Journal: HDY Disk used Despatch Date: 10/4/2007
Article : npg_hdy_6800979 Pages: 1–13 OP: KGU ED: POULY

Heredity (2007) , 1–13

www.nature.com/hdy

ORIGINAL ARTICLE

Allelic penetrance approach as a tool to model two-locus interaction in complex binary traits

N Sepúlveda[1], CD Paulino[2], J Carneiro[1] and C Penha-Gonçalves[1]
[1]*Instituto Gulbenkian de Ciência, Oeiras, Portugal and* [2]*Department of Mathematics, Instituto Superior Técnico, Lisboa, Portugal*

Many binary phenotypes do not follow a classical Mendelian inheritance pattern. Interaction between genetic and environmental factors is thought to contribute to the incomplete penetrance phenomena often observed in these complex binary traits. Several two-locus models for penetrance have been proposed to aid the genetic dissection of binary traits. Such models assume linear genetic effects of both loci in different mathematical scales of penetrance, resembling the analytical framework of quantitative traits. However, changes in phenotypic scale are difficult to envisage in binary traits and limited genetic interpretation is extractable from current modeling of penetrance. To overcome this limitation, we derived an allelic penetrance approach that partitioned incomplete penetrance into the alleles controlling the phenotype and into the genetic background and environmental factors. We applied this approach to formulate dominance and recessiveness in a single biallelic locus and to model different genetic mechanisms for the joint action of two biallelic loci. We fit the models to data on the susceptibility of mice following infections with *Listeria monocytogenes* and *Plasmodium berghei*. These models gain in genetic interpretation, because they specify the alleles that are responsible for the genetic (inter)action and their genetic nature (dominant or recessive), and predict genotypic combinations determining the phenotype. Further, we show via computer simulations that the proposed models produce penetrance patterns not captured by traditional two-locus models. This approach provides a new analysis framework for dissecting mechanisms of interlocus joint action in binary traits using genetic crosses.
Heredity (2007) **0,** 000–000. doi:10.1038/sj.hdy.6800979

Keywords: reduced penetrance; allelic penetrance; external penetrance; epistasis; heterogeneity; allelic liability

Introduction

In quantitative genetics, the joint action of two loci is usually assessed by models assuming linear genetic effects on a given mathematical scale of the quantitative trait (for example, Fisher, 1918; Cockerham, 1954; Cheverud and Routman, 1995; Kao and Zeng, 2002; Zeng *et al*., 2005). If models do not agree with experimental data, one may say that an interaction – epistasis – is present. The same reasoning is followed when analyzing penetrance – defined as the probability of a phenotype given a genotype – in the study of complex binary traits (for example, Hodge, 1981; Risch, 1990; Risch *et al*., 1993; Vieland and Huang, 2003). Although the choice of scale and linearity can be supported by some biological mechanisms (Omholt *et al*., 2000), there is no 'natural' scale for penetrance in which current models have a specific biological or causal interpretation (Cordell *et al*., 2001). Moreover, the assumption of linearity on penetrance is more of a statistical convenience rather than a true genetic description. Therefore, clear genetic information cannot be extracted from current two-locus models for complex binary traits.

Correspondence: C Penha-Gonçalves, Instituto Gulbenkian de Ciência, Apartado 14, P-2781-901 Oeiras, Portugal.
E-mail:cpenha@igc.gulbenkian.pt
Received 4 April 2006; revised 26 September 2006; accepted 16 February 2007

A study on the genetic interaction of Idd3 and Idd10 loci in the control of diabetes in mice thoroughly discusses the difficulties of interpreting the traditional two-locus models for penetrance (Cordell *et al*., 2001). Bagot *et al*. (2002) experienced the same difficulties in a study of the interaction between Berr1 and Berr2 loci in the resistance to experimental cerebral malaria in mice. Despite showing that some current models could fit the data, both Cordell *et al*. (2001) and Bagot *et al*. (2002) were unable to unravel which kind of genetic interaction actually underlies the traits.

This paper aims to improve interpretability of current two-locus models for complex binary traits. For this purpose, we developed an allelic penetrance approach that models dominance and recessiveness for a single diallelic locus. Then, the framework is extended to the two-locus case to describe three different genetic (inter)-action mechanisms: (i) genetic heterogeneity, in which an individual manifests the phenotype by possessing a predisposing genotype at either locus; (ii) inhibition action, whereby an allele of a given locus prevents an allele of another locus from manifesting its effect; (iii) allelic liability, in which the inheritance of the phenotype is controlled by a given number of certain alleles at the combined genotype of the two loci. These allelic penetrance models have improved genetic interpretation, because they specify at each locus which allele is responsible for the inheritance of the phenotype (hereafter referred to as phenotype-conferring alleles) and their genetic behavior (dominant or recessive). This

information is instrumental in designing the experiments that may confirm suggested genetic (inter)actions.

We illustrate the advantage of using the allelic penetrance models with data from an intercross and a backcross experiment on the genetic control of susceptibility of mice following infections with *Listeria monocytogenes* (Boyartchuk *et al*., 2001) and *Plasmodium berghei* (Bagot *et al*., 2002), respectively. We further show by simulation that allelic penetrance models can generate data that cannot be fitted by current models, and therefore cover at least a different range of genetic interaction.

Statistical modeling

Allelic penetrance approach

The inheritance of complex binary traits often shows reduced penetrance. That is, individuals bearing the same genotype can manifest either presence or absence of the phenotype of interest. The classical interpretation for this phenomenon postulates the existence of other genetic and environmental factors that can modify the action of the phenotype-conferring alleles (Nadeau, 2001). However, experimental genetics is rich in examples of reduced penetrance in pure lines that are maintained under controlled laboratory conditions designed to minimize environmental variation. One example is the nonobese diabetes mouse strain, which spontaneously develops autoimmune diabetes, but with reduced penetrance (reviewed, for example, in Anderson and Bluestone, 2005). Another example is given by Lalucque and Silar (2004) that show that loss-of-function mutations of two K^+ transporters in the euascomycete fungus *Podospora anserina* exhibit reduced penetrance of crippled growth, even when genetic and environmental effects were virtually eliminated. These observations suggest that somehow the genotype has an intrinsic stochastic property of being expressed at the level of the phenotype. In fact, Rakyan *et al*. (2002) suggested that reduced penetrance can actually be attributed to a stochastic expression of the alleles themselves. In this regard, these authors introduced the concept of metastable epialleles when studying the coat color of mice controlled by the agouti locus. An epiallele is an alllele that can stably exist in more than one epigenetic state, resulting in different phenotypes. A metastable epiallele is an epiallele at which the epigenetic state can switch and its establishment is a probabilistic event. Once established, the state is mitotically inherited. Therefore, there is a need to account for probabilistic allelic effects in current genetic models.

We propose an allelic penetrance approach to model penetrance of a single diallelic locus. Following the above observations, penetrance is decomposed in a sum of two components: an *internal component* attributable to the probability of the alleles of the genotype expressing the phenotype and an *external component* pertaining to the probability of the phenotype being affected by (genetic and/or environmental) factors other than the locus under study. The probability of an allele expressing the phenotype is hereafter referred to as allelic penetrance.

Consider a diallelic locus A with a dominant allele A over an allele a with respect to the phenotype. Let π_g be the penetrance of genotype $g = AA$, Aa, aa, respectively. Denote the penetrances of alleles A and a by π_A and π_a, respectively. Since allele A is dominant, the phenotype is determined by the expression of at least one allele A. Assuming independent allelic expressions, dominance is equivalent to an independent action of the alleles A towards the expression of the phenotype. As defined above, the internal component of penetrance is given by

$$\pi_g^{\text{int}} = \begin{cases} \pi_A^2 + 2\pi_A(1-\pi_A) & \text{if } g = AA \\ \pi_A & \text{if } g = Aa \\ 0 & \text{if } g = aa \end{cases} \quad (1)$$

Two comments can be made regarding the above equation: (i) π_a does not model the internal component of penetrance, because the recessive allele a cannot contribute to the expression of phenotype; (ii) in the homozygous genotype AA, both alleles A can be expressed, and thus monoallelic expression phenomena are excluded from the present modeling.

In this framework, we also include the action of factors external to the locus under study. This was based on observations that disease phenotypes are expressed in absence of disease-conferring genotypes at a given locus. These observations suggest a phenocopy mechanism by which genetic or nongenetic factors are expressed when phenotype-conferring alleles are not present. Thus, we postulate that the effect of external factors is only relevant for penetrance when phenotype-conferring alleles, in this case alleles A, are not expressed. This implies that external factors do not affect the allelic penetrance of the phenotype-conferring allele and are not accountable when calculating internal penetrance.

In this line of thought, the external component of penetrance refers to the probability of the action of external factors to the locus in promoting the phenotype. It is described mathematically by the product of the probability of having no expression of alleles A towards the phenotype $(1-\pi_g^{\text{int}})$ with the probability of the external factors favoring the phenotype of interest (π_{ext})

$$\pi_g^{\text{ext}} = (1-\pi_g^{\text{int}})\pi_{\text{ext}} \begin{cases} (1-\pi_A)^2\pi_{\text{ext}} & \text{if } g = AA \\ (1-\pi_A)\pi_{\text{ext}} & \text{if } g = Aa \\ \pi_{\text{ext}} & \text{if } g = aa \end{cases} \quad (2)$$

The final expression of the penetrance is then the sum of internal and external components, that is,

$$\begin{aligned} \pi_g &= \pi_g^{\text{int}} + \pi_g^{\text{ext}} \\ &= \begin{cases} \pi_A^2 + 2\pi_A(1-\pi_A) + (1-\pi_A)^2\pi_{\text{ext}} & \text{if } g = AA \\ \pi_A + (1-\pi_A)\pi_{\text{ext}} & \text{if } g = Aa \\ \pi_{\text{ext}} & \text{if } g = aa \end{cases} \end{aligned} \quad (3)$$

Consider now that the phenotype is controlled by the expression of the recessive allele a, and is inhibited by the expression of the dominant allele A. In this situation, the phenotype is observed only when the dominant allele A is not being expressed and at least one allele a is being expressed. In this line of thought, the internal component of penetrance becomes

$$\pi_g^{\text{int}} = \begin{cases} 0 & \text{if } g = AA \\ \pi_a(1-\pi_A) & \text{if } g = Aa \\ \pi_a^2 + 2\pi_a(1-\pi_a) & \text{if } g = aa \end{cases} \quad (4)$$

Following similar rationale as described for the dom-

inance situation, the external component of penetrance refers to the probability of having phenotypic expression of external factors when the phenotype is not caused by the expression of alleles *a*. Thus, we have

$$\pi_g^{\text{ext}} = (1-\pi_g^{\text{int}})\pi_{\text{ext}} = \begin{cases} \pi_{\text{ext}} & \text{if } g = AA \\ [1-\pi_a(1-\pi_a)]\pi_{\text{ext}} & \text{if } g = Aa \\ (1-\pi_a)^2\pi_{\text{ext}} & \text{if } g = aa \end{cases} \quad (5)$$

Finally, summing (4) and (5) the penetrance is given by

$$\pi_g = \pi_g^{\text{int}} + \pi_g^{\text{ext}} = \begin{cases} \pi_{\text{ext}} & \text{if } g = AA \\ \pi_a(1-\pi_A) + [1-\pi_a(1-\pi_A)]\pi_{\text{ext}} & \text{if } g = Aa \\ \pi_a^2 + 2\pi_A(1-\pi_a) + (1-\pi_a)^2\pi_{\text{ext}} & \text{if } g = aa \end{cases} \quad (6)$$

It is worth noting that recessiveness is modeled by as many parameters as genotypes, and thus it can be regarded as a default model for the action of a single locus. Moreover, the above equation can describe dominance when $\pi_A = 0$ (compare with Equation (3)). Therefore dominance is a special case of recessiveness.

The comparison of Equations (3) and (6) shows two important features of the modeling. One feature is that when a locus is dominant or recessive, the penetrance of the heterozygous genotype is not equal to the penetrance of either homozygous genotype, except when the allelic penetrance of one of the alleles is equal to 1. This is in contrast to the current modeling assumption that penetrance in the heterozygous genotype is equal to one of the homozygous genotypes (Vieland and Huang, 2003). Moreover, penetrance patterns of dominance and recessiveness are derived under the allelic penetrance approach, while in current models are assumed but not derivable. Interestingly, when both alleles have complete penetrance in absence of external influence ($\pi_{\text{ext}} = 0$), Equations (3) and (6) give rise to traditional Mendelian dominant and recessive patterns of inheritance. Therefore, the allelic penetrance approach can be seen as a generalization of Mendelian inheritance for reduced penetrance scenarios. Another feature is that Equations (3) and (6) are not complementary. This is explained by two reasons: (i) the absence of expression of a dominant allele does not necessarily imply the expression of the recessive allele, because both alleles have allelic penetrances; (ii) since the external factors are only relevant in the absence of the expression of the phenotype-conferring alleles, different conditions are found for the action of external factors with respect to the dominant or recessive phenotype-conferring allele. Thus, under the allelic penetrance approach, dominance and recessiveness are not symmetrical concepts.

Two-locus models based on allelic penetrance

Several two-locus models for penetrance have been discussed in the past (reviewed in Cordell *et al*., 2001). They can be generically written as

$$g(\pi_{g_Ag_B}) = \alpha_{g_A} + \beta_{g_B} \quad (7)$$

where $\pi_{g_Ag_B}$ is the penetrance of genotype g_A (*AA*, *Aa* and *aa*) of locus A and of genotype (*BB*, *Bb* and *bb*) of locus B, $g(\cdot)$ is a function that defines each model as follows: (i) identity for the additive model (Risch, 1990); (ii) logarithm for the multiplicative model (Hodge, 1981); (iii) complementary logarithm for the heterogeneity model (Risch, 1990; Vieland and Huang, 2003); (iv) logarithm of the odds for the logistic model (Baxter, 2001; Stewart, 2002); and (v) cumulative distribution of a standard normal variable for the liability model (for example, Pearson, 1900; Dempster and Lerner, 1950; Risch *et al*., 1993). Some authors declare dominance and/or recessiveness at each locus if $\alpha_{AA} = \alpha_{Aa} = \alpha_A$ and $\beta_{BB} = \beta_{Bb} = \beta_B$ (Strauch *et al*., 2003; Vieland and Huang, 2003).

However, in this situation, one cannot distinguish whether the phenotype is produced by the recessive or by the dominant allele at each locus. Therefore, the above models cannot *per se* identify either the phenotype-conferring alleles or their genetic nature (dominant or recessive).

To overcome this problem, we use the allelic penetrance approach in alternative to classical two-locus models. To this end, we extended the decomposition of penetrance for the two-locus case, that is,

$$\pi_{g_Ag_B} = \pi_{g_Ag_B}^{\text{int}} + \pi_{g_Ag_B}^{\text{ext}} \quad (8)$$

where $\pi_{g_Ag_B}^{\text{int}}$ and $\pi_{g_Ag_B}^{\text{ext}}$ are the internal and the external components of penetrance for the combined genotype (g_A, g_B), respectively. As stated in the single locus case, external factors are only relevant when the alleles of the two interacting loci are not expressing the phenotype. Therefore, the external component of penetrance can be factorized as

$$\pi_{g_Ag_B}^{\text{ext}} = (1-\pi_{g_Ag_B}^{\text{int}})\pi_{\text{ext}} \quad (9)$$

when substituted in Equation (8) leads to the following general formula of two-locus penetrance

$$\pi_{g_Ag_B} = \pi_{g_Ag_B}^{\text{int}} + (1-\pi_{g_Ag_B}^{\text{int}})\pi_{\text{ext}} \quad (10)$$

Different genetic interaction mechanisms can be considered by specifying the internal penetrance as follows.

Independent action models (IAMs): The independent action models (IAMs) are based on genetic heterogeneity, in which the expression of two loci are independent causes of the phenotype. We consider that each locus has a phenotype-conferring allele, which can be either dominant or recessive. Thus, there are four types of IAMs according to the genetic behavior of the phenotype-conferring alleles at each locus: dominant–dominant, dominant–recessive, recessive–dominant and recessive–recessive.

Derivation of the penetrances according to IAM is almost straightforward. The internal component of penetrance is simply defined by the probabilities of expression of the phenotype-conferring alleles at either locus toward the phenotype of interest. Since heterogeneity means independent action of both loci, the internal component of penetrance satisfies the probabilistic relationship for the union of two independent events, each one referring to the allelic expression of each locus, that is,

$$\pi_{g_Ag_B}^{\text{int}} = \phi_{g_A} + \phi_{g_B} - \phi_{g_A}\phi_{g_B} \quad (11)$$

where ϕ_{g_A} and ϕ_{g_B} are the probabilities of expression of the phenotype-conferring alleles at genotypes g_A and g_B,

Table 1 Penetrance tables illustrating independent action models (IAMs), inhibition models (IMs) and cumulative action models (CAMs) when the alleles have complete penetrance (that is, equal to 1) in the absence of external factors ($p_{ext}=0$)

Genotypes	*Models*					
	IAM(A/B)[a]	*IAM(A/b)*[b]	*IM(A/B)*[c]	*IM(A/b)*[d]	*CAM$_2$(A/B)*[e]	*CAM$_3$(A/B)*[f]
AA						
BB	1	1	0	1	1	1
Bb	1	1	0	1	1	1
bb	1	1	1	0	1	0
Aa						
BB	1	1	0	1	1	1
Bb	1	1	0	1	1	0
bb	1	1	1	0	0	0
aa						
BB	1	0	0	0	1	0
Bb	1	0	0	0	0	0
bb	0	1	0	0	0	0

[a]IAM with dominant phenotype-conferring alleles *A* and *B* at loci A and B, respectively.
[b]Same as in [a] but with a recessive allele *b* at locus B.
[c]IM with a dominant phenotype-conferring allele *A* at locus A and a dominant inhibiting allele *B* at locus B.
[d]Same as in [c] but with a recessive inhibiting allele *b* at locus B.
[e]CAM requiring the expression of at least two phenotype-conferring alleles *A* and *B* at loci A and B, respectively.
[f]Same as in [e], but requiring at least three phenotype-conferring alleles *A* and *B*.

respectively. If the phenotype-conferring allele at one locus is dominant, the corresponding ϕ_{g_i} then follows Equation (1). Analogously, if the phenotype-conferring allele is recessive, the respective ϕ_{g_i} is then determined by Equation (4). Finally, external factors are included in the model through (10) with $\pi^{int}_{g_Ag_B}$ determined by (11). Thus, in addition to external penetrance π_{ext}, IAMs are parameterized as follows: by π_A and π_B for IAM dominant–dominant; by π_A, π_B and π_b for IAM dominant–recessive; by π_A, π_a and π_B for IAM recessive–dominant; and by π_A, π_a, π_B and π_b for IAM recessive–recessive. Following the comments about recessiveness and dominance given in the single-locus modeling, IAM recessive–recessive can be converted either into an IAM dominant–recessive or in an IAM recessive–dominant through $\pi_A=0$ and $\pi_B=0$, respectively. Moreover, the IAM recessive–dominant and IAM dominant–recessive are IAM dominant–dominant if $\pi_A=0$ and $\pi_B=0$. Therefore, the IAM shows a nested structure.

The mathematical formulation of IAM is also able to generate deterministic inheritance patterns (corresponding to complete penetrance scenarios), when allelic penetrances of phenotype-conferring alleles at either locus are made equal to one and the external penetrance equal to 0. Table 1 illustrates two examples of IAMs under this condition, where the combined genotypes of the two loci determining phenotype inheritance become easily identifiable. For instance, in the IAM with a dominant-conferring phenotype allele *A* at locus A and a recessive phenotype-conferring allele *b* at locus B, the phenotype is inherited when an individual has either the homozygous or the heterozygous genotype *AA* and *Aa* at locus A or has the homozygous genotype *BB* at locus B. Thus, IAM can account simultaneously for reduced and complete penetrance scenarios. It is worth noting that classical two-locus models do not possess this feature as they cannot account for deterministic situations.

Inhibition models (IMs): Bateson (1907, 1909) described a phenomenon termed epistasis whereby an allele of an *epistatic locus* prevents an allele of a *hypostatic locus* from manifesting its effect. Since then, epistasis has been used in many different contexts, often with conflicting meanings, that led to several discussions about its formal definition (Philips, 1998; Cordell, 2002; Strauch *et al*., 2003; Moore and Williams, 2005). As a consequence, many authors proposed different models to detect epistatic effects: additive models (Risch, 1990), multiplicative models (Hodge, 1981), and heterogeneity models (Risch, 1990; Vieland and Huang, 2003). Failure to fit the models is often claimed to be evidence of epistatic interaction and the same can be said regarding failure to fit the independent action models described above.

Here, we recovered the original Bateson's definition of epistasis to develop inhibition models (IMs). In these models, a locus confers the phenotype through the expression of the respective phenotype-conferring allele and the other locus inhibits the phenotypic expression of the former by its inhibiting allele. The alternative alleles have no conferring or inhibiting action on the phenotype. Phenotype-conferring and -inhibiting alleles can be considered either dominant or recessive.

Consider that locus A confers the phenotype of interest and locus B inhibits the expression of the former. In this case, the internal component of penetrance relates to the probability of the phenotype-conferring alleles at locus A being expressed when the phenotype-inhibiting alleles at locus B are not expressing their inhibiting action. Thus, the internal component of penetrance satisfies

$$\pi^{int}_{g_Ag_B} = \phi_{g_A}(1-\phi^*_{g_B}) \quad (12)$$

where ϕ_{g_A} is the probability of genotype g_A expressing the phenotype of interest and $\phi^*_{g_B}$ is the probability of genotype g_B having an inhibitory behavior. Dominance

and recessiveness are included in the model by replacing ϕ_{g_A} and $\phi^*_{g_B}$ by the single-locus internal penetrances (1) and (4), respectively. Finally, we include the action of external factors in penetrance through Equation (10) with internal component of penetrance given by the above equation.

IMs have different numbers of parameters depending on the dominance and recessiveness nature of the phenotype-conferring and -inhibiting alleles. For example, IM with a dominant phenotype-conferring allele at locus A and a recessive inhibiting allele *b* at locus B is parameterized by π_A, π_B, π_b and π_{ext}. Parameterization of the remaining models follows the same reasoning. Like in IAMs, the IMs also show a nested structure, where IMs with some (phenotype-conferring or -inhibiting) dominant alleles are nested in IMs with both recessive alleles, and IMs with both dominant alleles are nested in IMs with some dominant alleles.

Comparing Equations (11) and (12), it is easy to see that IM and IAM produce distinct penetrance patterns. This is better illustrated in complete penetrance scenarios when the allelic penetrances are equal to one and the external penetrance is equal to zero (see examples in Table 1). In this table, the phenotype in a IM with a dominant phenotype-conferring allele *A* at locus A and a dominant inhibiting allele *B* at locus B is inherited when an individual has either the homozygous or the heterozygous genotype *AA* and *Aa* at locus A (genotypes that confer the phenotype) and the homozygous genotype *bb* at locus B (genotype that does not induce inhibition).

Cumulative models: Some models for penetrance are based on a latent Gaussian quantitative trait, the so-called liability (Falconer, 1965). One approach advocates that a fixed threshold exists in the liability that divides individuals with or without the phenotype (Pearson, 1900; Wright, 1937; Dempster and Lerner, 1950; Risch *et al*., 1993). Another approach is to say that each individual has its own liability threshold, which follows a Gaussian distribution in the population (Curnow, 1972; Curnow and Smith, 1975). One can also use logistic regression models, which approximate the Gaussian liability distribution by the (standard) logistic distribution (Baxter, 2001). However, most geneticists use these models in the same way as epidemiologists use them to estimate risk factors for a disease in a given population (for example, Cordell, 2002; Whittemore and Halpern, 2003; North *et al*., 2005).

Hagen and Gilbertson (1973) suggest that the inheritance of 'lateral plate morphs' in freshwater threespine sticklebacks is determined by the number of certain alleles at the combined genotype of two loci. In the same line of thought, Stewart (2002) hypothesized that the inheritance of multifactorial diseases is controlled by a given number of disease-conferring alleles in the genotype of an individual. In view of these ideas, we propose the cumulative action models (CAMs) where the expression of the phenotype is controlled not only by the presence of a certain number of phenotype-conferring alleles in the combined genotype of the two loci, but also by their expression. Note that dominance and recessiveness are not included in the model, because what matters here is the number of the phenotype-conferring alleles when expressed. This implies that alleles that are not conferring the phenotype are considered to have a null action for the expression of the phenotype.

Let x_i represent the number of phenotype-conferring alleles in the genotype of locus $i = A, B$. Let also Y_i be the random variable that indicates the number of those alleles expressing the phenotype at locus $i = A, B$. According to the allelic penetrance approach, $Y_i \mid x_i$ has a binomial distribution with x_i trials and probability of success given by the allelic penetrance π_i of the phenotype-conferring allele at locus $i = A, B$. Assuming independence between Y_A and Y_B, the probability mass function of the total number Y of phenotype-conferring alleles expressing the phenotype given by the combined genotype (x_A, x_B) is determined by

$$\begin{aligned} P[Y = & y|x_A, x_B] \\ &= \sum_{l=0}^{x_A} P[Y_A = l|x_A, x_B] \\ &\quad P[Y_B = y - l|x_A, x_B] \end{aligned} \tag{13}$$

where

$$P[Y_i = y_i|x_A, x_B] = \binom{x_i}{y_i} \pi_i^{yi} (1 - \pi_i)^{x_i - y_i} \tag{14}$$

Consequently, CAM entails the following internal component of penetrance

$$\begin{aligned} \pi^{\text{int}}_{x_A x_B} &= P[Y \geqslant k|x_B, x_B] \\ &= \sum_{y=k}^{x_B + x_B} P[Y = y|x_A, x_B],\ k = 1, \ldots, 4 \end{aligned} \tag{15}$$

where $P[Y = y \mid x_A, x_B]$ is given by (13) with $k \leqslant x_A + x_B$. As in IAM, the effect of the external factors is included in the model by (10) with $\pi^{\text{int}}_{g_A g_B}$ as calculated by (15). Thus, CAMs have three parameters: the two allelic penetrances of the phenotype-conferring alleles at each locus and external penetrance π_{ext}. In contrast with IAM and IM, CAM have no nested structure, and thus they have to be fitted accordingly.

It is worth noting that the assumed independence between Y_A and Y_B implies that CAM is conceptually akin to independent action of the two loci at the level of the allelic expression. However, one cannot say that the action of the two loci is independent at the level of the phenotype. Take for example the CAM requiring the simultaneous expression of at least three phenotype-conferring alleles at the combined genotype of the two loci. In spite of the allelic expression being independent, the action of a single locus is insufficient to confer the phenotype, as opposed to IAM. The phenotype is only observed when the two loci act together.

As in previous models, when the alleles have complete penetrance in absence of external factors, CAM can also describe specific complete penetrance scenarios (see examples in Table 1). For instance, in the CAM requiring the simultaneous expression of at least two phenotype-conferring alleles *A* and *B* at loci A and B at the combined genotype, the phenotype is inherited when an individual has either the homozygous genotypes *AA* or *BB* at each locus or has the combined genotype *Aa*/*Bb*. However, when the allelic threshold of CAM is 1 or 4, the correspondent complete penetrance tables are coincident to an IAM with dominant phenotype-conferring alleles at both loci and to IM with a recessive phenotype-

conferring allele at one locus and a dominant inhibiting allele at the other locus, respectively. This coincidence has been pointed out by Li and Reich (2000), who showed that different genetic interaction mechanisms can be specified by the same requisite for the inheritance of phenotype.

Statistical inference

Some statistical considerations have to be put forward when fitting the above models to experimental data. We assume that the likelihood of the data is

$$\mathcal{L} = \prod_{g_A, g_B} \left(n_{g_A g_B} n_{g_A g_{B^1}} \right) (\pi_{g_A g_B})^{n_{g_A g_{B^1}}} (1 - \pi_{g_A g_B})^{n_{g_A g_{B^2}}} \qquad (16)$$

where $n_{g_A g_{Bk}}$ is the number of sampled individuals with combined genotype (g_A, g_B) and phenotype $k = 1,2$, $n_{g_A g_B} = n_{g_A g_{B^1}} + n_{g_A g_{B^2}}$, and $\pi_{g_A g_B}$ is the penetrance of combined genotype (g_A, g_B). The (log-)likelihood is then maximized with respect to the allelic and external penetrances. At this point, it is worth noting that one cannot know whether a particular allele is being expressed in each individual, and thus data is incomplete (or missing) under the allelic penetrance models. In this case, the maximization of the likelihood is easily achieved via the Expectation–Maximization algorithm that is particularly suitable for missing data problems (Dempster *et al*., 1977). The algorithm is described in the appendix and was implemented in R language (Ihaka and Gentleman, 1996).

We perform the traditional Wilks' likelihood ratio test to evaluate the goodness –of fit of the models. The test is based on the following test statistic

$$Q_v = -2 \sum_{g_A, g_B} \sum_{k=1,2} n_{g_A g_{Bk}} (\ln \hat{\mu}_{g_A g_{Bk}} - \ln n_{g_A g_{Bk}})$$

where $\hat{\mu}_{g_A g_{Bk}}$ is the expected value of $n_{g_A g_{Bk}}$ under the model being tested. Standard statistical theory predicts that, for large samples, the distribution of Q_v under the null hypothesis follows a χ^2_p where p is the difference between the number of combined genotypes in the data and the parameter number of the model under test. The level of significance was setup at 5%.

As alluded previously, the IAM and IM have nested structures. In this situation, one may first evaluate the goodness –of fit of the most general IAM and IM, which are the ones with recessive allelic action at each locus. Then, if these models agree with the data, they can be compared with models specifying some dominant allelic action. In general, one can compare models M_1 and M_2 with $M_1 \subset M_2$ using the Wilks' likelihood ratio conditional statistic

$$Q_v(M_1|M_2) = -2 \sum_{g_A, g_{Bk}} n_{g_A g_{Bk}} (\ln \tilde{u}_{g_A g_{Bk}} - \ln \hat{\mu}_{g_A g_{Bk}})$$

where $\tilde{u}_{ij\ k}$ an1d $\hat{\mu}_{g_A g_{Bk}}$ are the expected values of $n_{g_A g_{Bk}}$ under models M_1 and M_2, respectively. Generically, the distribution of $Q_v(M_1 \mid M_2)$ for large samples is under the null hypothesis χ^2_p, where p is the diference between the number of parameters of M_1 and M_2. However, in the IAM and IM, M_1 is obtained from M_2 via an allelic penetrance equal to 0. For example, testing IAM dominance–recessive against IAM recessive–recessive is equivalent to test the penetrance of the dominant allele at locus A equal to 0. Since an allelic penetrance is a probability, testing a probability equal to 0 falls into a situation where the χ^2 approximation to the Wilks' test statistic may not be accurate (Self and Liang, 1987). A good approximation is to cut in half the P-value obtained from a χ^2_1 (Self and Liang, 1987). CAMs are not nested and they are fitted using the above-described unconditional testing. On the other hand, the mathematical structures of the three classes of models do not allow for fitting comparison by classical likelihood ratio test.

In practical terms, several models may fit the data well, and thus one should have good criteria to help to decide which genetic joint actions should be evaluated experimentally. To do this, we devised a measure to compare competing genetic models based on the decomposition of penetrance shown in Equation (10). For a given model, the idea is to compare the internal component of penetrance with total penetrance. One way to do it is to consider the ratio between the penetrance of the population attributable to the internal component of penetrance (that is, attributable to joint action of the two loci itself) and the penetrance of the population explained by both internal and external components of penetrance, that is,

$$\mathrm{IPC} = \frac{\sum_{g_A, g_B} \pi^{\mathrm{int}}_{g_A g_B} f_{g_A g_B}}{\sum_{g_A g_B} (\pi^{\mathrm{int}}_{g_A g_B} + (1 - \pi^{\mathrm{int}}_{g_A g_B}) \pi^{\mathrm{ext}}) f_{g_A, g_B}} \qquad (17)$$

where $f_{g_A g_B}$ is the frequency of the combined genotype (g_A, g_B) in the population. High IPC values mean that the observed penetrance is well explained by the joint action of the two loci, while low IPC values show that external factors are the main contribution for the observed penetrance. In this line of thought, if one is interested in the two-locus joint action *per se*, one should confirm experimentally the models exhibiting the highest IPC values.

Examples

The following examples refer to data derived from experimental crosses between a susceptible and a resistance strain. Our analysis used the following notation for the allelic penetrance models. There are two loci A and B with alleles a_1 and a_2 and with alleles b_1 and b_2, respectively, where the alleles a_1 and b_1 are derived from the susceptible strain and the alleles a_2 and b_2 are derived from the resistance strain. In the models, we use when necessary upper and lower cases to represent dominant and recessive alleles, respectively. IAM(A_1/b_2) represents an independent action model with phenotypeconferring alleles A_1 and b_2 at each locus. IM (A^i_1/B^c_2) is an IM with a phenotype-conferring allele B_2 and a -inhibiting allele A_1, where the superscripts c and i denote a phenotype-conferring and a -inhibiting action, respectively. CAM$_k$ (a_2/b_1) is a CAM requiring jointly the expression of at least k phenotype-conferring alleles a_2 and b_1.

Application to intercross data

Boyartchuk *et al*. (2001) reported the genetic mapping of susceptibility to infection by *Listeria monocytogenes* in mice. These authors performed an intercross experiment between the susceptible strain BALB/cByJ and the resistant strain C57BL/6. Two loci at chromosomes 5 and 13 (here referred as A and B, respectively) were identified to be strongly associated with susceptibility.

The observed phenotype data for each genotype combination is shown in Table 2. Boyartchuk *et al*. (2001) pointed out that the alleles a_2 and b_1 seem to contribute to susceptibility. This observation suggests that these two alleles might be the phenotype-conferring alleles for susceptibility, and a_1 or b_2 the inhibiting alleles. In contrast, the phenotype-conferring alleles for resistance might be a_1 and b_2, and a_2 or b_1 the inhibiting alleles.

The allelic penetrance approach is designed to study reduced penetrance phenotypes. As both resistance and susceptibility showed complete pene trance in parental strain (Boyartchuk *et al*., 2001), there was no clear choice for the phenotype of interest. In this case, we decided to separately analyze susceptibility and resistance, and look for the most suitable models (see Table 3).

With respect to resistance, we first fitted IAM(a_1/b_2). The *P*-value of the goodness-of-fit test for this model was 0.03, and thus this model does not fit the data well. However, estimates of π_{b_1} π_b and π_{ext} are close to 0, a situation where the *P*-value may not be accurate (Self and Liang, 1987). To overcome this problem, we simulated 1000 data sets under the fitted model, calculating in each one the corresponding likelihood ratio statistic. An empirical *P*-value is given by the proportion of tests that accepted the model. In this way, we obtained an empirical *P*-value of 0.07. In spite of being accepted at the 5% level of significance, we rejected this model, because the quality of the fit is far from being satisfactory. Owing to nested structure of IAM, IAM with some dominant phenotype-conferring or -inhibiting alleles were not tested.

Next, we fitted IM (a_1^{ς}/b_1^{i}) and IM (a_2^{i}/b_2^{ς}). The goodness-of-fit tests for these models revealed that both models do not describe the data well ($P=0.02$ for IM (a_1^{ς}/b_1^{i}) and $P=0.01$ for IM (a_2^{i}/b_2^{ς})). Once again IMs with some dominant alleles were not fitted owing the same reason as in IAM. Therefore, there is no evidence for an inhibition action between the two loci. Finally, we evaluated the fit of different CAM with phenotype-conferring alleles a_1 and b_2. The goodness-of-fit tests show that CAM do not capture reasonably the observed penetrance pattern ($P<0.05$). Thus, none of the proposed models was able to fit reasonably the penetrance pattern of resistance.

In relation with susceptibility, we fitted IAM(a_2/b_1) as opposed to the IAM(a_1/b_2) fitted to resistance. This model shows a reasonable fit to data (*P*-value = 0.13). This was also confirmed by an empirical *P*-value of 0.14 obtained as described above for resistance. Conditionally to IAM(a_2/b_1), we asked whether the alleles a_2 or b_1 were dominant rather than recessive. The *P*-values of goodness-of-fit tests for IAM(A_2/b_1) and IAM(a_2/B_1) show evidence for the first model and against the second. Then, conditionally to IAM(A_2/b_1), we evaluate the hypothesis of the allele b_1 being dominant. The *P*-value of the goodness-of-fit test disfavors this hypothesis ($P<10^{-3}$). Therefore, there is evidence for IAM(A_2/b_1). The internal component of penetrance in this model accounts for 90% of the total penetrance, additionally supporting the existence of other 'minor' loci in the genetic background. In contrast to IM specified for

Table 2 Data of Listeria experiment regarding loci A and B at chromosomes 5 and 13, respectively, where a_1 and b_1 denote the alleles derived from the susceptible strain, whereas a_2 and b_2 represent the alleles derived from the resistant strain

Genotypes		*Susceptibility*	*Resistance*	*Penetrance*
Locus A	*Locus B*			
a_1a_1	b_1b_1	4	3	0.57
a_1a_1	b_1b_2	4	15	0.21
a_1a_1	b_2b_2	1	11	0.08
a_1a_2	b_1b_1	23	1	0.95
a_1a_2	b_1b_2	10	21	0.32
a_1a_2	b_2b_2	6	7	0.46
a_2a_2	b_1b_1	10	0	1.00
a_2a_2	b_1b_2	8	4	0.66
a_2a_2	b_2b_2	5	4	0.55

Penetrance refers to susceptibility.

Table 3 Likelihood ratio (Q_v) tests and maximum likelihood parameter estimates for the allelic penetrance models to the Listeria data set

Phenotype	*Model*	Q_v	*d.f.*	P-*value*	$\hat{\pi}_{a_1}$	$\hat{\pi}_{a_2}$	$\hat{\pi}_{b_1}$	$\hat{\pi}_{b_2}$	$\hat{\pi}_{ext}$
Resistance	IAM(a_1/b_2)	10.75	4	0.03	0.34	0.79	0.00	0.44	0.00
	IM(a_1/b_1)	17.15	4	0.002	0.61	0.04	0.98	0.93	0.17
	IM(b_2/a_2)	11.39	4	0.02	0.22	0.38	0.00	0.78	0.08
	CAM1(a_1/b_2)	16.03	6	0.01	0.21	—	—	0.43	0.00
	CAM_2(a_1/b_2)	15.10	6	0.02	0.71	—	—	0.63	0.12
	CAM_3(a_1/b_2)	36.83	6	$<10^{-3}$	1.00	—	—	0.62	0.35
	CAM_4(a_1/b_2)	49.48	6	$<10^{-3}$	0.96	—	—	0.96	0.44
Susceptibility	IAM(a_2/b_1)	7.72	4	0.13	0.22	0.35	0.89	0.60	0.11
	IAM(A_2/b_1)[a]	0.91	—	0.17	—	0.32	0.60	0.89	0.10
	IAM(a_2/B_1)[a]	16.93	—	$<10^{-3}$	0.15	0.37	0.32	—	0.15
	IAM(A_2/B_1)[b]	16.73	—	$<10^{-3}$	—	0.34	0.32	—	0.05
	IM($a_2/b2$)	14.65	4	0.005	0.07	1.00	0.00	0.57	0.18
	IM(b_1/a_1)	13.99	4	0.007	0.41	0.85	0.99	0.77	0.24
	CAM1(a_2/b_1)	24.80	6	$<10^{-3}$	—	0.34	0.32	—	0.05
	CAM_2(a_2/b_1)	15.69	6	0.02	—	0.33	0.89	—	0.24
	CAM_3(a_2/b_1)	13.01	6	0.04	—	0.85	1.00	—	0.33
	CAM_4(a_2/b_1)	46.98	6	$<10^{-3}$	—	1.00	1.00	—	0.48

Abbreviations: CAM, cumulative action model; d.f., degrees of freedom; IAM, independent action model; IM, inhibition model.
[a]Conditional tests given IAM(a_2/b_1).
[b]Conditional test given IAM(A_2/b_1).

resistance, we fitted IM(a_2^c/b_2^c) and IM(a_1^i/b_1^c). Both models do not fit the data (P-values <0.01), and thus as in resistance there is no inhibition action between the two loci for susceptibility. Finally, CAM with phenotype-conferring alleles A_2 and B_1 cannot also fit the data well. With these results, we conclude that susceptibility might be controlled by an independent action of a dominant allele derived from the resistant strain at locus A and a recessive allele derived from the susceptible strain at locus B.

We also fitted the traditional two-locus models: additive, multiplicative, heterogeneity, liability and logistic models. The results revealed that all models could describe the data (data not shown). One usually interprets these results as lack-of-interaction between the loci. Thus, a classical heterogeneity model holds where the two loci seem to act independent of each other, in close agreement with the best-fitted allelic penetrance model IAM(A_2/b_1). This shows that the IAM added information to the genetic action in determining the alleles causing the phenotype as well as their genetic nature. Finally, it is worth noting that the likelihoods of current and allelic penetrance models are not comparable, because the parameter number among these models is different.

Application to backcross data

Bagot *et al*. (2002) performed a genetic mapping study in mice for the susceptibility to experimental cerebral malaria (ECM) following *Plasmodium berghei ANKA* infection. A cross was performed between the ECM resistant strain WLA and the ECM susceptible strain C57BL/6J, where F_1 progeny was ECM resistant and was backcrossed with the susceptible parental strain.

Two disease-associated loci in different autosomes (Berr1 and Berr2) were detected using standard genetic mapping tools. The observed phenotypic frequencies for each genotype combination of Berr1 and Berr2 are shown in Table 4. Since the F_1 progeny was backcrossed with the susceptible parental strain, we chose susceptibility as the phenotype of interest susceptibility.

Data from backcross experiment have only four combined genotypes of the two loci. Thus, the variants of independent action and IMs that include recessive alleles have more parameters than the number of genotypic combination. In this case, the model is not estimable (see Discussion). This is a consequence of modeling internal penetrance of recessive alleles at each locus by two allelic penetrances, one for the dominant allele and another for the recessive allele (see Equation (4)). Since the WLA strain was observed to be 100% resistant to the disease, while the other strain was not fully susceptible (Bagot *et al*., 2002), we could avoid overparameterization by assuming that the internal component of penetrance of a recessive phenotype-conferring (or inhibiting) allele follows Equation (4) with $\pi_A = 1$. In this case penetrances for all models add up to three parameters: one allelic penetrance for each locus and the external penetrance π_{ext}. With this assumption, it is easy to see IAM and IM will not have a nested structure, and thus we evaluated the fit of these models using unconditional Wilks' likelihood ratio tests.

A careful observation of Table 4 shows that penetrance increases with the number of alleles derived from the susceptible strain. This suggests that phenotype-conferring alleles should be a_1 and b_1, while inhibiting alleles should be a_2 and b_2. The goodness-of-fit tests revealed that IAM(a_1/b_1), CAM$_2$(A_1/B_1) and CAM$_3$(A_1/B_1) provide the best fit for the ECM data (Table 5); other models were fitted but exhibited lack of fit at the level of significance of 5% (data not shown). The first model suggests an independent action of two recessive alleles of the susceptible strain, while the other two support a cooperative action of both loci, requiring simultaneous expression of at least two or three conferring alleles of the susceptible strain at the combined genotype. With respect to IPC (see last row of Table 5), the internal component of penetrance of IAM(a_1/b_1) and CAM$_3$(a_1/b_1) explains around 80% of the total penetrance, whereas in the case of CAM$_2$(A_1/B_1) it explains 100% of the total penetrance. However, these estimates might be inflated, because many genotypic combinations are missing from the data by experimental design.

Table 4 Phenotypic data of experimental cerebral malaria backcross experiment for loci Berr1 and Berr2, where a_1 and b_1 denote the alleles derived from the susceptible strain, whereas a_2 and b_2 represent the alleles derived from the resistant strain

Genotypes		*Susceptibility*	*Resistance*	*Penetrance*
Berr1	*Berr2*			
a_1a_1	b_1b_1	35	10	0.78
a_1a_1	b_1b_2	25	23	0.56
a_1a_2	b_1b_1	27	21	0.52
a_1a_2	b_1b_2	9	40	0.18

Penetrance refers to susceptibility.

Table 5 Likelihood ratio (Q_v) tests and maximum likelihood parameter estimates for the best-fitted allelic penetrance models to the experimental cerebral malaria data set

Fitted penetrances	*Best-fitted models*		
	IAM(a_1/b_1)	*CAM*$_2$(a_1/b_1)	*CAM*$_3$(a_1/b_1)
a_1a_1/a_2a_2	0.76	0.73	0.79
a_1a_1/b_2a_2	0.53	0.49	0.51
a_1a_2/a_2a_2	0.58	0.58	0.55
a_1a_2/b_2a_2	0.17	0.27	0.19
$\hat{\pi}_{a_1}$	0.66	0.44	0.78
$\hat{\pi}_{b_1}$	0.70	0.61	0.89
π_{ext}	0.18	0.00	0.17
Q_v	0.161	2.720	0.087
P-value (1 d.f.)	0.688	0.099	0.769
IPC	0.79	1.00	0.77

Abbreviations: d.f., degrees of freedom.

Simulations

The previous section illustrated the application of the allelic penetrance models to intercross and backcross data. An important issue is to assess whether the traditional models would capture the penetrance patterns generated by the allelic penetrance models. To address this issue, we performed a small simulation study. We generated 250 F_2 individuals from an intercross experiment, where two loci A and B located in

different autossomes (with independent Mendelian segregation) interact according to the proposed models. Then, we fitted the traditional models (additive, multiplicative, heterogeneity, liability and logistic) to 1000 data sets simulated from the proposed models with a given parameter set. We then computed the proportion of times that the Wilks' likelihood ratio test provided evidence for these models. The level of significance of each test was setup at 5%. The implementation of the simulation procedure was performed in R language (Ihaka and Gentleman, 1996) and was tested by fitting the model to the data it generated (see the control acceptance ratio in Figure 1).

In the simulation, we assume that the internal component of penetrance of a recessive allele is given by Equation (4) with $\pi_A = 1$, while the internal component of penetrance of a dominant allele follows Equation (1). Thus, all allelic penetrance models have only three parameters: one allelic penetrance for each locus plus the external penetrance. We studied the situation where the allelic penetrances of each locus are equal and $\pi_{ext} = 0.10$ and 0.20. Figure 1 shows the results of varying the allelic penetrance for some of the proposed models. The simulations for the remaining models show similar patterns and are available from the authors upon request.

One general observation is that the ability of the traditional models to fit the simulated data decreases as the allelic penetrance increases. In one extreme, the traditional models can fit simulated data well from low allelic penetrances, because the penetrances of the proposed models are mostly controlled by π_{ext}, and thus they can be described by a linear model including only a global effect (equal to π_{ext}). In the other extreme, the traditional models cannot reproduce simulated data from high allelic penetrances, because such data are close to a deterministic situation, and therefore any (statistical) linear model is expected to fail in those cases.

The other general observation is that the liability and logistic models are quite similar in terms of fitting the simulated data. This is in close agreement with Chambers and Cox (1967), who reported that these two models can only be distinguished for large sample sizes. Furthermore, these two models seem to be the ones that best fit the data simulated from CAM. Since CAMs are based on a latent trait representing the number of phenotype-conferring alleles being expressed, we speculate that the discrete distribution of this trait may be reasonably approximated by the Normal or the Logistic distribution on which the liability and logistic models are based.

The heterogeneity and multiplicative models are the ones that best reproduce data generated from IAM and IM, respectively. This can be explained by the fact that the heterogeneity and the multiplicative model are described by $\pi_{ij} = \alpha_i + \beta_j - \alpha_i\beta_j$ and $\pi_{ij} = \alpha_i\beta_j$, respectively, which resemble the internal penetrances of the above-mentioned proposed models (see Equations (11) and (12), respectively).

In summary, the traditional models do not have the full ability to describe data generated from the proposed models, particularly when the two interacting loci have moderate to high effects on penetrance (corresponding to moderate to high allelic penetrances). Therefore, in addition to the possibility of describing reasonable genetic interaction mechanisms, the proposed models produce penetrance patterns not covered by the traditional two-locus models, showing their potential usefulness in exploring genetic interaction in experimental data.

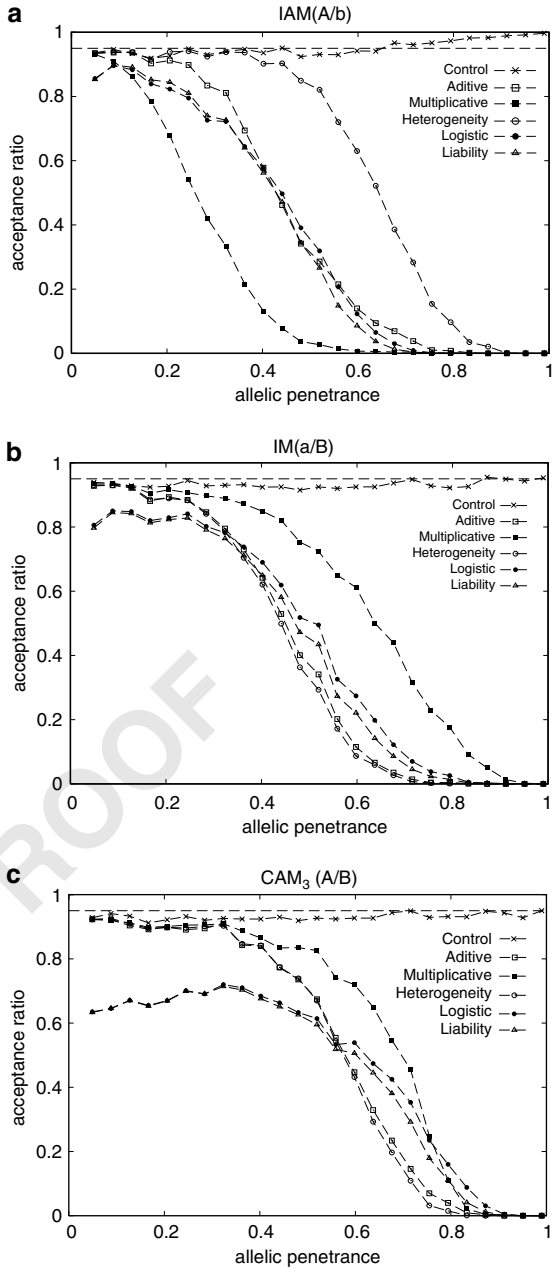

Figure 1 Examples of simulations results: estimated proportion of the Wilks' likelihood ratio tests accepting the traditional models for data simulated by three allelic penetrance models. Each data point represents estimations for 1000 datasets at given allelic penetrance with external penetrance equal to 0.10. The control line represents the fitting of the model generating the data.

Discussion

An allelic penetrance approach was proposed to model genetic interactions in complex binary traits. The whole framework is based on a decomposition of penetrance in a sum of an internal component pertaining to the

expression of the alleles of the genotypes and an external component attributable to other factors acting on the phenotype. The modeling presupposes the same 'mean effect' of the genetic background and/or environmental factors on penetrance of all genotypes. This assumption seems only reasonable when the genetic background is not homogeneous, such as in the second-generation crosses analyzed. Thus, the modeling of external component needs to be refined to describe better the effect of different genetic backgrounds or environmental factors on penetrance. It would be particularly interesting to apply the proposed approach to congenic mice strains and human data. Refinement of the external component will also be useful to include data concerning parental strains and F_1 generation in the analysis of genetic interaction. Interesting guidelines to deal with genetic backgrounds are provided by Hansen and Wagner (2001).

We proposed two-locus allelic penetrance models based on genetic heterogeneity, Bateson's epistasis and allelic liability. It is worth noting that these three classes of genetic joint actions do not cover all possible two-locus interactions. In fact, Li and Reich (2000) alluded to other types of interactions, namely, jointly dominant/recessive models, modifying effect and interference models. We did not consider them in this paper, because they are rarely reported in the literature.

We showed that the proposed modeling is sensitive to the choice of the phenotype of interest. In the cerebral malaria example, we chose susceptibility as the phenotype of interest, because, the F_1 generation was backcrossed with the susceptible strain to increase disease incidence. In the case of intercrosses, the *a priori* choice of the phenotype may not be so straightforward as illustrated by the Listeria data set. In general, if one is studying a binary trait using wild-type versus mutant strains, the concept of canalization may be used to identify the phenotype of interest as a deviation of the wild-type phenotype. However, in cases of phenotypes generated by exogeneous stimuli, such as resistance or susceptibility to infections, it could be difficult to define *a priori* the phenotype of interest. In such cases, we suggest to analyze separately resistance and susceptibility and search the models that best fit the data.

We illustrated the fitting of the proposed models in two data sets from mouse experiments. In both examples, the two-locus joint action of the best-fitted models explains *per se* 80–100% of total penetrance. Therefore, the external factors have a minor contribution to the fitting.

In the Listeria example, susceptibility appears to be mediated by an independent action with a dominant allele of the susceptible strain in the chromosome 5 locus and a recessive allele of the resistant strain in the chromosome 13 locus. To validate the model, one can breed two single congenic mice, one with the allele at chromosome 5 locus from the resistant strain being bred in the susceptible strain and the other with the allele at chromosome 13 locus from susceptible strain being bred in the resistant strain. These congenic strains should be more susceptible to Listeria infection than the respective parental strain. Boyartchuk *et al*. (2001) showed that the C57BL/6 strain was fully resistant to a Listeria infection. This observation may be explained by a putative phenotype-inhibiting role of the genetic background in the resistant parental strain, which is not included in the proposed modeling.

In the experimental cerebral malaria example, our results indicate that susceptibility seems to be controlled by an independent action of recessive alleles derived from the susceptible strain at Berr1 and Berr2. They also suggest that the phenotype may be inherited by a genetic action requiring the simultaneous expression of at least two or three alleles of the susceptible strain at the combined genotype of the two loci. Bagot *et al*. (2002) show that F_1 generation mice were all resistant to the disease. This result seems to rule out the CAM requiring jointly at least two phenotype-conferring alleles being expressed, whereas it favors the other two models. To validate the models, our results predict that single-locus congenic mice in a resistance background should be susceptible or remain resistant to the disease if the two-locus interaction follows either IAM or CAM_3, respectively.

One limitation of the allelic penetrance models relates with the fitting of backcross data: some models show more parameters than the data allow. As a consequence, parameters cannot be uniquely determined and estimated. In statistical terms, the models are said to be nonidentifiable, and thus nonestimable (Paulino and Pereira, 1994). In the cerebral malaria example, this problem was overcome by attributing complete allelic penetrance to the nonconferring alleles, because the susceptible parental strain exhibited reduced penetrance for susceptibility, while its resistant parental strain showed complete penetrance. Therefore, usage of the models in backcross data is possible when the phenotype of interest shows reduced penetrance in the respective parental strain whereas the absence of it has complete penetrance in the other strain. However, to avoid this kind of assumptions, we will study in the future in which situations the models are identifiable.

The standard models for genetic interaction are commonly used to detect the existence of some kind of epistasis, but they cannot infer its nature (Cordell, 2002). One advantage of the models presented here is the possibility of identifying the nature of the genetic interaction, even in the absence of prior evidence for interaction. This suggests a two-step procedure: (i) fit classical two-locus models, and if they do not fit the data, (ii) use models proposed here to infer plausible mechanisms of epistasis. Notwithstanding, our models may also be useful in cases where epistasis is absent. In fact, when the traditional heterogeneity model holds, our independent action models can infer which are the phenotype-conferring alleles at each locus as well as their genetic behavior. This is illustrated by the Listeria and cerebral malaria examples discussed above, for which there is evidence for the classical heterogeneity model (data not shown).

The simulation study corroborated the notion that the proposed models produce penetrance patterns not captured by the traditional two-locus models, especially when the two loci have a moderate to strong impact on penetrance. In conclusion, the allelic penetrance models produce genetic interpretations of real data and can simulate data that is not explained by any of the traditional models of binary traits. Therefore, these models may prove useful in revealing genetic interactions that may have been hitherto undetectable.

Acknowledgements

The authors are grateful to Dan Holmberg's laboratory, specially to Susana Campino from the Instituto Gulbenkian de Ciência (IGC), and Victor Boyartchuk from the Department of Genetics, Howard Hughes Medical Institute, for kindly providing the data from cerebral malaria and Listeria experiments, respectively. The authors thank Henrique Teotónio, Tiago Paixão and Rui Gardner from the IGC for valuable discussions. Nuno Sepúlveda acknowledges financial support from Fundação Calouste Gulbenkian, and Fundação para a Ciência e Tecnologia (fellowship SFRH/BD/19810/2004).

References

Anderson MS, Bluestone JA (2005). The NOD mouse: a model of immune dysregulation. *Annu Rev Immunol* **23**: 447–485.

Bagot S, Campino S, Penha-Gonçalves C, Pied S, Cazenave PA, Holmberg D (2002). Identification of two cerebral malaria resistance loci using an inbred wild-derived mouse strain. *Proc Natl Acad Sci USA* **99**: 9919–9923.

Bateson W (1907). Facts limiting the theory of heredity. *Science* **26**: 649–660.

Bateson W (1909). *Mendel's Principles of Heredity.* Cambridge University Press: Cambridge.

Baxter AG (2001). Modelling the effects of genetic and environmental factors on the risk of autoimmune disease. *J Autoimmun* **16**: 331–335.

Boyartchuk VL, Broman KW, Mosher RE, D'Orazio SE, Starnbach MN, Dietrich WF (2001). Multigenic control of *Listeria monocytogenes* susceptibility in mice. *Nat Genet* **27**: 259–260.

Chambers EA, Cox DR (1967). Discrimination between alternative binary responses. *Biometrika* **54**: 573–578.

Cheverud JM, Routman EJ (1995). Epistasis and its contribution to genetic variance components. *Genetics* **139**: 1455–1461.

Cockerham CC (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**: 859–882.

Cordell H (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* **11**: 2463–2468.

Cordell HJ, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB *et al*. (2001). Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type I diabetes. *Genetics* **158**: 357–367.

Curnow RN (1972). The multifactorial model for the inheritance of liability to disease and its implications for relatives at risk. *Biometrics* **28**: 931–946.

Curnow RN, Smith C (1975). Multifactorial models for familial diseases in man (with discussion). *J Roy Stat Soc Ser A* **138**: 131–169.

Dempster A, Laird N, Rubin D (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc Ser B* **39**: 1–38.

Dempster ER, Lerner IM (1950). Heritability of threshold characters. *Genetics* **35**: 212–236.

Falconer DS (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet* **29**: 51–76.

Fisher RA (1918). The correlations between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* **52**: 399–433.

Hansen TF, Wagner GP (2001). Modeling genetic architecture: a multilinear theory of gene interaction. *Theor Popul Biol* **59**: 61–86.

Hagen DW, Gilbertson LG (1973). The genetics of plate morphs in freshwater threespine sticklebacks. *Heredity* **31**: 75–84.

Hodge S (1981). Some epistatic two-locus models of disease. i. relative risks and identity-by-descent distributions in affected sib pairs. *Am J Hum Genet* **33**: 381–395.

Ihaka R, Gentleman R (1996). R: a language for data analysis and graphics. *J Comp Graph Stat* **5**: 299–314.

Kao CH, Zeng ZB (2002). Modeling epistasis of quantitative trait locus using Cockerham's model. *Genetics* **160**: 1243–1261.

Lalucque H, Silar P (2004). Incomplete penetrance and variable expressivity of a growth defect as a consequence of knocking out two K^+ transporters in the euascomycete fungus *Podospora anserina*. *Genetics* **166**: 125–133.

Li W, Reich J (2000). A complete enumeration and classification of twolocus disease models. *Hum Hered* **50**: 334–349.

Moore JH, Williams SM (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* **27**: 637–646.

Nadeau JH (2001). Modifier genes in mice and humans. *Nat Rev Genet* **2**: 165–174.

North BV, Curtis D, Sham PC (2005). Application of logistic regression to case–control association studies involving two causative loci. *Hum Hered* **59**: 79–87.

Omholt SW, Plahte E, Oyehaug L, Xiang K (2000). Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* **155**: 969–980.

Pearson K (1900). Mathematical contributions to the theory of evolution. VIII. On the inheritance of characters not capable of exact quantitative measurement. *Philos Trans R Soc Lond Ser A* **195**: 79–121.

Paulino CD, Pereira CAB (1994). On identifiability of parametric statistical models. *J Ital Stat Soc* **3**: 125–151.

Philips PC (1998). The language of gene interaction. *Genetics* **149**: 1167–1171.

Rakyan V, Blewitt M, Druker R, Preis J, Whitelaw E (2002). Metastable epialleles in mammals. *Trends Genet* **18**: 348–351.

Risch N (1990). Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* **46**: 222–228.

Risch N, Ghosh S, Todd JA (1993). Statistical evaluation of multiple locus linkage data in experimental species and relevance to human studies: application to murine and human IDDM. *Am J Hum Genet* **53**: 702–714.

Self S, Liang KY (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* **82**: 605–610.

Stewart J (2002). Towards the genetic analysis of multifactorial diseases: the estimation of allele frequency and epistasis. *Hum Hered* **54**: 118–131.

Strauch K, Fimmers R, Baur MP, Wienker TF (2003). How to model a complex trait. II. Analysis with two disease loci. *Hum Hered* **56**: 200–211.

Vieland V, Huang J (2003). Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected-sib-pair data. *Am J Hum Genet* **73**: 223–232.

Whittemore AS, Halpern J (2003). Logistic regression of family data from retrospective study designs. *Genet Epidemiol* **25**: 177–189.

Wright S (1937). An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics* **19**: 506–536.

Zeng ZB, Wang T, Zou W (2005). Modeling quantitative trait loci and interpretation of models. *Genetics* **169**: 1711–1725.

Appendix A

EM Algorithm

Here, we provide an EM algorithm to estimate the different allelic penetrance models by maximum likelihood. As a simplification, we consider models where the internal penetrance of a locus is modeled by a single parameter (that is, Equation (1) for dominance and Equation (4) with $\pi_A = 1$ for recessiveness). In this context, all models have three parameters: π_{ext} and two

allelic penetrances for each locus, that will be denoted hereafter by π_A and π_B. The general case follows the same reasoning as described here and is available from the authors.

The EM algorithm was originally developed to facilitate the maximum likelihood estimation in missing data problems (Dempster *et al*., 1977). This algorithm was also found useful for cases where real data is not missing, but are conceptually incomplete, as in the case under the allelic penetrance models. Indeed, the observed data have limited information about the allelic expression events occurred in the construction of the phenotype of each individual.

To simplify the derivations, we denote the phenotype by $z=0$, 1 and the combined genotype of an individual by xy, where x and y represent the number of phenotype-conferring alleles in the genotype of locus A and B, respectively ($x=0$, 1, 2 and $y=0$, 1, 2 for an intercross, and $x=1$, 2 and $y=1$, 2 for a backcross). Following this notation, the observed data are the frequencies n_{xyz} of individuals with genotype xy and phenotype z. The frequency of individuals with genotype xy is represented by $n_{xy}=n_{xy0}+n_{xy1}$.

Recall the decomposition of penetrance given in Equation (10). The internal penetrance of the allelic penetrance models is based on the following random variables: $A^*\,|\,xy$ and $B^*\,|\,xy$ are independently distributed as $Bin(x, \pi_A)$ and $Bin(y, \pi_B)$, representing the number of alleles being expressed at locus A and B, respectively. Therefore, the joint probability of $(A^*, B^*)=(v, w)$ is given by

$$\theta_{xyvw} = P[(A^*, B^*) = (v, w)|xy] = \binom{x}{v}\pi_A^x(1-\pi_A)^{x-v}\binom{y}{w}\pi_B^w(1-\pi_B)^{y-w} \quad (18)$$

In this context, the internal component of penetrance is

$$\pi_{xy}^{\text{int}} = \sum_{(v,w)\in\mathcal{P}xy}\theta_{xyvw}$$

where P_{xy} is the set of values (v, w) given xy that produce the phenotype of interest (see Table 6 for the definition of P_{xy} for the different two-locus interaction models). The external penetrance is modeled by a Bernoulli trial E that indicates whether the external factors are expressing the phenotype of interest given that the phenotype-conferring allelic expression is absent. That is,

$$P[E=e|(v,w)\notin\mathcal{P}_{xy}] = \pi_{\text{ext}}^e(1-\pi_{\text{ext}})^{1-e}$$

Thus, we have

$$\pi_{xy}^{\text{ext}} = \pi_{\text{ext}}\sum_{(v,w)\notin\mathcal{P}_{xy}}\theta_{xyvw}$$

If one could have the information of $A^*\,|\,xy$, $B^*\,|\,xy$ and $E\,|\,(v,w)\notin P_{xy}$ for all the individuals, the complete data would be the vector of frequencies $\{n^*_{xyvw}, n^+_{xyvwe}\}$, where n^*_{xyvw} is the frequency of individuals with $(v, w)\notin P_{xy}$, and n^+_{xyvw} is the frequency of individuals with $(v, w)\notin P_{xy}$, and $E=e$. Its sampling distribution (or the likelihood function) would then be a multinomial-product distribution, that is, one multinomial distribution for each combined genotype xy

Table 6 Definition of $P_{xy}=\{(v, w)\in\{0,\ldots, x\}\times\{0,\ldots, y\}$: conditions$\}$ for the different allelic penetrance models

Model	*Conditions*
IAM(*A*/*B*)	$v\geqslant 1\vee w\geqslant 1$
IAM(*A*/*b*)	$v\geqslant 1\vee(y=2\wedge w\geqslant 1)$
IAM(*a*/*B*)	$(x=2\wedge v\geqslant 1)\vee w\geqslant 1$
IAM(*a*/*b*)	$(x=2\wedge v\geqslant 1)\vee(x=2\wedge w\geqslant 1)$
IM(*A*/*B*)	$v\geqslant 1\wedge w=0$
IM(*A*/*b*)	$v\geqslant 1\wedge[y\leqslant 2\vee(y=2\wedge w=0)]$
IM(*a*/*B*)	$(x=2\wedge v\geqslant 1)\wedge w=0$
IM(*a*/*b*)	$(x=2\wedge v\geqslant 1)\wedge[y<2\vee(y=2\wedge w=0)]$
CAM_1(*A*/*B*)	$v+w\geqslant 1$
CAM_2(*A*/*b*)	$v+w\geqslant 2$
CAM_3(*a*/*B*)	$v+w\geqslant 3$
CAM_4(*a*/*b*)	$v+w=4$

$$\mathcal{L} = \prod_{x,y} n_{xy}! \prod_{(v,w)\in\mathcal{P}_{xy}}\frac{\theta_{xyvw}^{n^*_{xyvw}}}{n^*_{xyvw}!} \prod_{(v,w)\notin\mathcal{P}_{xy}}\frac{(\theta_{xyvw}\pi_{\text{ext}})^{n^*_{xyvw1}}[\theta_{xyvw}(1-\pi_{\text{ext}})]^{n^*_{xyvw0}}}{n^+_{xyvw1}!n^+_{xyvw0}!}$$

where θ_{xyvw} is given by Equation (18). One can easily prove that the maximum likelihood estimators of π_A, π_B and π_{ext} are

$$\hat{\pi}_A = \frac{\sum_{x,y}\left[\sum_{(v,w)\in\mathcal{P}_{xy}} vn^*_{xyvw}+\sum_{(v,w)\notin\mathcal{P}_{xy}} v(n^+_{xyvw1}+n^+_{xyvw0})\right]}{\sum_{x,y} xn_{xy}}$$

$$\hat{\pi}_B = \frac{\sum_{x,y}\left[\sum_{(v,w)\in\mathcal{P}_{xy}} wn^*_{xyvw}+\sum_{(v,w)\notin\mathcal{P}_{xy}} w(n^+_{xyvw1}+n^+_{xyvw0})\right]}{\sum_{x,y} yn_{xy}} \quad (19)$$

$$\hat{\pi}_{\text{ext}} = \frac{\sum_{x,y}\sum_{(v,w)\notin\mathcal{P}_{xy}} n^+_{xyvw1}}{\sum_{x,y}\sum_{(v,w)\notin\mathcal{P}_{xy}}(n^+_{xyvw1}+n^+_{xyvw0})}$$

The complete data relate to the observed data as follows

$$n_{xy1} = \sum_{(v,w)\in\mathcal{P}_{xy}} n^*_{xyvw}+\sum_{(v,w)\notin\mathcal{P}_{xy}} n^+_{xyvw1}$$

$$n_{xy0} = \sum_{(v,w)\notin\mathcal{P}_{xy}} n^+_{xyvw0}$$

Thus, the E-step of the *k-th* iteration refers to the calculation of the expected values of $\{n^*_{xyvw}, n^+_{xyvwe}\}$, conditional to $\{n_{xyz}\}$. It is easy to verify that, for all $(v,w)\in P_{xy}$

$$m^{+(k)}_{xyvw1} = E[n^*_{xyvw}|\{n_{xyz}\}; \pi_A^{(k-1)}, \pi_B^{(k-1)}, \pi_{\text{ext}}^{(k-1)}] = n_{xy1}\frac{\theta_{xyvw}^{(k-1)}}{\sum_{(v,w)\in\mathcal{P}xy}\theta_{xyvw}^{(k-1)}+\sum_{(v,w)\notin\mathcal{P}xy}\theta_{xyvw}^{(k-1)}\pi_{\text{ext}}^{(k-1)}}$$

and, for all (v,w)$\notin P_{xy}$

$$m^{+(k)}_{xyvw1} = E[n^{+}_{xyvw1} | \{n_{xyz}\}; \pi_A^{(k-1)}, \pi_B^{(k-1)}, \pi_{\text{ext}}^{(k-1)}]$$
$$= n_{xy1} \frac{\theta^{(k-1)}_{xyvw} \pi_{\text{ext}}^{(k-1)}}{\sum_{(v,w)\in \mathcal{P}xy} \theta^{(k-1)}_{xyvw} + \sum_{(v,w)\notin \mathcal{P}xy} \theta^{(k-1)}_{xyvw} \pi_{\text{ext}}^{(k-1)}}$$

$$m^{+(k)}_{xyvw0} = E[n^{+}_{xyvw0} | \{n_{xyz}\}; \pi_A^{(k-1)}, \pi_B^{(k-1)}, \pi_{\text{ext}}^{(k-1)}]$$
$$= n_{xy0} \frac{\theta^{(k-1)}_{xyvw} (1 - \pi_{\text{ext}}^{(k-1)})}{\sum_{(v,w)\in \mathcal{P}xy} \theta^{(k-1)}_{xyvw} (1 - \pi_{\text{ext}}^{(k-1)})}$$

where $\theta^{(k-1)}_{xyvw}$ is given by (18) with π_A and π_B replaced by $\pi_A^{(k-1)}$ and $\pi_B^{(k-1)}$, respectively. Then, in the M-step of the kth iteration, $\pi_A^{(k)}$, $\pi_B^{(k)}$ and $\pi_{\text{ext}}^{(k)}$ follow Equation (19) with n^{*}_{xyvw}, n^{+}_{xyvw0} and n^{+}_{xyvw1} replaced by $m^{*(k)}_{xyvw}$, $m^{+(k)}_{xyvw0}$ and $m^{+(k)}_{xyvw1}$, respectively. The E- and M-steps are alternated repeatedly until the difference between log-likelihood functions of two consecutive iterations changes by an arbitrary small amount (that is, 10^{-6}).