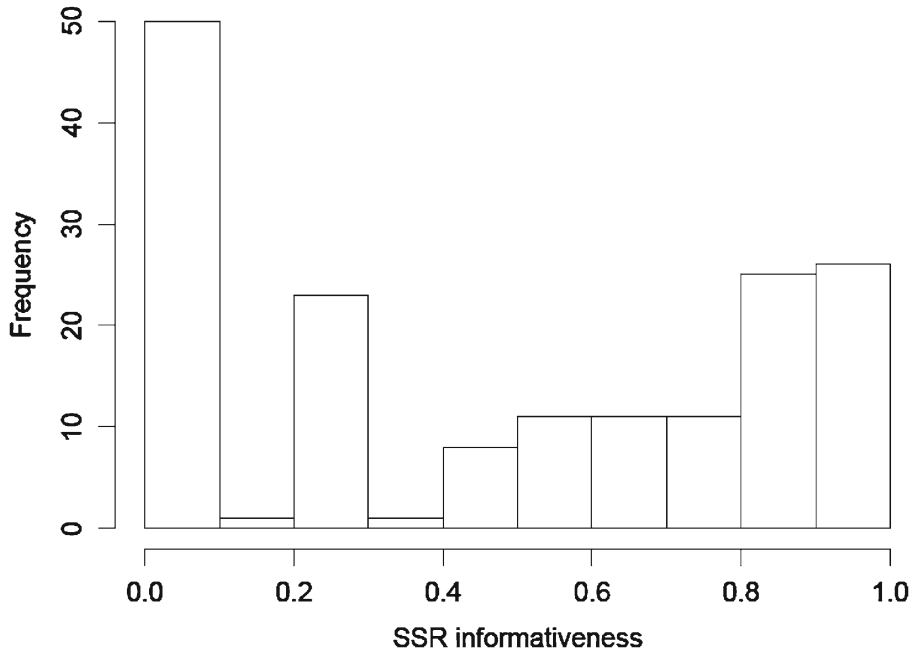

Fig. 1 Entropy-based SSR informativeness for 167 biallelic loci in soybean. Data provided by Stella Kantartzi

Assume a set of cultivars, with marker frequencies 0.091, 0.008, 0.005, 0.022, 0.086, 0.029, 0.090, 0.036, 0.047, 0.040, 0.012, 0.011, 0.087, 0.030, 0.034, 0.059, 0.042, 0.189, 0.013, 0.069. Proceed as follows:

```
> entropy(c(0.091,0.008,0.005,0.022,0.086,0.029,0.090,
0.036,
0.047,0.040,0.012,0.011,0.087,0.030,0.034,0.059,0.042,
0.189,
0.013, 0.069))
[1] 3.86779
```

This means that the marker information available for cultivar discrimination is 3.86779 bits, equivalent to the same number of fully informative independent binary markers, and enough to discriminate among $2^{3.86779} = 14.6$ cultivars. Obviously, there cannot be 14.6 cultivars, but this pictures the availability of information.

As a further example, I analyze homozygous SSR data on soybean lines from a biparental cross, kindly provided by Dr. Stella Kantartzi. In this case, information ranges from 0 to a maximum of 1, given that the biparental origin allows a maximum of two alleles. An informativeness of 0 is obtained for the SSR loci with the same genotype across lines, whereas the value of 1 is calculated for loci with a 50 % frequency of each genotype. In Fig. 1, we can appreciate the informativeness distribution of 167 SSR loci.

3.4 Marker Informativeness for Inference of Coancestry

The software *infocalc* (18) is used to calculate marker informativeness for coancestry, with one of the main parameters being informativeness for assignment (I_n). The used instructions are in the respective Perl script. The data file follows the STRUCTURE format, whose first line denotes the names of marker loci. The following lines include the genotype data of individuals, with the first five columns being individual identifiers, followed by the allele code for each locus. Each individual genotype is represented by two lines, with the order of the two alleles being irrelevant. Missing data are marked with a particular value, –9. The following five lines, taken from *infocalc*, represent codification for two individuals:

```
D9S1779   D9S1825   D7S2477   D17S784   D16S403   D3S1262
D10S189
854 86 Maya Mexico AMERICA 124 129 152 -9 138 112 186
854 86 Maya Mexico AMERICA 142 135 156 -9 140 124 186
855 86 Maya Mexico AMERICA 124 129 156 230 138 112 186
855 86 Maya Mexico AMERICA 124 129 164 234 140 112 186
```

The first line contains the names of seven marker loci. For each of the subsequent lines, the first five columns are individual code, population code, population name, country, and geographical region. The numbers that follow are either allele codes or the code –9 for missing data. Thus, for the Mayan individual coded with 854, the marker genotype for locus D9S1779 is the set of alleles 124 and 142, whose order does not indicate phase, thus being interchangeable. Weights can be defined, so a nonuniform prior for the populations can be accommodated.

If you use Unix, Linux, or MacOS X, Perl is most likely already installed. To get information about your Perl version, type perl –v at a command prompt. For Windows operating systems, the current standard Perl distribution is ActivePerl, from ActiveState, at http://www.activestate.com/ActivePerl/.

Example: I use the dataset provided by the *infocalc* site, mksp.stru, for data on four human populations: Maya, Karitiana, Suri, and Pima. For an unweighted analysis, proceed as follows:

Make the directory containing the dataset your home folder. Then type

```
./infocalc -column 3 -numpops 4 -input mksp.stru -output mksp.stru.out.txt <Enter>
```

The option `-column 3` states the population identifier column, –numpops 4 is the number of populations, `-input mksp.stru` is for the input file, and `-output mksp.stru.out.txt` is for the output file. The results are displayed as follows:

```
Locus    I_n       I_a       ORCA[1-allele]  ORCA[2-allele]
D10S189 0.761877  0.130766  0.61756         0.727457
D16S403 0.854949 0.167937  0.745536        0.87205
D17S784 0.342555  0.0572477 0.4625          0.599987
D3S1262 0.23707   0.0472184 0.460417        0.552591
```

```
D7S2477 0.332763 0.0614692 0.494048 0.607001
D9S1779 0.259119 0.0530722 0.416667 0.537326
D9S1825 0.0531744 0.0111935 0.327083 0.377795
Command: infocalc -column 3 -numpops 4 -input mksp.stru
-output mksp.stru.out.txt -weightfile [none]
PriorWeights: Karitiana 0.25 Maya 0.25 Pima 0.25 Surui
0.25
```

Besides calculating informativeness for assignment (I_n) for each locus, *infocalc* performs calculation for I_a, the informativeness for ancestry coefficients in the admixture model, and ORCA, the optimal rate of correct assignment (3). The last two lines recapitulate the options and the weights.

3.5 Information for QTL Mapping

Informativeness maps for QTL analysis can be drawn across linkage maps through entropy-based founder informativeness "EFI" (7). Since calculation requires probabilities of putative QTL genotypes across the linkage map, the R/qtl package may be used for common mapping populations. The trick is to extract those probabilities. In the example below, I use an anonymous recombinant inbred line (RIL) dataset and show how to perform this calculation. Since there are only two possible QTL genotypes, one from each parent, the maximum entropy is 1; thus, the entropy of the distribution of the putative QTL genotypes must be subtracted from 1 across the linkage map. I will not extend on details of how to use R/qtl, since the subject is extense and excellently covered by several manuals and one book (20).

Once the map file and the genotype file are saved on the working directory, the following script is executed in an R console:

```
> #Drawing a QTL information map
> dat<-
+
read.cross("mm",file="genotypes.raw",mapfile="mapR.map")
+      #Retrieving data in a MapMaker format
> class(dat)[1]<-"riself" #Declaring RILs
+   jittermap(dat) #Rectify markers located at the
+   same position
> dat <- calc.genoprob(dat, step=1,
+   error.prob=0.01,map.function="kosambi")#Calculate
+   probabilities, assuming a genotyping error rate of
    0.01
> attach(dat$geno[[3]])#Attaching probabilities to
  linkage
+   group 3
> dim(prob)#Check dimensions of the probability data
> MyLog2p<-function(x){if(x==0) 0 else x*log(x,2)}
  #Define
+   function plog2p
> entropy<-function(x){-sum(sapply(x,MyLog2p))}
+  a<-NULL;length(a)<-dim(prob)[2];for(i in 1:dim(prob)
+   [2])a[i]<-1-mean(apply(prob[,i,1:2],1,entropy))
```

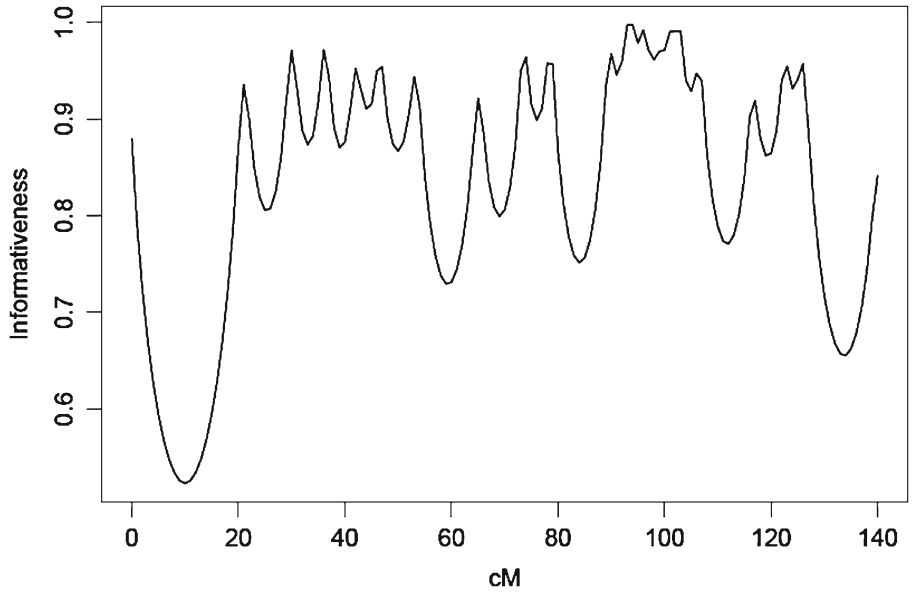

Fig. 2 Informativeness for QTL mapping along a 140 cM linkage group. Peaks correspond to SSR marker positions

```
> plot(cbind(c(0:140),a),type="l",xlab="cM",
+  ylab="Informativeness+   ")#Information   map   for
   linkage
+  group 3
```

The informativeness map for linkage group 3 is depicted in Fig. 2. The upward peaks on the plot represent local maxima for average EFI, and those points are usually located at the marker loci. One may note that even at the marker sites, EFI has not its maximum attainable of 1. There are two reasons: the first one is that we are setting a genotyping error rate of 0.01 and the second one is the presence of missing data records.

4 Notes

4.1 Recommendations on the use of R

The R object names are case sensitive; thus, one must be very careful when writing down commands and variables. One of the main problems with analyzing data is to have a correctly structured dataset in the correct directory and with a fairly simple name. In general, for file names one must keep in mind the following recommendations: (i) consider that file names are case sensitive in several systems; (ii) you can use upper and lower case letters, dots, numbers, and underscore symbol; (iii) it is better to avoid blank spaces; (iv) avoid the following characters in file names: "/," "&," "|," ":," ">," and "<." The character "/" is reserved as a directory and file name separator in a pathname; (iv) start your names with a letter or a number; and (v) make your names short but not

cryptic. Datasets saved in tab separated plain text files, and in comma separated (csv) files work very well, and can be exported from Excel, OpenOffice, and LibreOffice, among others.

Sometimes things go wrong in R, because the columns in a data frame are not in the desired format. Therefore, one must make sure that we are dealing either with a factor or a numeric vector, etc., by using the class command. For example, if we are interested in analyzing a numerical vector *x*, and the command `class(x)` gives `factor` as the output, then we need to convert the variable. The following instruction has worked fine on my experience: `x<-as.numeric(as.vector(x))`.

Acknowledgements

I am thankful to Stella Kantarzi, who provided soybean SSR data to be used in one of the examples; to Noah Rosenberg, who reviewed the material related to his developments in marker informativeness; and to José Reyes, who checked my R scripts.

The R functions used in this book chapter can be accessed through the following link: http://www.uaaan.mx/~mhreyes/FunctionsChapterSSR.html.

References

1. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(379–423):623–656
2. Nothnagel M, Fürst R, Rhode K (2002) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered 54:186–198
3. Rosenberg NA, Li LM, Ward R et al (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422
4. Hampe J, Schreiber S, Krawczak M (2003) Entropy-based SNP selection for genetic association studies. Hum Genet 114:36–43
5. Butler JM, Bishop DT, Barrett JH (2005) Strategies for selecting subsets of single-nucleotide polymorphisms to genotype in association studies. BMC Genet. doi:10.1186/1471-2156-6-S1-S72
6. Zhao J, Boerwinkle E, Xiong M (2005) An entropy-based statistic for genomewide association studies. Am J Hum Genet 77:27–40
7. Reyes-Valdés MH, Williams CG (2005) An entropy-based measure of founder informativeness. Genet Res 85:81–88
8. Martínez O, Reyes-Valdés MH (2008) Defining diversity, specialization, and gene specificity in transcriptomes through information theory. Proc Natl Acad Sci USA 105:9709–9714
9. Botstein D, White RL, Skolnick M et al (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331
10. Hildebrand CE, Torney DC, Wagner RP (1992) Informativeness of polymorphic DNA markers. Los Alamos Sci 20:100–102
11. Anderson JA, Churchill GA, Autrique JE et al (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181–186
12. Weir BS (1990) Genetic data analysis. Methods for discrete genetic data. Sinauer Associates, Inc., Sunderland, MA
13. Simpson EH (1949) Measurement of diversity. Nature 163:688
14. Jost L, Baños T (2006) Entropy and diversity. Oikos 113:373–375
15. Tessier C, David J, This P et al (1999) Optimization of the choice of molecular markers for varietal identification in *Vitis vinifera* L. Theor Appl Genet 98:171–177
16. R Development Core Team (2012) R: a language and environment for statistical comput-

ing. R Foundation for Statistical Computing. http://www.R-project.org
17. Broman KW, Wu H, Sen S et al (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890
18. Rosenberg N (2006) Infocalc—a program for calculating marker informativeness statistics. Version 1.1. http://www.stanford.edu/group/rosenberglab/infocalc.html
19. Kemp S (2002) PIC calculator. http://www.stanford.edu/group/rosenberglab/infocalc.html
20. Broman KW, Sen S (2009) A guide to QTL mapping with R/qtl. Springer, New York