Genet. Res., Camb. (2005), **85**, *pp.* 81–88. *With 5 figures.* © 2005 Cambridge University Press
doi:10.1017/S0016672305007354 Printed in the United Kingdom

An entropy-based measure of founder informativeness

M. HUMBERTO REYES-VALDÉS[1]* AND CLAIRE G. WILLIAMS[2]
[1] *Departamento de Fitomejoramiento, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo, Coah., Mexico, C.P. 25315*
[2] *Nicholas School of the Environment and Earth Sciences, Duke University, Durham NC 27708, USA*

(*Received 22 June 2004 and in revised form 18 October and 22 December 2004*)

Summary

Optimizing quantitative trait locus (QTL) mapping experiments requires a generalized measure of marker informativeness because variable information is obtained from different marker systems, marker distribution and pedigree types. Such a measure can be derived from the concept of Shannon entropy, a central concept in information theory. Here we introduce entropy-based founder informativeness(*EFI*), a new measure of information content generalized across pedigrees, maps, marker systems and mating configurations. We derived equations for inbred- and outbred-derived mapping populations. Mathematical properties of *EFI* include enhanced sensitivity to mapping population type and extension to any number of founders. To illustrate the use of *EFI*, we compared experimental designs for QTL mapping for three examples: (i) different marker systems for an F_2 pedigree, (ii) different marker densities and sampling sizes for a BC_1 pedigree and (iii) a comparison of haplotypic versus zygotic analyses of an outbred pedigree. As an *a priori* generalized measure of information content, *EFI* does not require phenotypic data for optimizing experimental designs for QTL mapping.

1. Introduction

Designing quantitative trait locus (QTL) mapping experiments requires a generalized measure of informativeness because real maps deviate from even marker saturation and fully informative marker configurations. Informativeness of a map or a segment of linked markers varies depending on mapping population, marker system, mating configuration, punctuate meiotic recombination and marker density. Thus, a practical approach is to ask how close a map or a DNA segment comes to optimizing QTL mapping for a given experimental design, and compare between maps and segments. Another use is measuring information content for a given DNA segment where many marker observations are missing, or several markers may have been scored in a dominant fashion.

At present, two measures of informativeness are used to determine the amount of information available for QTL analysis. The first is the R^2 value, which was developed for evaluating genetic maps of human pedigrees (Kruglyak & Lander, 1995). The R^2 measures the linear fit between the actual and predicted identity-by-descent (IBD) sharing distribution among sib pairs for each map point along the chromosome (Kruglyak & Lander, 1995). R^2 thus quantifies the predictive ability of marker information to infer IBD sharing. Mapped regions with small R^2 values require increased marker density to optimize QTL mapping. In the BC_1 case, it is shown in Appendix 1 that R^2 for random full-sib pairs is a function of the variance of the genotypic conditional probabilities as follows:

$$R^2(m) = 16\sigma_p^4, \tag{1}$$

where p is the conditional probability of a given individual being heterozygous at a given map point m, as calculated from marker information. This p value is also defined as the founder-origin probability because it is the probability of a given haplotype as a descendant of the donor founder versus a recurrent founder (Reyes-Valdés & Williams, 2002).

The second informativeness measure, which will be called Ψ, was proposed for an outbred three-generation forest tree pedigree with a large sibship

* Corresponding author. Departamento de Fitomejoramiento, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo, Coah., Mexico, C.P. 25315. Tel: +52 (844)411 02 96. Fax: +52 (844) 411 02 11. e-mail: mhreyes@uaaan.mx

(Knott *et al*., 1997). Pairwise matings between two sets of founders are defined as $H_1H_2 \times L_1L_2$ and $H_3H_4 \times L_3L_4$ so re-labelling these four founders as $A_1 \times A_2$ on the maternal side and $A_3 \times A_4$ on the paternal side shows Ψ to be a measure of marker information content at a given map location m. The measure $\Psi(m)$ for the maternal side is defined as follows:

$$\Psi(m) = \frac{4\sum_{i=1}^{n}(P_{i(A1,A3)} + P_{i(A1,A4)} - 0.5)^2}{n}, \quad (2)$$

where n is the offspring size. Here P_i is defined as the probability of a founder-origin combination and $P_{i(A1,A3)} - P_{i(A1,A4)} = p$, where p is the founder-origin probability of donor founder A_1 for a given haplotypic location m on the maternal map. The Mendelian expected value of p is 0·5. The founder-origin probability concept as used here is based on treating each offspring genome as a mosaic of chromosomal segments contributed by different founders (Reyes-Valdés & Williams, 2002). Marker information is used to infer probability of descent for the transmission of a donor's haplotype to its offspring. The parameter Ψ, as originally defined by Knott *et al*. (1997), is haplotypic in nature because it considers only a single founder at a time. In its original outbred application, the parameter Ψ was not intended for inbred-derived pedigree cases such as F_2 intercross and recombinant inbred lines.

On the other hand, R^2 was developed for IBD mapping applications and its calculation is cumbersome for large sibships. Neither method was intended as a generalized measure, but each was well suited to specific cases. Another drawback to both measures is that they both assign a maximum value of 1 regardless of the mapping population. Neither method thus detects differences in maximum informativeness between different pedigree types. For example, neither measure would detect an increase in information content when comparing a BC_1 with a F_2 pedigree with fully informative markers, although the latter clearly has more information.

Another interesting property, as shown in Appendix 1, is that R^2, like Ψ, uses founder-origin probabilities from marker information for the single-founder case. Here Ψ is related to R^2. Note that in the backcross case, Ψ becomes an estimate of $4\sigma_p^2$, the square root of R^2, and the expected relationship is $\Psi^2 = R^2$ assuming Mendelian segregation. This relationship was not apparent when Ψ was first introduced (Knott *et al*., 1997).

2. Information content based on Shannon entropy

We propose a new measure of marker information content, entropy-based founder informativeness (*EFI*), for optimizing QTL mapping designs. *EFI* is based on Shannon entropy, a central concept in information theory (Shannon, 1948). The Shannon entropy, designed for modelling transmission across communication channels, measures the lack of information in a system. Some desirable mathematical properties for measuring informativeness include non-negativity, continuity and symmetry (Taneja, 2001). The value of Shannon entropy has been recognized for other genetics applications (Yockey, 1992) including evolution of biological complexity (Adami *et al*., 2000), linkage disequilibrium (Nothnagel *et al*., 2002) and inference of ancestry (Rosenberg *et al*., 2003).

For a random discrete variable, the Shannon entropy is defined in terms of its probability distribution as:

$$H(p_1, p_2, \ldots, p_k) = -\sum_{i=1}^{k} p_i \log_2 p_i \quad (3)$$

where p_1, p_2, …, p_k are probabilities assigned to the possible values x_1, x_2, …, x_k of the random variable X. Here, the Shannon entropy will have a maximum of $\log_2(k)$ for $p_1 = p_2 = \ldots = p_k$ and a minimum of 0 for any $p_i = 1$.

To illustrate the use of Shannon entropy in evaluating information content for QTL mapping, we describe the case of a single marker locus in a linkage interval. The single marker is used to infer the presence of a given allele in an anonymous or hidden locus within a chromosomal segment or haplotype. The information about the presence/absence of a given allele travels from the index (Q) locus defined as the putative QTL, to the marker (M) locus through a 'noisy channel' (Fig. 1) Noise is caused either (i) by independent chromosome assortment which occurs when M and Q are not linked or (ii) by crossing-over if both M and Q loci are linked within a haplotype.

To illustrate this concept, assume M and Q loci are biallelic. If one allele at the M locus is fixed then the two possible alleles at locus Q have probabilities r and $(1-r)$, where r is the frequency of recombination between loci M and Q. As shown in Appendix 2, the conditional entropy at locus Q given a specific allele at locus M is defined as follows:

$$\begin{aligned} H(Q|M) &= H(M,Q) - H(M) \\ &= -r\log_2(r) - (1-r)\log_2(1-r), \end{aligned} \quad (4)$$

where $H(Q|M)$ is the degree of uncertainty about the allele present in Q given that a certain allele is present at M in a backcross experiment, $H(M)$ is the entropy of the *a priori* distribution of the alleles at M, and $H(M, Q)$ is the joint entropy. Joint entropy is defined as the entropy of the joint probabilistic distribution of alleles at M and Q. The bounds of $H(Q|M)$ are 0 (for $r = 0$) and 1 (for $r = \frac{1}{2}$).

https://doi.org/10.1017/S0016672305007354 Published online by Cambridge University Press

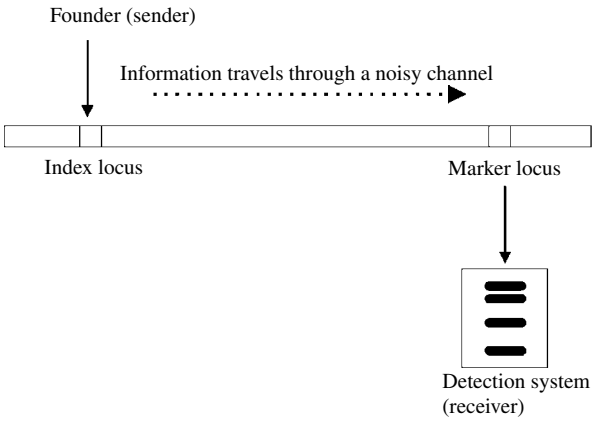

Fig. 1. An application of Shannon entropy to QTL mapping designs. A chromosome segment is shown here as a noisy channel. Information about founder origin travels from the index locus to a marker locus. Meiotic recombination introduces noise into the channel and the resulting uncertainty is measured by Shannon entropy.

The mutual information $I(X;Y)$ is defined as the difference in entropy for X generated by the knowledge of Y, and it is calculated by the difference between $H(X)$ and $H(X|Y)$. Thus, for the single-marker locus example, the mutual information between M and Q can be expressed as follows:

$$I(Q;M) = H(Q) - H(Q|M) = 1 - H(Q|M)$$
$$= 1 + r\log_2(r) + (1-r)\log_2(1-r), \quad (5)$$

where $I(Q;M)$ measures the average reduction in entropy, given in units of information bits, about Q that directly results from inferring the value of M, assuming that the two alleles at locus Q have expected Mendelian probabilities of 0·5.

Inference about the putative QTL increases if calculated from information provided by two or more flanking marker loci. The simultaneous use of multiple marker information is defined as *redundancy* in information theory. Redundancy is a source of multipoint inference in QTL mapping and increases the accuracy of information recovery for index locus Q. For the multipoint case, Shannon entropy will depend on recombination frequencies between the index locus Q and its informative marker loci $M_1, M_2, \ldots, M_n$.

Shannon entropy can be generalized to founder-origin probabilities (Reyes-Valdes & Williams, 2002). In this case, let p_i $(i=1, 2, 3,\ldots, f)$ be the probability of the ith founder or founder-combination at a given map location, conditional on marker information for a given individual in a mapping population. The uncertainty for a given map locus can be measured as follows:

$$H(p_1, p_2, \ldots, p_f) = -\sum_{i=1}^{f} p_i \log_2 p_i. \quad (6)$$

The concept of founder-origin probability is applied here in a broad sense. It can refer to any of the following conditions: (i) the probability of haplotypes descending from an individual founder (Reyes & Williams, 2002), (ii) the probability of a putative QTL genotype in the case of biparental populations (Haley *et al*., 1992; Martínez & Curnow, 1992) or (iii) the zygotic probabilities of combined founders (Haley *et al*., 1994; Knott *et al*., 1997).

EFI(m), as a measure of mutual information at point m, measures the amount of information that loci $M_1, M_2, \ldots, M_n$ convey about the inferred index locus Q located at map position m. *EFI*(m) is defined as follows:

$$EFI(m) = Max(H) + \sum_{i=1}^{f} p_i \log_2 p_i, \quad (7)$$

where $Max(H)$ is the maximum entropy of the system. $Max(H)$ is the Shannon entropy value for the founder-origin probabilities calculated *without* marker information. If $p_1 = p_2 = \ldots = p_f$ is a property of the founder-origin probabilities at m in the absence of marker information then $Max(H) = \log_2(f)$. Note that equality of probabilities does not apply to all mapping populations but that $Max(H) = \log_2(f)$ only occurs when there is an equality of Mendelian probabilities (Table 1).

Equations for estimating *EFI*(m) for several types of QTL mapping populations are presented in Table 1. Equations included the outbred pedigree with an array of full-sib offspring, termed CP pedigrees (Maliepaard *et al*., 1997; Van Ooijen, 2004). The probabilities p_i for founder-origin combinations are also given for the inbred-derived populations as the probabilities of putative QTL genotypes. In the case of the F_2 intercross population (Table 1), the putative QTL genotypes were defined as Q_1Q_1, Q_1Q_2 and Q_2Q_2 with assigned probabilities 0·25, 0·5 and 0·25, respectively, in the absence of any marker information.

The definition of *EFI*(m) applies to a specific map location for a given individual within a mapping population. To draw an information content map it is necessary to combine *EFI*(m) across the individuals of the mapping population. This can be done in two ways: first by obtaining an average of *EFI*(m), defined here as $EFI_A(m)$, or, second, by using a total *EFI* defined here as $EFI_T(m)$, across individuals at a map location m. $EFI_A(m)$ represents the average information or bits per individual at a given map location:

$$EFI_A(m) = \frac{\sum_{j=1}^{n} EFI_j(m)}{n}, \quad (8)$$

where $EFI_j(m)$ is the informativeness at a point m in the jth individual of a mapping population of size n. This is consistent with conventional presentation of information content maps.

https://doi.org/10.1017/S0016672305007354 Published online by Cambridge University Press

Table 1. *Entropy-based founder informativeness, EFI(m), for a range of mapping population types*

Population	No. of founder-combinations	Probabilities without marker information	*Max*(*H*)	*EFI*(*m*)
BC_1	2	[0·5, 0·5]	1	$1-\sum_{i=1}^{2} p_i \log_2(p_i)$
F_2	3	[0·25, 0·50, 0·25]	1·5	$1{\cdot}5-\sum_{i=1}^{3} p_i \log_2(p_i)$
$RIL(F_{t:t+1})^a$	3	$\left[\frac{1}{2}-\left(\frac{1}{2}\right)^t, \left(\frac{1}{2}\right)^{t-1}, \frac{1}{2}-\left(\frac{1}{2}\right)^t\right]$	$(t-1)\left(\frac{1}{2}\right)^{t-1}-\left[1-\left(\frac{1}{2}\right)^{t-1}\right] \times \log_2\left[1-\left(\frac{1}{2}\right)^{t-1}\right]$	$Max(H)-\sum_{i=1}^{3} p_i \log_2(p_i)$
CP	4	[0·25, 0·25, 0·25, 0·25]	2	$2-\sum_{i=1}^{4} p_i \log_2(p_i)$

BC_1 refers to a backcross derived from two inbred lines, F_2 refers to intercross derived from two inbred lines and RIL refers to recombinant inbred lines developed by single-seed descent. The population CP (or cross-pollinated) is equivalent to the full-sib offspring of a three-generation outbred pedigree. In all cases, p_i refers to the probability of the ith founder-origin combination or putative QTL genotype, calculated at a given map position m of an individual with marker information.
[a] It is assumed that marker data come from the tth generation and phenotypes from the $(t+1)$th generation.

The total, EFI_T, represents the number of bits in the population at a given map location, and it has the desirable property of sensitivity to population sample size:

$$EFI_T(m) = \sum_{j=1}^{n} EFI_j(m). \quad (9)$$

The average *EFI* across a linkage group for a population can be obtained by integrating either $EFI_A(m)$ or $EFI_T(m)$. For the last case we have:

$$EFI_{gT} = \frac{1}{l}\int_0^l EFI_T(m)dm, \quad (10)$$

where l is the length of linkage group g and EFI_{gT} is the expected total number of bits across the linkage group. If $EFI_T(m)$ was not integrated across the linkage group then it can be calculated as an arithmetic mean EFI_{gT} from the set of $EFI_T(m)$ values calculated at regular map intervals. The same methods apply to EFI_{gA}, the expected average information for linkage group g.

(i) *Example 1: Comparing marker types in an F_2 pedigree*

Consider a 20 cM chromosome interval between two marker loci in an F_2 intercross mapping population (Table 1). The $EFI_A(m)$ values are expected to be higher for the marker interval between two co-dominant markers than for two dominant markers, which in this example were linked in coupling phase (Fig. 2).

$EFI_A(m)$ values were obtained from the theoretical distributions of putative QTL genotypes along the segment. EFI_{gA} values were calculated by numerical integration of $EFI_A(m)$. The expected average informativeness, EFI_{gA}, was higher for the interval with two co-dominant markers than for the interval with two dominant markers (1·140 vs 0·671, respectively). Maximum differences in $EFI_A(m)$ values occurred at the extreme ends, coinciding with marker locations (Fig. 2).

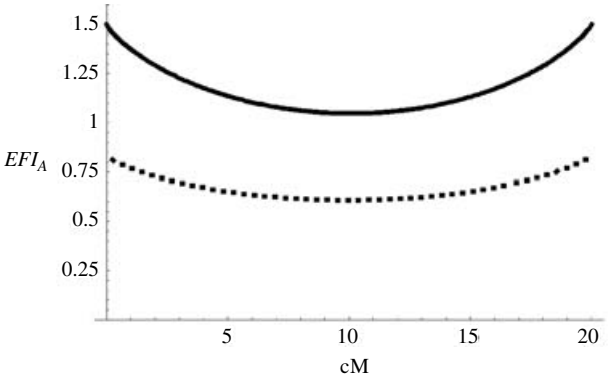

Fig. 2. Comparison of the average entropy-based founder informativenes (EFI_A) through a 20 cM interval for flanking co-dominant (continuous line) and dominant (dotted line) markers in coupling phase. The mapping population is an F_2 intercross.

In the case of an F_2 mapping population, it is formally possible to obtain a negative *EFI*(*m*) value for a given individual in the mapping population. Negative *EFI*(*m*) values indicate high uncertainty about founder origin. This would occur, for example, if the probabilities of a putative QTL are 1/3 for each genotype Q_1Q_1, Q_1Q_2 and Q_2Q_2. This is unlikely to be encountered in practice, as noted for another measure of informativeness, R^2 (Kruglyak & Lander, 1995).

https://doi.org/10.1017/S0016672305007354 Published online by Cambridge University Press

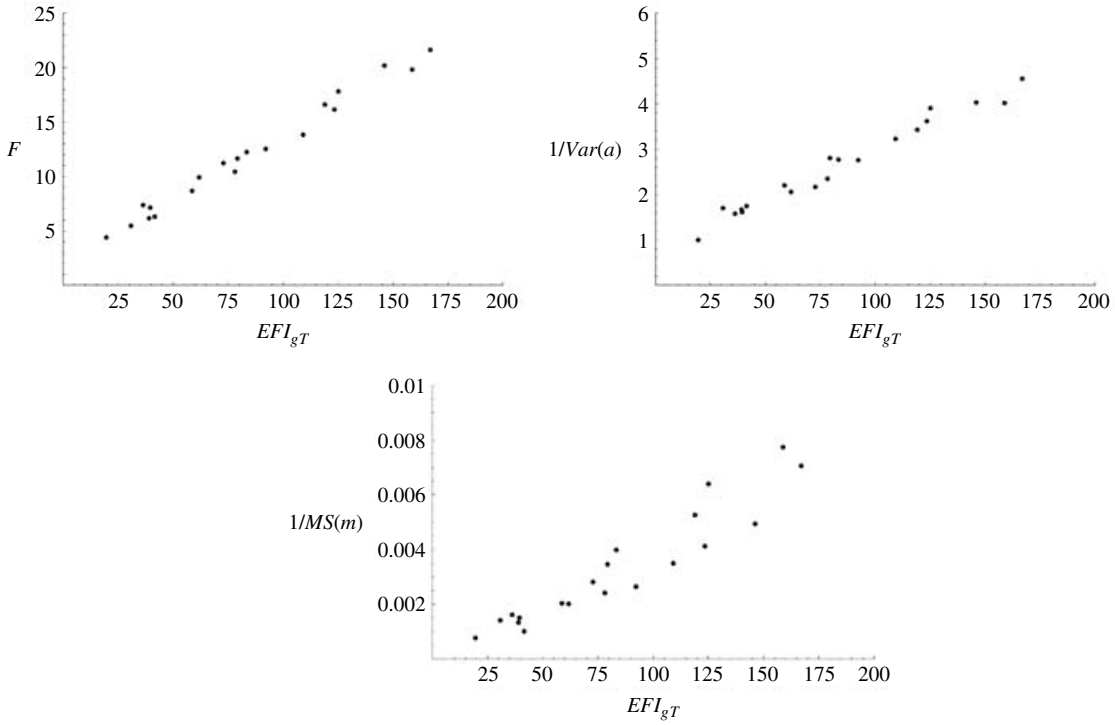

Fig. 3. Relationship between the total, chromosome-wide measure of information content, EFI_{gT}, and a suite of *a posteriori* measures of interval QTL mapping efficacy for the backcross (BC_1) case from simulated data. The indicators are as follows: regression *F* statistic (*F*), inverse of the variance of the estimated effect (1/Var(*a*)) and inverse of the mean square deviation of the estimated map location ($1/MS(m)$).

(ii) *Example 2: Comparing marker density and sample size among BC_1 pedigrees*

Consider a set of simulated backcross mapping populations for a QTL with $h^2 = 0{\cdot}1$ randomly placed in a 100 cM linkage group. The numbers of evenly spaced marker loci were 2, 3, 4, 5 and 6 with population sizes of 50, 100, 150 and 200 individuals. A set of 100 mapping populations or replicates was simulated for each of the 20 combinations. Each replicate was analysed for presence of a QTL using least squares interval mapping.

For each set of 100 replicates, the following statistics were measured: (i) average total informativeness EFI_{gT}, (ii) average *F* test statistic, (iii) precision in the estimation of additive effect, and (iv) precision in the estimation of map location. The precision of QTL effect estimation was recorded as the reciprocal of Var(*a*), where Var(*a*) is the sampling variance of the estimate of *a*. The precision of the map location estimate was measured as the reciprocal of $MS(m)$, where $MS(m)$ is the average squared deviation between real and estimated map location.

The BC_1 equation (Table 1) was used to calculate individual $EFI(m)$ values and then $EFI_T(m)$ at 1 cM increments for each mapping population (replicate). EFI_{gT} was obtained for each replicate by averaging $EFI_T(m)$ across the 100 cM in the linkage group. For each one of the 20 combinations, EFI_{gT} was averaged across the 100 replicates. As expected, the QTL mapping design using six markers and 200 individuals had the highest average EFI_{gT}. The relationships between mapping statistics and average EFI_{gT} are shown in Fig. 3.

The average regression test statistic *F* indicates the power of detection for an interval mapping experiment. Similarly, the inverse of the sampling variance of the QTL effect estimate measures the precision of the effect estimation, and the inverse of the mean quadratic deviation of the estimated QTL from the true QTL position indicates the precision with which positions are estimated. *EFI* is an *a priori* indicator of these *a posteriori* QTL statistics so they varied concomitantly with *EFI* (Fig. 3). Note that *EFI* is calculated without phenotypic data and that it carries no implicit assumptions about the heritability of the phenotypic trait. *EFI* measures marker informativeness without regard to the nature of the trait and its measurement methods, both of which ultimately affect the efficacy of QTL analysis.

(iii) *Example 3: Comparing QTL analysis methods in an outbred pedigree*

Here, $EFI_A(m)$ was used to compare two methods of founder-origin probability use in outbred pedigrees: (i) the zygotic method (Haley *et al.*, 1994) and (ii) its

https://doi.org/10.1017/S0016672305007354 Published online by Cambridge University Press

haplotype-based approximation (Reyes-Valdés & Williams, 2002). Consider a simulated three-generation outbred pedigree with 100 full sibs and a 100 cM linkage group with six linked marker loci. Grandparents or founders were represented by A_1, A_2, A_3 and A_4. The unrelated matings $A_1 \times A_2$ and $A_3 \times A_4$ generated parents P_1 and P_2 respectively. The full-sib array was generated by the $P_1 \times P_2$ mating. Each offspring has one of the following founder-combinations at each of the six marker loci: A_1A_3, A_1A_4, A_2A_3 or A_2A_4. The probabilities for those combinations are p_1, p_2, p_3 and p_4.

The multiple-allele marker genotypic data for grandparents and female (P_1) and male (P_2) parents, separated by commas, were as follows:

A_1	5	6,	3	3,	2	3,	7	6,	7	6,	1	1
A_2	6	6,	3	3,	3	2,	1	8,	3	3,	2	2
A_3	6	6,	1	4,	3	3,	5	5,	1	3,	2	1
A_4	6	5,	3	3,	2	3,	6	7,	3	3,	2	1
P_1	5	6,	3	3,	3	3,	6	1,	6	3,	1	2
P_2	6	5,	4	3,	3	2,	5	6,	3	3,	1	2

For the haplotypic case, $EFI(m)$ was calculated by adding the informativeness on the female side of the pedigree (8) to the informativeness of the male side of the pedigree as follows (9):

$$EFI_{(\text{female})}(m) = 1 + p(A_1)\log_2[p(A_1)] + p(A_2)\log_2[p(A_2)], \quad (11)$$

$$EFI_{(\text{male})}(m) = 1 + p(A_3)\log_2[p(A_3)] + p(A_4)\log_2[p(A_4)], \quad (12)$$

$$EFI_{(\text{haplotypes})}(m) = EFI_{(\text{female})}(m) + EFI_{(\text{male})}(m), \quad (13)$$

where $p(A_i)$ is the probability of founder-origin for grandparent A_i. The $EFI_A(m)$ values were calculated at 1 cM intervals. For the zygotic case, the $EFI(m)$ values were calculated by the simultaneous use of the four founder-origin probabilities for A_1A_3, A_1A_4, A_2A_3, A_2A_4 (Table 1, CP), rather than summing haplotypic values.

The two QTL analysis approaches, zygotic and haplotypic, gave nearly identical $EFI_A(m)$ values (Fig. 4) except that the haplotypic method gave less information at the extreme ends of the simulated linkage group. *EFI* will decrease further for the haplotypic method if marker loci are constrained to the two-allele intercross configuration (Reyes-Valdés & Williams, 2002). The haplotypic and zygotic analyses are equal only if haplotypes for all individuals in the mapping population can be determined without ambiguity. However, for some haplotypes in this example, the founder origin of the first and last marker loci cannot be resolved, a situation caused by their grandparental and parental configuration rather than by their map position. In these cases, founder-origin probability relies on information from other marker loci.

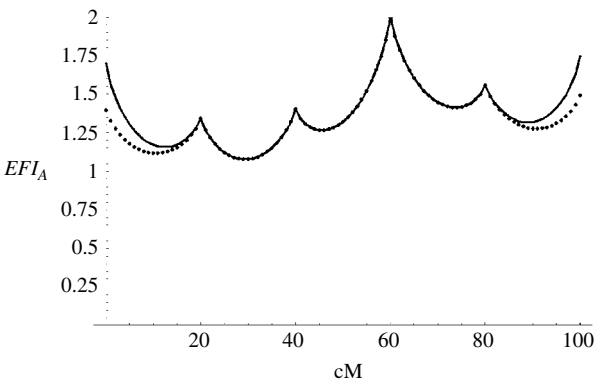

Fig. 4. Information content map for a 100 cM linkage group with six marker loci in a three-generation outbred pedigree. The continuous line represents the zygotic informativeness, whereas the dotted line represents the haplotypic informativeness.

3. Discussion

EFI, a new measure of informativeness for optimizing QTL mapping, can be generalized across marker types, map density and pedigree types and it has enhanced sensitivity for direct comparison between pedigree types. As such, *EFI* offers two advantages over previous measures, Ψ and R^2: (i) *EFI* is general for any number of founders from the single- to multiple-founder case and (ii) the maximum value of *EFI* is sensitive to the type of mapping population. Although R^2 applies to the multiple-founder case, it requires complex computations for large sibships.

The informativeness measure Ψ can be applied as a function of a single-founder probability but it has not been formally extended to several independent probabilities for multiple founders simultaneously. It was designed for outbred pedigrees and not originally intended for F_2 or recombinant inbred lines. Rather than extending to the multiple-founder case, the measure Ψ was originally offset by averaging the informativeness of the female side and the male side of the pedigree separately. This means that the measure Ψ is inherently haplotypic in nature and thus parallels *EFI* values only when used as a function of a single founder-origin probability, e.g. the BC_1 case. The measure Ψ(m), equivalent to the square root of R^2 for BC_1, closely parallels $EFI(m)$ as a measure of individual informativeness only in the single-founder case (Fig. 5). In the multiple-founder case, Ψ(m) remains haplotypic whereas *EFI* considers all zygotic information which carries equal or higher information content than haplotypic information. It is interesting to note that either *EFI*(m) and Ψ(m) can be used to calculate informativeness per individual

https://doi.org/10.1017/S0016672305007354 Published online by Cambridge University Press

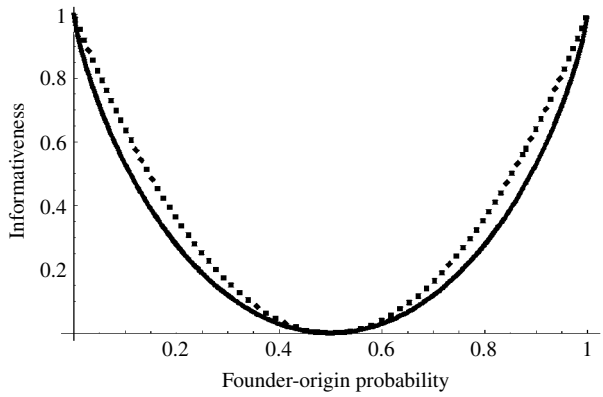

Fig. 5. Comparison of two measures for founder-origin informativeness applied to an individual progeny within a backcross (BC_1) mapping population. The dotted line represents the Knott *et al.* (1991) individual informativeness from $\Psi(m)$, whereas the continuous line represents the entropy-based founder informativeness, *EFI*(m).

(see equation 7 for *EFI* and the terms composing equation 2 for Ψ). However R^2 lacks a definition for individual measures of informativeness, instead using a sib-pair definition as a function of the variance of the *a posteriori* IBD distribution (Kruglyak & Lander, 1995).

EFI permits increased sensitivity for comparing information content among mapping population types. Both R^2 and Ψ are theoretically bounded by a maximum value of 1, independent of the type of mapping population. By comparison, the upper bound of *EFI*(m) is *Max*(H), where *Max*(H) depends on the type of mapping population, whether inbred- or outbred-derived pedigrees. This is a desirable property for any measure of map informativeness because higher values for *Max*(H) are directly related to higher informativeness for a given QTL mapping design. Sensitivity of *EFI*(m) to mapping populations is useful in determining the most appropriate population structure, either by comparing Max(H) or by using the theoretical distributions of founder-origin probabilities for the markers system to be applied.

As a cautionary note, any errors inherent to linkage map estimates will also be present in *EFI* as a measure of informativeness. A linkage map in practical terms is subject to sampling error so information content maps drawn from R^2, Ψ or *EFI* provide estimates prone to the same kind of error.

In summary, information theory and Shannon entropy are introduced to genetic and QTL mapping through the concept of entropy-based founder informativeness (*EFI*), a new measure of marker information content. *EFI* provides information content maps and measures global information from any genetic map for any type of mapping population with defined founders. *EFI* is general for the case of multiple founder-origin probabilities. A strong advantage of *EFI* is its enhanced sensitivity for comparing pedigree types such as direct comparison among types of inbred- or outbred-derived pedigrees. Several examples based on computer-simulated datasets showed the merit of *EFI* for evaluating optimal QTL mapping design.

Appendix 1. Relationship between the IBD R^2 and informativeness based on founder-origin probabilities for a backcross (BC_1) population

Let us assume a population generated by crossing two homozygous inbred lines A and B, and then their F_1 crossed to A. The lines A and B will be considered to be the founders of this population. Let x be the number of alleles IBD shared by two members of the BC_1 population. If both individuals bear alleles with same origin, i.e. both have AA or AB, then $x=2$. On the other hand, if one of the individuals has a set AA and the other AB, then $x=1$, with the IBD allele being shared coming from the recurrent parent. Before genotyping is performed, the *a priori* IBD distribution for the ith sib pair is $(\frac{1}{2}, \frac{1}{2})$. After genotyping has been performed, the *a posteriori* IBD distribution at point m is $[\pi_1(m), \pi_2(m)]$, with $\pi_1(m)$ and $\pi_2(m)$ being the probabilities of $x=1$ and $x=2$, respectively. Let $\sigma^2_{\text{initial}}$ be the expected variance of the *a priori* IBD distribution, and $\sigma^2_{\text{residual}}(m)$ be the expected variance of the *a posteriori* IBD distribution at point m. The R^2 for the predicted IBD distribution is a measure of marker informativeness at point m (Kruglyak & Lander, 1995) and its expected value at point m can be defined as:

$$R^2(m) = 1 - \frac{\sigma^2_{\text{residual}}(m)}{\sigma^2_{\text{initial}}}.$$

For a BC_1 population the expected $\sigma^2_{\text{initial}} = \frac{1}{4}$, because $E[x] = 3/2$, and $E[x^2] = 5/2$. Therefore $E[x^2] - (E[x])^2 = \frac{1}{4}$. Then, the expression for R^2 is:

$$R^2(m) = 1 - 4\sigma^2_{\text{residual}}(m).$$

For the sake of simplicity, to give an expression for $\sigma^2_{\text{residual}}(m)$ we will use now the variance of $x-1$, instead of x, having in mind that Variance(x) = Variance($x+k$), where k is any constant.

Let p and q be the probabilities, calculated with marker information, of two independent sibs being heterozygous for a putative QTL locus, respectively. They are equivalent to probabilities of a donor parent on a given map site in gametes coming from the F_1 to form the BC_1. Then, the probability of the two sibs being simultaneously heterozygous is pq, whereas the probability of two individuals being homozygous is $(1-p)(1-q)$. Hence the IBD distribution is

https://doi.org/10.1017/S0016672305007354 Published online by Cambridge University Press

$[p+q-2pq, pq+(1-p)(1-q)]$. The expected variance of $x-1$ is:

$$\sigma^2_{\text{residual}}(m) = E[pq+(1-p)(1-q)-(pq+(1-p)(1-q))^2].$$

Expanding this expression we obtain:

$$\sigma^2_{\text{residual}}(m) = E[p+q-p^2-q^2+4p^2q+4pq^2-4p^2q^2].$$

Now, the expectations for the terms inside the expression are: $\mathrm{E}[p]=\mathrm{E}[q]=\frac{1}{2}$; $\mathrm{E}[p^2]=\mathrm{E}[q^2]=\sigma_p^2+\frac{1}{4}$; $\mathrm{E}[4pq]=1$; $\mathrm{E}[4pq^2]=\mathrm{E}[4p^2q]=2\ \ \sigma_p^2+\frac{1}{2}$; $\mathrm{E}[4p^2q^2]=4(\sigma_p^2+\frac{1}{4})^2$. After substitution and simplification we obtain:

$$\sigma^2_{\text{residual}}(m) = \frac{1}{4} - 4\sigma_p^4.$$

Then the expected informativeness parameter at point m for a backcross population is:

$$R^2(m) = (4\sigma_p^2)^2,$$

where p is the probability of a given individual being heterozygous for a putative QTL allele or, in other words, the probability of genetic material coming from the recurrent founder of the backcross population through the F_1 at site m.

Appendix 2. Derivation of conditional entropy for the example of two biallelic, haplotypic loci

Let M and Q be two biallelic loci with a recombination coefficient r. Here we obtain an expression for $H(Q|M)$, i.e. the degree of uncertainty about the allele present in Q given a certain allele present at M in a testcross experiment. Alleles will be coded as (0, 1) in both loci, and they will be assumed to be in coupling phase. Symbols M and Q, used to name the marker and index locus, are also used here as random variables that can take the values 0 or 1 for allele coding.

The conditional uncertainty of Q given M is the average entropy $H(Q|M=x_i)$, where x_i is a given allele. In terms of the entropies of the *a priori* allele distributions at M and (M, Q) we have (Taneja, 2001):

$$H(Q|M) = H(M,Q) - H(M).$$

Now, the *a priori* distribution of the two alleles at M is $[\frac{1}{2}(0), \frac{1}{2}(1)]$, whose entropy is 1. Then:

$$H(Q|M) = H(M,Q) - 1.$$

Let $z(x, x')$ be the multivariate allele distribution at (M, Q). Thus

$$\begin{aligned} H(Q|M) = &-z(0,0)\log[z(0,0)] - z(0,1)\log[z(0,1)] \\ &-z(1,0)\log[z(1,0)] - z(1,1)\log[z(1,1)] - 1. \end{aligned}$$

With r being the coefficient of recombination between M and Q, the last expression becomes:

$$\begin{aligned} H(Q|M) &= -(1-r)\log\left[\frac{1}{2}(1-r)\right] - r\log\left[\frac{1}{2}r\right] - 1 \\ &= -(1-r)\log(1-r) - r\log(r) - \log\left(\frac{1}{2}\right) - 1 \\ &= -r\log(r) - (1-r)\log(1-r). \end{aligned}$$

We thank Dr David Gwaze for his assistance with exploratory computing at the start of this project. This work was sponsored by the CONACYT-Texas A&M University #3-050702-3, USDA-Forest Service #SRS 04-11333010 and Universidad Autónoma Agraria Antonio Narro #02·03·0203·2401.

References

Adami, C., Ofria, C. & Collier, T. C. (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences of the USA* **97**, 4463–4468.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci of linked factors. *Journal of Genetics* **8**, 299–309.

Haley, C. S., Knott, S. A. & Elsen, J. M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195–1207.

Knott, S. A., Neale, D. B., Sewell, M. M. & Haley, C. S. (1997). Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. *Theoretical and Applied Genetics* **94**, 810–820.

Kruglyak, L. & Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439–454.

Maliepaard, C., Jansen, J. & van Ooijen, J. W. (1997). Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. *Genetical Research* **70**, 237–250.

Martínez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

Nothnagel, M., Fürst, R. & Rhode, K. (2002). Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Human Heredity* **54**, 186–198.

Reyes-Valdés, M. H. & Williams, C. G. (2002). A haplotypic approach to founder-origin probabilities and outbred QTL analysis. *Genetical Research* **80**, 231–236.

Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics* **73**, 1402–1422.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423, 623–656.

Taneja, I. J. (2001). Generalized information measures and their applications. <http://www.mtm.ufsc.br:~taneja/book/book.html>.

Van Ooijen, J. W. (2004). MapQTL 5·0, Software for the mapping of quantitative trait loci in experimental populations. Wageningen, The Netherlands: Kyazma.

Yockey, H. (1992). *Information Theory and Molecular Biology*. Cambridge: Cambridge University Press.