



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

## Computational Statistics and Data Analysis

journal homepage: [www.elsevier.com/locate/csda](http://www.elsevier.com/locate/csda)

## Bayesian analysis of allelic penetrance models for complex binary traits

Nuno Sepúlveda<sup>a,b,c,\*</sup>, Carlos Daniel Paulino<sup>c,d</sup>, Carlos Penha-Gonçalves<sup>a</sup><sup>a</sup> Instituto Gulbenkian de Ciência, Portugal<sup>b</sup> Escola Superior de Saúde Egas Moniz, Portugal<sup>c</sup> Center of Statistics and Applications, University of Lisbon, Portugal<sup>d</sup> Department of Mathematics, Instituto Superior Técnico, Portugal

## ARTICLE INFO

## Article history:

Received 19 May 2008

Received in revised form 24 October 2008

Accepted 30 October 2008

Available online 6 November 2008

## ABSTRACT

Complex binary traits result from an intricate network of genetic and environmental factors. To aid their genetic dissection, several generalized linear models have been described to detect interaction between genes. However, it is recognized that these models have limited genetic interpretation. To overcome this problem, the allelic penetrance approach was proposed to model the action of a dominant or a recessive allele at a single locus, and to describe two-locus independent, inhibition, and cumulative actions. Classically, a recessive inheritance requires the expression of both recessive alleles in homozygotes to obtain the phenotype (type I recessiveness). In previous work, recessiveness was defined alternatively as a situation where a recessive allele is able to express the phenotype when the dominant allele is not active (type II recessiveness). Both definitions of recessiveness are then discussed under the allelic penetrance models. Bayesian methods are applied to analyze two data sets: one regarding the effect of the haplotype [HLA-B8, SC01, DR3] on the inheritance of IgD and IgG4 immunoglobulin deficiencies in humans, and other related to two-locus action in the control of *Listeria* infection susceptibility in mice.

© 2008 Elsevier B.V. All rights reserved.

## 1. Introduction

A binary trait is a biological character with two possible outcomes (for example, susceptibility or resistance to a certain infection). In the simplest case, the trait is affected by a single gene that often exhibits either classical dominant or recessive inheritance. However, most interesting binary traits, such as common human diseases, show a complex pattern of inheritance, since many genes and environmental factors are involved. Because of this complexity, a single gene seems neither sufficient nor necessary to the manifestation of the phenotype. Given a genotype, an individual has only a probability of expressing the phenotype, known as penetrance (Griffiths et al., 2000).

Genetic dissection of these complex binary traits aims to determine the underlying genetic architecture. Hitherto, this problem has been tackled either by generalized linear models in experimental populations (Cordell et al., 2001) or generalized linear mixed models in outbred populations and familial studies (Yi and Xu, 1999; Houwing-Duistermaat et al., 2003). In general, these models assume linear genetic effects of the loci on a convenient penetrance scale, as in most quantitative trait models (Lynch and Walsh, 1998). However, they show four caveats in terms of interpretation. First, the alleles conferring the phenotype are not explicitly specified. Second, the penetrance pattern describing the action of a dominant (or a recessive) allele is rather assumed than derived, as in Vieland and Huang (2003), where the penetrance

\* Corresponding address: Instituto Gulbenkian de Ciência, Apartado 14, 2781-901 Oeiras, Portugal. Tel.: +351 214464614; fax: +351 214407970.

E-mail addresses: [nunosep@igc.gulbenkian.pt](mailto:nunosep@igc.gulbenkian.pt) (N. Sepúlveda), [dpaulino@math.ist.utl.pt](mailto:dpaulino@math.ist.utl.pt) (C.D. Paulino), [cpenha@igc.gulbenkian.pt](mailto:cpenha@igc.gulbenkian.pt) (C. Penha-Gonçalves).

of one of the homozygotes equals to that of the heterozygotes. Third, it is difficult to know the most biologically appropriate scale for penetrance (Cordell et al., 2001). Often, it is determined by the model that best fits the data. Fourth, these models cannot be easily connected to a biological mechanism that might explain the inheritance of the phenotype (Cordell et al., 2001). Therefore, there is an urgent need to develop different statistical tools to analyze complex binary traits, as attempted in Di Serio and Vicard (2005) with the usage of graphical chain models to dissect complex diseases.

Recently, we proposed an allelic penetrance approach to model penetrance in experimental populations (Sepúlveda et al., 2007). Under this new framework, the alleles themselves are assumed to have a probability of being expressed at the level of the phenotype. By doing this, we could define different models for the action of dominant or recessive alleles as distinct conditions of allelic expressions. A dominant allele would lead to the phenotype when that allele is being expressed in the respective homozygotes and heterozygotes. Conversely, a recessive allele would manifest the phenotype when it is being expressed with no expression of the dominant allele. In this regard, some heterozygotes with the expression of the recessive allele might have the phenotype, a situation usually not considered by geneticists. In fact, the classical definition of a recessive trait requires the expression of both recessive alleles in the homozygotes (Alper and Awdeh, 2000). Here, we extend the previous modeling approach to contemplate this situation. We consider not only the single locus case, but also the joint action of two diallelic loci. We fit single-locus models to data on two immunoglobulin deficiencies in humans (Alper and Awdeh, 2000). Two-locus interaction models are used to unravel the genetics underlying the susceptibility to *Listeria* infection in mice (Boyartchuk et al., 2001). Bayesian methods were applied to draw inferences of interest, as an alternative to the classical ones previously adopted. Since they consistently represent all existing uncertainty through probability distributions, it is possible to use all the available information about what is unknown (e.g., model parameters) and to produce clear, direct and meaningful inferences on it. Furthermore, modern MCMC simulation methods make practical implementation of the allelic penetrance models much less troubled.

The structure of the paper is the following: Section 2 describes the allelic penetrance approach and the respective models; Section 3 proposes a Bayesian analysis via Markov Chain Monte Carlo to fit the models; Section 4 illustrates the methodology with the analysis of two examples; Section 5 presents concluding remarks.

## 2. The allelic penetrance approach

When analyzing a particular set of genes, or more generally, loci, one may think that the phenotype is the manifestation of either the genotypes of those loci (*internal factors*) or other loci with minor effect in the genetic background plus environmental factors (*external factors*). Often, genetically identical individuals manifest different phenotypes, even under constrained environmental conditions, as happens in experimental populations. This suggests that internal factors have somehow an intrinsic stochastic action with respect to the expression of the phenotype (Alper and Awdeh, 2000; Rakyant et al., 2002). Autoimmune diseases, such as type I diabetes and multiple sclerosis, seem to have this intrinsic stochastic property. In general, an autoimmune disease results from erroneous immune responses against body components of an individual. T and B lymphocytes, two key players of the immune system, recognize and react to antigens through a receptor at the cell surface. Collectively, these cells exhibit a large but random receptor repertoire in every individual. When studying autoimmune diseases, some authors have found differences in these repertoires between non-concordant monozygotic twins (Zipris et al., 1991; Hohler et al., 1999; Haegert et al., 2003). In this context, the intrinsic nature of penetrance might be related to the randomness of T and B cell receptor repertoire formation. Classically, the external factors are considered to have a stochastic effect on the expression of the phenotype as well. Taking again autoimmune pathologies as an example, disease can be prevented or triggered by certain kind of infections (Léon et al., 2004), which might occur or not throughout lifetime of an individual. Therefore, a phenotype should result either from internal or external factors, both with stochastic expression.

The internal and external factors are considered to act independently of each other. This assumption deserves some comments. On the one hand, it is a matter of mathematical convenience, following closely what is often done in the analysis of quantitative traits, where interaction terms between genetic and environmental factors are usually not included in the respective linear models (Lynch and Walsh, 1998). On the other hand, it is reasonable to assume independence between internal and external factors when there is no information on environmental factors, as in this work, or when dealing with data from experimental populations, where the environment is under strict control. However, it is worth noting that external factors also reflect the role of other loci in the genetic background on penetrance, and thus the above assumption implies that the other loci in the genetic background should be located in different chromosomes, and they should have negligible interaction with the loci under study.

As mentioned above, the phenotype can be acquired either by the action of internal or external factors. Since we assume independence between internal and external factors, the penetrance of a genotype  $g$  follows the probabilistic relationship of two independent events

$$\pi_g = \pi_g^{\text{int}} + \pi_g^{\text{ext}} - \pi_g^{\text{int}} \pi_g^{\text{ext}}, \quad (1)$$

where  $\pi_g^{\text{int}}$  and  $\pi_g^{\text{ext}}$  are the internal and external penetrances of a genotype  $g$  attributable to action of internal and external factors, respectively.

The above equation can be simplified if one assumes further a general mean effect of external factors on penetrance (i.e.,  $\pi_g^{\text{ext}} = \pi^{\text{ext}}$ ). This is in close agreement to many observations regarding autoimmune diseases. In fact, it is thought that

these pathologies might be triggered by certain infections, and if so, this exposure should not depend on the genotype  $g$  under loci. Thus, Eq. (1) can be simplified into

$$\pi_g = \pi_g^{\text{int}} + (1 - \pi_g^{\text{int}})\pi^{\text{ext}}. \quad (2)$$

This decomposition of penetrance is then the core of the whole modeling. The internal penetrance  $\pi_g^{\text{int}}$  will throughout take different forms appropriately describing some genetic mechanisms.

Some authors suggest that, in an individual, the expression of the phenotype is intimately related to the expression of each individual allele at the genotype (Rakyan et al., 2002; Lalucque and Silar, 2004). Moreover, the phenotypic expression of the alleles seems to be a stochastic event, which can be explained by the epigenetic state of the alleles (Rakyan et al., 2002) or to be an intrinsic property of loss-of-function alleles (Lalucque and Silar, 2004). Therefore, it is reasonable to put forward the notion of allelic penetrance, which embodies the probability of an allele being expressed at the level of the phenotype (Sepúlveda et al., 2007). By doing this, internal penetrance can be modeled as a function of allelic penetrances, describing different genetic mechanisms, as we will see below.

### 2.1. Single-locus allelic penetrance models

Let us consider a diallelic locus with alleles  $a$  and  $b$ , which have allelic penetrances  $\theta_a$  and  $\theta_b$ , respectively, wherein the allele  $a$  has a dominant role with respect to the phenotype. The phenotype is then acquired when there is expression of at least one dominant allele  $a$  at the homozygotes and heterozygotes. If one assumes independent allelic expressions, the action of the dominant alleles  $a$  is equivalent to an intra-locus independent action of these alleles. In this situation, the internal penetrance of each genotype is

$$\pi_{aa}^{\text{int}} = \theta_a^2 + 2\theta_a(1 - \theta_a), \quad \pi_{ab}^{\text{int}} = \theta_a, \quad \text{and} \quad \pi_{bb}^{\text{int}} = 0. \quad (3)$$

The internal penetrance of genotype  $aa$  is based on a binomial distribution with 2 trials and probability of success  $\theta_a$ , reflecting the probability of having a single or both alleles  $a$  being expressed at the level of the phenotype. Note that the internal penetrance of genotype  $bb$  is equal to 0, because the expression of alleles  $b$  does not lead to the phenotype. To avoid any confusion with previous definitions of dominance, we refer to Eq. (3) as the dominant allele model.

The above equation shows that heterozygotes and homozygotes have distinct penetrances when external factors are included in the calculations (Eq. (2)). Therefore, under the allelic penetrance approach, the action of a dominant allele cannot be captured by assuming penetrance of one of the homozygotes equal to that of heterozygotes, as in Vieland and Huang (2003).

Let now allele  $a$  be recessive with respect to the phenotype. Classically, a recessive phenotype is only acquired when both recessive alleles are being expressed in the respective homozygotes (type I recessive allele model). Under this definition of a recessive allele, heterozygotes can only express the phenotype by the action of external factors. The internal penetrance is then given by

$$\pi_{aa}^{\text{int}} = \theta_a^2 \quad \text{and} \quad \pi_{ab}^{\text{int}} = \pi_{bb}^{\text{int}} = 0. \quad (4)$$

Since heterozygotes and homozygotes  $bb$  have null internal penetrance, their penetrance are equal and attributed to external factors. In this case, type I recessive allele model agrees with the assumption of Vieland and Huang (2003).

At this point, it is worth noting that both classical dominant and recessive definitions are intimately related to the number of phenotype-conferring alleles being expressed. Dominant allele model requires at least one phenotype-conferring allele to be expressed, while type I recessive allele model relies on the expression of two phenotype-conferring alleles (at the respective homozygotes). These classical concepts can be regarded as intra-locus cumulative action models as defined in Section 2.2.3.

In previous work, we adopted an alternative definition for the action of a recessive allele (Sepúlveda et al., 2007). There, the heterozygotes might manifest the phenotype by the expression of recessive allele  $a$  when the dominant allele  $b$  is not active. As in the dominant allele model, the homozygotes  $aa$  might have the phenotype by expressing at least one allele  $a$ , and the homozygotes  $bb$  cannot intrinsically express the phenotype, because alleles  $b$  cannot confer the phenotype. This idea is embodied in the following equation

$$\pi_{aa}^{\text{int}} = \theta_a^2 + 2\theta_a(1 - \theta_a), \quad \pi_{ab}^{\text{int}} = \theta_a(1 - \theta_b), \quad \text{and} \quad \pi_{bb}^{\text{int}} = 0. \quad (5)$$

Three important comments should be made to the above equation (type II recessive allele model). First, the heterozygotes and homozygotes have distinct penetrances, as opposed to what is assumed in Vieland and Huang (2003). Therefore, single-locus allelic penetrance models agree with the assumption of Vieland and Huang (2003) only under classical definition of a recessive allele. Second, when the dominant allele is fully penetrant ( $\theta_b = 1$ ), Eq. (5) leads to  $\pi_{aa}^{\text{int}} = \theta_a^2 + 2\theta_a(1 - \theta_a)$ ,  $\pi_{ab}^{\text{int}} = \pi_{bb}^{\text{int}} = 0$ . In this case, both Eqs. (4) and (5) can be rewritten as  $\pi_{aa}^{\text{int}} = \eta$ ,  $\pi_{ab}^{\text{int}} = \pi_{bb}^{\text{int}} = 0$ , where  $\eta$  can be defined either by  $\theta_a^2 + 2\theta_a(1 - \theta_a)$  or  $\theta_a^2$ . Therefore, type II recessive allele model with a fully penetrant dominant allele is statistically indistinguishable from the type I recessive allele model. Third, when  $\theta_b = 0$ , Eq. (5) converts into the dominant allele model (Eq. (3)), and thus dominance is a special case of type II recessive allele model.

The above allelic penetrance models show a remarkable feature. When  $\theta_a = 1$ , Eqs. (3) and (4) result in Mendelian dominance and recessiveness inheritance, respectively. Eq. (5) with  $\theta_a = 1$  and  $\theta_b = 1$  also imply Mendelian recessiveness. Therefore, these models might be viewed as a stochastic version of these classical genetic concepts, as opposed to current linear models.

## 2.2. Two-locus allelic penetrance models

Here we address the joint action of two diallelic loci A and B. To this end, we use the decomposition of penetrance in Eq. (2) with genotype  $g$  regarding the combined genotype  $g_{AgB}$  of the two loci. Different genetic interaction models can be obtained by modeling the internal penetrance appropriately. Some of them are described as follows.

### 2.2.1. Independent action models

The simplest genetic action between two loci is genetic heterogeneity, where each locus manifests independently the phenotype. Using the probabilistic relationship for two independent events, Risch (1990) proposed the following model

$$\pi_{g_{AgB}} = \pi_{g_A} + \pi_{g_B} - \pi_{g_A}\pi_{g_B} \Leftrightarrow 1 - \pi_{g_{AgB}} = (1 - \pi_{g_A})(1 - \pi_{g_B}), \quad (6)$$

where  $\pi_{g_i}$  is the penetrance attributable to a genotype  $g$  at locus  $i = A, B$ . It is worth noting that  $\{\pi_{g_i}\}$  have no particular structure. Thus, Eq. (6) is the most general model for genetic heterogeneity. However, one might think that there are either dominant or recessive alleles at each locus. One way to include their genetic nature in the above model is to impose restrictions on  $\{\pi_{g_i}\}$ . Vieland and Huang (2003) interpret either dominance (or recessiveness) when the heterozygotes of a locus have equal penetrance to one of the homozygotes at the same locus. That is, if locus A has alleles  $a_1$  and  $a_2$ , then  $\pi_{a_1a_1} = \pi_{a_1a_2}$  or  $\pi_{a_2a_2} = \pi_{a_1a_2}$ . However, this restriction does not clearly imply which alleles at each locus are actually conferring the phenotype.

The independent action models (IAMs) aim to explicitly specify the phenotype-conferring allele at each locus and its genetic nature (dominant and recessive). The models obey to the decomposition of penetrance given in Eq. (2). Because of this, IAMs are also able to distinguish the effect of external factors on penetrance, as opposed to previous heterogeneity models. Since the two loci act independently of each other, the internal penetrance of genotype  $g_{AgB}$  satisfies Risch's model (Eq. (6)), but with  $\pi_{g_i}$  representing the internal penetrance of a genotype  $g$  at locus  $i$ . After some algebra, penetrance of a combined genotype  $g_{AgB}$  according to IAMs follows

$$1 - \pi_{g_{AgB}} = (1 - \pi_{g_A}^{\text{int}})(1 - \pi_{g_B}^{\text{int}})(1 - \pi^{\text{ext}}). \quad (7)$$

Finally, the action of dominant and recessive alleles are included in the models. If the phenotype-conferring allele at one locus is dominant, then the corresponding  $\pi_{g_i}^{\text{int}}$  follows Eq. (3). Analogously, if the phenotype-conferring allele is recessive, then the respective  $\pi_{g_i}^{\text{int}}$  is given either by Eqs. (4) or (5) (type I or type II recessive allele models, respectively).

### 2.2.2. Inhibition models

Classically, epistasis describes a genetic mechanism whereby an allele of a given locus prevents an allele of another locus from manifesting its effect (Griffiths et al., 2000). For example, Sepúlveda et al. (2005) show that the rearrangement process of genes encoding the T cell receptor follows this kind of mechanism. However, the term epistasis has been used in many different contexts, which led to a great confusion about its formal definition (Cordell, 2002). Here the inhibition models (IMs) are developed to capture the classical definition of epistasis. Thus, one locus confers the phenotype by the expression of its respective phenotype-conferring allele, whereas the other locus simply inhibits the phenotypic expression of the former by its inhibiting-action alleles. Phenotype-conferring or inhibiting-action alleles can be considered either dominant or recessive.

Let locus A have an allele conferring the phenotype and locus B an allele inhibiting the expression of the former allele. In this case, the internal penetrance is given by the probability of the phenotype-conferring alleles at locus A being expressed when there is no expression of inhibiting-action alleles at locus B. Thus, the internal penetrance for combined genotype  $g_{AgB}$  in Eq. (2) satisfies

$$\pi_{g_{AgB}}^{\text{int}} = \pi_{g_A}^{\text{int}} (1 - \pi_{g_B}^{\text{int}}), \quad (8)$$

where  $\pi_{g_A}^{\text{int}}$  is the probability of genotype  $g_A$  expressing the phenotype and  $\pi_{g_B}^{\text{int}}$  is the probability of genotype  $g_B$  having an inhibiting action. The action of dominant and recessive alleles are included in the model by replacing  $\pi_{g_A}^{\text{int}}$  and  $\pi_{g_B}^{\text{int}}$  with the respective single-locus internal penetrances (Eqs. (3)–(5)).

### 2.2.3. Cumulative action models

Genetic liability refers to a latent quantitative trait underlying the inheritance of a binary trait (Lynch and Walsh, 1998). In this scenario, the phenotype is acquired when an individual has its liability above (or below) a certain limit. In the same line of thought, allelic liability can be put forward, but now regarding the overall number of phenotype-conferring alleles in an individual (Stewart, 2002). The cumulative action models (CAMs) go further regarding allelic liability, extending it to allelic expression. This means that the phenotype is inherited when the joint expression of the phenotype-conferring alleles



at both loci exceeds a certain threshold  $t$ . Note that dominant and recessive alleles are not included in the model because what matters here is the cumulative expression of the phenotype-conferring alleles.

Let  $x_i$  represent the number of phenotype-conferring alleles in the genotype of locus  $i = A, B$ . For sake of simplicity, the combined genotype of the two loci is now denoted by  $x_A x_B$ . Let also  $Y_i$  be the random variable corresponding to the number of those alleles expressing the phenotype at locus  $i$ . According to the allelic penetrance approach,  $Y_i | x_i$  has a Binomial distribution with  $x_i$  trials and probability of success given by the allelic penetrance  $\theta_i$  of the phenotype-conferring allele at locus  $i$ . 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 a combined genotype  $x_A x_B$  is determined by

$$P[Y = y | x_A, x_B] = \sum_{(y_A, y_B) \in \mathcal{P}} P[Y_A = y_A | x_A] P[Y_B = y_B | x_B], \quad (9)$$

where  $\mathcal{P} = \{(y_A, y_B) \in \{0, \dots, x_A\} \times \{0, \dots, x_B\} : y_A + y_B = y\}$  and

$$P[Y_i = y_i | x_i] = \binom{x_i}{y_i} \theta_i^{y_i} (1 - \theta_i)^{x_i - y_i}. \quad (10)$$

Thus, for each CAM, the internal penetrance of a combined genotype  $x_A x_B$  entails the following expression

$$\pi_{x_A x_B}^{\text{int}} = P[Y \geq t | x_A, x_B] = \begin{cases} \sum_{y=t}^{x_A + x_B} P[Y = y | x_A, x_B], & \text{if } t \leq x_A + x_B \\ 0, & \text{otherwise,} \end{cases} \quad (11)$$

where  $P[Y = y | x_A, x_B]$  is given by Eq. (9) for some  $t = 1, \dots, 4$ . It is worth noting that, for the same phenotype-conferring alleles at each locus, there are 4 CAMs by changing the allelic threshold  $t$  in the expression of the phenotype.

### 3. Bayesian analysis

Data from experimental populations, or of unrelated individuals, are usually described as a contingency table  $G \times 2$ , where  $G$  is the number of genotypes (of one or more loci) under analysis. Denoting  $n_g$  the number of sample units with genotype  $g$  that exhibit the trait (out of  $m_g$ ), the sampling model assumed for  $\{n_g\}$  is the product of  $G$  independent binomial distributions  $\{Bin(m_g, \pi_g)\}$ , where  $\pi_g$  follows Eq. (2) with  $\pi_g^{\text{int}}$  defined by the allelic penetrance model to be fitted to data. Hence, the respective probability function is

$$f(\{n_g\} | \{\pi_g^{\text{int}}\}, \pi^{\text{ext}}) = \prod_g \binom{m_g}{n_g} [\pi_g^{\text{int}} + (1 - \pi_g^{\text{int}}) \pi^{\text{ext}}]^{n_g} [(1 - \pi_g^{\text{int}})(1 - \pi^{\text{ext}})]^{m_g - n_g}. \quad (12)$$

A Bayesian analysis is adopted to compare, select, and estimate allelic penetrance models by taking advantage of its ability to account for all uncertainty. All models rely on allelic and external penetrances (overall denoted by  $\eta$ ), which are assumed to be independent *a priori*. Since allelic and external penetrances are novel concepts in Genetics, it is reasonable to use non-informative priors for them, namely uniform distributions. Note that, as a consequence, the corresponding priors for the genotypic penetrances can hardly be regarded as non-informative from this point of view, as illustrated in the next section.

We used Markov Chain Monte Carlo (MCMC) methods via WinBUGS (Spiegelhalter et al., 2003) to simulate posterior distributions, because these distributions do not have closed form expressions. Sepúlveda (2004) showed that the full conditional distributions of the parameters are, in general, log-concave, which allows the software to use the adaptive rejection method (Gilks, 1992).

Since many models can be entertained, we designed a feasible strategy of model comparison and selection. First, the effects of the alleles of each locus are empirically assessed. In this way, we define a first set of models with the most plausible phenotype-conferring or inhibiting action alleles at each locus. Second, for each model, we compute in WinBUGS the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and the posterior mean of Pearson's parametric function (PMP). We select then the models with lowest values for these two measures. Finally, the selected models are compared according to the prior predictive probability  $p(\{n_g\})$  (PPP) and the sum of the conditional predictive ordinates in log-scale (SLNCPO). In this regard, one should select models that show the highest values for these two measures.

Estimation of PPP can present a serious computational problem, because it is often difficult to have accurate results. When model complexity is not so high, as in the single-locus models, one can use numerical integration to estimate PPP. However, two-locus models may have a more complex parametric structure, and thus numerical integration might not lead to good approximations to the exact solution. Alternative methods are then needed.

A simple way to estimate PPP is to use the ordinary Monte Carlo method based on the average of the likelihood values related to values of  $\eta$  simulated from their prior distribution. In spite of its simplicity, this approach may not produce trustworthy estimates, as we will see in the next section. Newton and Raftery (1994) showed that PPP is the posterior harmonic mean of the likelihood, which suggests the following estimator

$$p_{NR}(\{n_g\}) = \left[ \frac{1}{K} \sum_{k=1}^K \frac{1}{f(\{n_g\} | \eta^{(k)})} \right]^{-1}, \quad (13)$$

**Table 1**

Penetrance of immunoglobulins D (IgD) and G4 (IgG4) deficiencies in homozygotes, heterozygotes and non-carriers of the [HLA-B8, SC01, DR3] haplotype (percentage in parenthesis).

[HLA-B8, SC01, DR3] status	IgD deficiency	IgG4 deficiency
Homozygotes	11/30 (37%)	9/30 (30%)
Heterozygotes	12/59 (20%)	2/59 (3%)
Non-carriers	3/61 (5%)	1/61 (2%)

where  $\eta^{(k)}$  is one of the  $K$  simulated values from the joint posterior distribution of the allelic and external penetrances. However, this estimator is not a reliable alternative due to its known instability across simulations. Resorting to the BIC approximation for the PPP, Raftery et al. (2007) proposed to use a posterior simulation-based version of BIC (BIC-MC), thus avoiding the computation of maximum log-likelihood, which yields the following approximation

$$\log p_{BIC-MC}(\{n_g\}) = \bar{l} - s_l^2(\log(m) - 1), \quad (14)$$

where  $\bar{l}$  and  $s_l^2$  are the posterior mean and variance of the log-likelihood, respectively, and  $m = \sum_g m_g$ . As we will see in the single-locus example, this seems a suitable approach to estimate PPP when comparing the allelic penetrance models in a Bayesian framework.

In our case, SLNCPO is given by

$$\sum_{g=1}^G [n_g \log p(z_g = 1|z_{(-g)}) + (m_g - n_g) \log p(z_g = 0|z_{(-g)})], \quad (15)$$

where  $p(z_g|z_{(-g)})$  denotes the predictive conditional probability of an observation referring to an individual with genotype  $g$  given the original data set without that observation. To estimate  $p(z_g|z_{(-g)})$ , we use the following formula (Gelfand, 1996)

$$\hat{p}(z_g|z_{(-g)}) = \left[ \frac{1}{K} \sum_{k=1}^K \frac{1}{f(z_g|z_{(-g)}, \eta^{(k)})} \right]^{-1}, \quad (16)$$

where  $f(z_g|z_{(-g)}, \eta^{(k)})$  is the conditional probability of  $z_g$  given  $\eta^{(k)}$  and the data set without observation  $z_g$ . Since one considers independence between observations, then  $f(z_g|z_{(-g)}, \eta^{(k)}) = f(z_g|\eta^{(k)})$ , where  $f(z_g|\eta^{(k)})$  is a Bernoulli distribution with success probability equal to the genotypic penetrance  $\pi_g$ .

Bayesian estimation of the allelic penetrance models involved the calculation of the posterior mean, median, and standard deviation (SD) for allelic, external, and genotypic penetrances. Highest posterior density (HPD) credible intervals were determined in BOA software (Smith, 2007) through a method proposed by Chen and Shao (1999). A  $100 \times \gamma\%$  credible region for  $G$  genotypic penetrances was obtained by the cartesian product of the  $100 \times (1 - (1 - \gamma/G)\%)$  individual HPD credible intervals for these parameters, according to Bonferroni inequality. The credibility level  $\gamma$  was set up at 0.95.

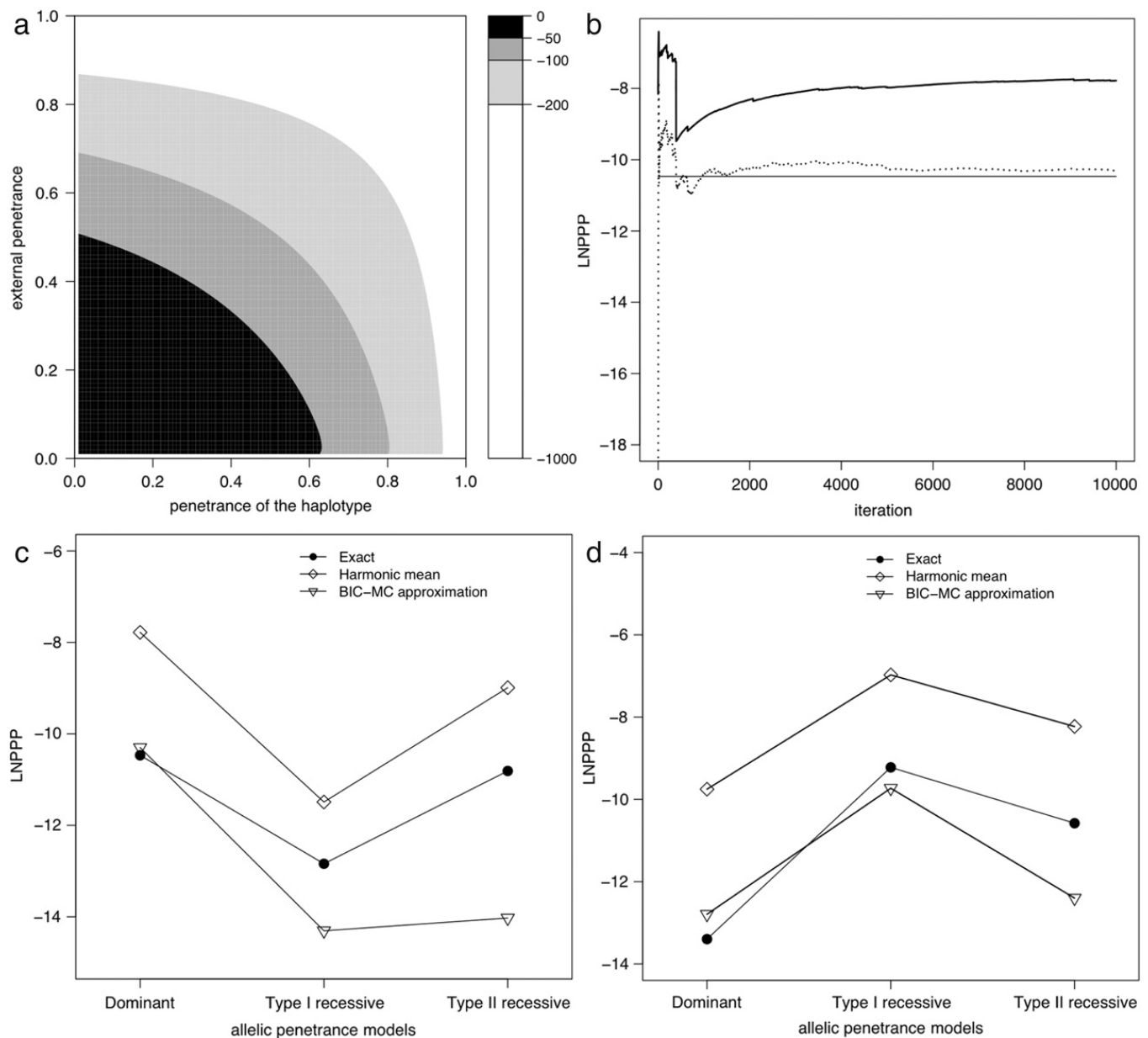
## 4. Applications

In this section, Bayesian analysis is illustrated with two data sets taken from the literature. In both examples, a good convergence to posterior distributions was obtained by simulating chains of length 110,000 with the first 10,000 iterations discarded as a burn-in period and a lag of 10 to remove autocorrelation across simulation to obtain  $K = 10,000$  values for posterior samples; see Sepúlveda (2004) for convergence analysis. The programs written in WinBUGS are available from the authors upon request.

For illustrative purposes about the construction of the joint posterior distribution for the allelic penetrance models, we shall detail its derivation for the second data set related to two-locus models.

### 4.1. Immunoglobulin deficiencies

Alper et al. (2000) studied the effect of the extended major histocompatibility complex (MHC) haplotype [HLA-B8, SC01, DR3] on the inheritance of several immunoglobulin deficiencies. In their study, there are three groups of unrelated patients: homozygotes for [HLA-B8, SC01, DR3], heterozygotes for this extended haplotype, and individuals who carried only other MHC haplotypes (non-carriers of [HLA-B8, SC01, DR3]). Here we analyze data of immunoglobulins D (IgD) and G4 (IgG4) deficiencies (Table 1). The aim of the analysis is to determine whether these traits are either dominant or recessive with respect to that haplotype. Previously, Alper et al. remarked that there is an increased frequency of IgD deficiency in the homozygotes and heterozygotes. Using similar arguments to those of the dominant allelic penetrance model, they considered IgD deficiency as a dominant trait, yet with no statistical support. With respect to IgG4 deficiency, penetrance seems only increased in the homozygous patients. The same authors stated, again without a supporting statistical analysis, that IgG4 deficiency is a recessive trait requiring the expression of two copies of the haplotype [HLA-B8, SC01, DR3], as in a type I recessive allele model.



**Fig. 1.** Estimating the logarithm of predictive prior probability (LNPPP): (a) log-likelihood contour plot of the dominant model for IgD deficiency; (b) stability of the posterior harmonic mean and BIC-MC estimators across MCMC simulation (solid and dotted lines, respectively) for dominant model regarding IgD deficiency, where the horizontal line represents the exact value of LNPPP; (c) and (d) different estimates of LNPPP for IgD and IgG4 deficiencies, respectively.

The first step in the analysis is to specify the disease-causing allele. Here it is assumed to be the haplotype [HLA-B8, SC01, DR3]. To infer its genetic nature, three single-locus models (dominant allele model, and both type I and type II recessive allele models) are compared with each other. As stated before, it is often hard to obtain reliable estimates for PPP. However, data refer to three genotypes only, and the three above-mentioned models do not have an extremely complex parametric structure. Thus, for each model, PPP could be best estimated by numerical integration of the likelihood, referred to as the “exact” solution. Because of this, we can evaluate the performance of the different estimators described in previous section (Fig. 1). For each model, simulation from prior distributions led to extremely low estimates for PPP (results not shown), because the mass of the log-likelihood is concentrated in a small region of the parameter space (Fig. 1(a)), which cannot be taken into account when generating values from independent uniform distributions. Although the posterior harmonic mean estimator does not show severe instability here (Fig. 1(b)), it overestimates the exact PPP (Fig. 1(c) and (d)). In most cases, the BIC-MC estimate is closer to the exact solution than the posterior harmonic mean (Fig. 1(c) and (d)), except for the type II recessive allele model regarding IgD deficiency. Moreover, it always shows stability (Fig. 1(b)). Nevertheless, both estimates agree in the ordering of the models in terms of plausibility (Fig. 1(c) and (d)). For all of this, we recommend the usage of the BIC-MC method to estimate PPP. Note that a better approximation for PPP could be obtained by the classical BIC approximation (results not shown). However, this requires additional calculations outside WinBUGS to obtain maximum likelihood estimates, which is not practical under a Bayesian analysis.



**Table 2**

Comparison of single-locus allelic penetrance models for IgD and IgG4 deficiency data. PMP is the posterior mean of Pearson's parametric function; DIC is the deviance information criterion; SLNCPO is the sum of the logarithms of conditional predictive ordinates; LNPPP is the logarithm of the predictive prior probability calculated by numerical integration of the likelihood.

Trait	Models	PMP	DIC	SLNCPO	LNPPP
IgD deficiency	Dominant allele	2.07	14.72	−63.47	−10.47
	Type I recessive allele	8.61	21.83	−67.04	−12.84
	Type II recessive allele	3.27	16.79	−64.64	−10.81
IgG4 deficiency	Dominant allele	7.99	18.27	−37.34	−13.40
	Type I recessive allele	2.44	12.45	−34.42	−9.22
	Type II recessive allele	2.97	13.20	−34.92	−10.58

**Table 3**

Posterior estimates for the relevant parameters of the dominant allele model and type I recessive allele model for IgD and IgG4 deficiencies, respectively. Posterior medians are similar to posterior means up to 1% (results not shown).

Trait	Parameters	Mean	SD	HPD CI
IgD deficiency	Haplotype penetrance	0.17	0.04	(0.09; 0.26) <sup>a</sup>
	External penetrance	0.06	0.03	(0.01; 0.12) <sup>a</sup>
	Genotypic penetrances			
	Homozygotes	0.35	0.06	(0.22; 0.50) <sup>b</sup>
	Heterozygotes	0.22	0.04	(0.14; 0.31) <sup>b</sup>
	Non-carriers	0.06	0.03	(0.01; 0.14) <sup>b</sup>
IgG4 deficiency	Haplotype penetrance	0.52	0.08	(0.36; 0.68) <sup>a</sup>
	External penetrance	0.03	0.02	(0.01; 0.07) <sup>a</sup>
	Genotypic penetrances			
	Homozygotes	0.30	0.08	(0.12; 0.49) <sup>b</sup>
	Heterozygotes	0.03	0.02	(0.01; 0.08) <sup>b</sup>
	Non-carriers	0.03	0.02	(0.01; 0.08) <sup>b</sup>

<sup>a</sup> HPD credible intervals at 95%.

<sup>b</sup> HPD credible intervals at 98.3%.

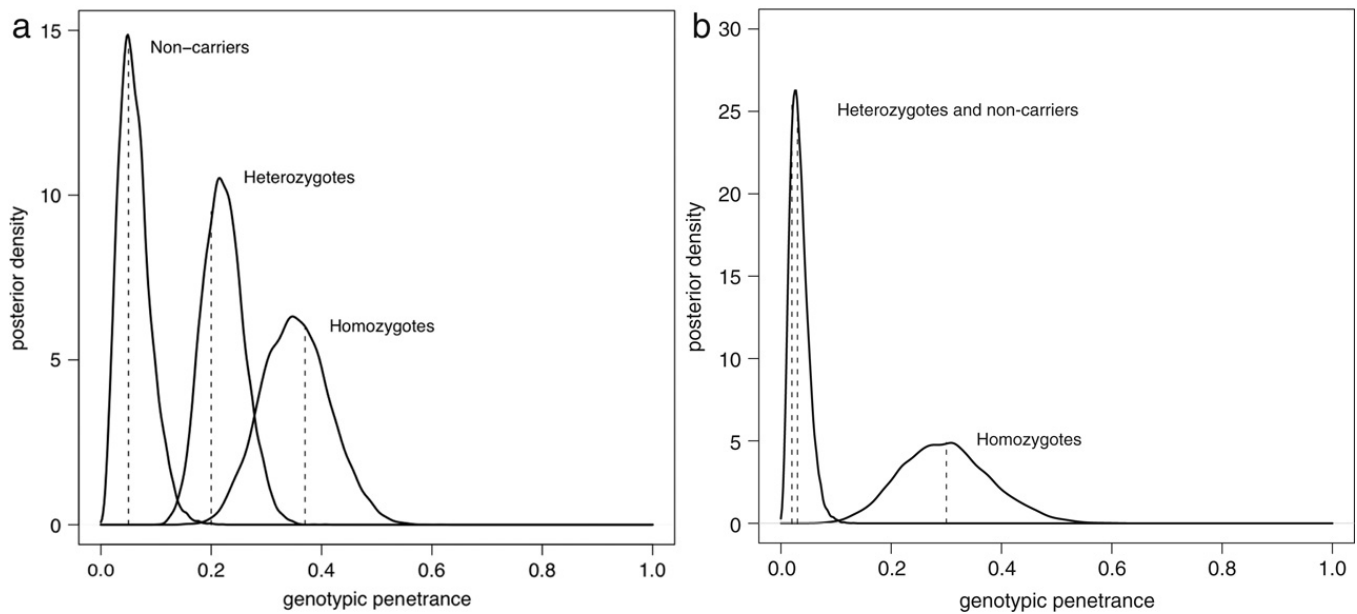
To compare the models, Table 2 shows the results for PMP, DIC, SLNCPO and PPP measures. All these measures favor a dominant and a type I recessive inheritance of IgD and IgG4 deficiencies, respectively. In fact, for the respective data set, these two models show the smallest PMP and DIC, and the highest SLNCPO and PPP. Therefore, IgD deficiency seems a dominant trait with respect to the haplotype [HLA-B8, SC01, DR3], while IgG4 deficiency might be a recessive trait requiring the simultaneous expression of two copies of the haplotype [HLA-B8, SC01, DR3].

The next step of the analysis is to estimate the above-selected models (Table 3). In IgD deficiency, the haplotype [HLA-B8, SC01, DR3] has low allelic penetrance, as demonstrated by its HPD credible interval (from 9% to 26%). The posterior mean of external penetrance is around 6% and the respective HPD credible interval ranges from 1% to 12%. Therefore, the effect of external factors in IgD deficiency is almost negligible. The posterior mean and median of genotypic penetrances are similar to the observed penetrances. The same happens with the posterior mode (Fig. 2(a)). For all of this, the dominant allele model seems to fit quite well the data. With respect to IgG4 deficiency, the haplotype [HLA-B8, SC01, DR3] has higher allelic penetrance than in IgD deficiency, but far from being fully penetrant, as suggested by the respective HPD credible interval (from 36% to 68%). As in IgD deficiency, external factors play a minor role in the inheritance of the trait; see posterior mean for external penetrance, and the respective HPD credible interval (from 1% to 7%). Finally, the different estimates for genotypic penetrances are in close agreement with the observed values (see also Fig. 2(b)). Therefore, the type I recessive allele model describes well IgG4 deficiency data.

#### 4.2. Susceptibility to *Listeria* infection

Boyartchuk et al. (2001) reported a genetic mapping of listeria infection susceptibility using an intercross between two mice strains, one susceptible to infection and other resistant. This study suggested that infection susceptibility might be under control of two loci A and B at chromosomes 5 and 13, respectively. Penetrance of susceptibility in each genotype combination of the two loci is shown in Table 4. The goal of the analysis is to infer the joint action of both loci with respect to susceptibility. Previously, we fitted the allelic penetrance models considering type II recessive alleles (Sepúlveda et al., 2007). We found an independent action between a dominant allele inherited from the resistant strain at locus A and a type II recessive allele derived from the susceptible strain at locus B. Here, we extend the previous analysis to account for type I recessive alleles in the models.

We use the following notation for the models. Upper and lower cases represent dominant and recessive alleles, respectively. For example, an independent action model with phenotype-conferring alleles  $A_1$  (dominant) and  $b_2$  (recessive) is denoted by IAM( $A_1/b_2$ ); an inhibition model with the same alleles is denoted by IM( $A_1^c/b_2^i$ ), where superscripts  $c$  and  $i$  denote a phenotype-conferring allele and an allele with an inhibiting action, respectively; and CAM $_k(a_1/b_1)$  stands for a cumulative action model requiring jointly the expression of at least  $k$  phenotype-conferring alleles  $a_1$  and  $b_1$ .



**Fig. 2.** Posterior densities of genotypic penetrances: (a) dominant allele model for IgD deficiency; (b) type I recessive allele model for IgG4 deficiency. Dashed lines show the observed penetrances.

**Table 4**

Penetrance of susceptibility to listeria infection in  $F_2$  mice (percentage in parenthesis), where  $a_1$  and  $b_1$  represent the alleles derived from the susceptible strain, while  $a_2$  and  $b_2$  are from the resistant strain. Loci A and B are at chromosomes 5 and 13, respectively.

Locus A genotype	Locus B genotype	Penetrance (%)
$a_1a_1$	$b_1b_1$	4/7 (57.1)
$a_1a_1$	$b_1b_2$	4/19 (21.1)
$a_1a_1$	$b_2b_2$	1/12 (8.3)
$a_1a_2$	$b_1b_1$	23/24 (95.8)
$a_1a_2$	$b_1b_2$	10/31 (32.3)
$a_1a_2$	$b_2b_2$	6/13 (46.2)
$a_2a_2$	$b_1b_1$	10/10 (100.0)
$a_2a_2$	$b_1b_2$	8/12 (66.7)
$a_2a_2$	$b_2b_2$	5/9 (55.6)

To construct the joint posterior distribution of the models, we recall that the sampling distribution is given by Eq. (12) with  $g = g_{AgB}$ . Since we consider uniform priors for the parameters, this equation defines the respective kernel of the joint posterior distribution. This distribution then assumes different forms according to the internal penetrances of the model under analysis. Thus, it is worthwhile to give some examples of internal penetrances to have a hint on the differences between the joint posterior distributions of the models. Let us focus on a particular genotype of Eq. (12), e.g.,  $g_{AgB} = a_1a_2|b_1b_2$ . For some allelic penetrance models, we have

$$\pi_{a_1a_2|b_1b_2}^{\text{int}} = \begin{cases} 0, & \text{IAM}(a_2|b_1)^{\text{I}} \\ \pi_{a_2}(1 - \pi_{a_1}) + [1 - \pi_{a_2}(1 - \pi_{a_1})]\pi_{b_1}(1 - \pi_{b_2}), & \text{IAM}(a_2|b_1)^{\text{II}} \\ \pi_{a_2}, & \text{IM}(A_2^c|b_2^i)^{\text{I}} \\ \pi_{a_2}[1 - \pi_{b_2}(1 - \pi_{b_1})], & \text{IM}(A_2^c|b_2^i)^{\text{II}} \\ \pi_{a_2}\pi_{b_1}, & \text{CAM}_2(a_2|b_1) \end{cases} \quad (17)$$

where  $\pi_{a_1}$ ,  $\pi_{a_2}$ ,  $\pi_{b_1}$  and  $\pi_{b_2}$  are the penetrance of the alleles  $a_1$ ,  $a_2$ ,  $b_1$  and  $b_2$ , respectively, and models with superscripts I and II have type I and type II recessive alleles, respectively. The above internal penetrances of IAMs were obtained according to Eq. (6) replacing both  $\pi_{g_A}$  and  $\pi_{g_B}$  by the heterozygote internal penetrance of the recessive allele models (either Eq. (4) or (5)). Both IMs are based on Eq. (8) with  $\pi_{g_A}^{\text{int}}$  given by the heterozygote internal penetrance of the dominant allele model (see Eq. (3)) and  $\pi_{g_B}^{\text{int}}$  given by the heterozygote internal penetrance of the recessive allele models. Finally, the internal penetrance of  $\text{CAM}_2(a_2|b_1)$  follows Eq. (11) with  $k = 2$ . The remaining internal penetrances of these and other models can be derived following similar reasoning.

If one takes into account all possible combinations of the alleles at the two loci, their genetic nature, and action, there are 100 allelic penetrance models under analysis. However, one can make this number of models, and also time of analysis, drop drastically by observing that in the data there is an increase of penetrance with the presence of either alleles  $a_2$  or  $b_1$

**Table 5**

Comparison of two-locus allelic penetrance models for listeria infection data. See further details in legend of Table 2. LNPPP was estimated by the BIC-MC approximation (Eq. (14)).

Models	Rec. <sup>a</sup>	PMP	DIC	SLNCPO	LNPPP
IAM( $A_2/B_1$ )	–	23.65	52.09	–	–
IAM( $A_2/b_1$ )	I	11.80	35.81	–71.62	–22.30
	II	12.61	36.43	–72.10	–23.84
IAM( $a_2/B_1$ )	I	31.24	59.64	–	–
	II	24.47	53.24	–	–
IAM( $a_2/b_1$ )	I	17.66	40.63	–	–
	II	13.38	37.28	–72.46	–24.80
IM( $A_1^i/B_1^c$ )	–	32.69	62.27	–	–
IM( $A_2^c/B_2^i$ )	–	15.64	39.85	–	–
IM( $A_1^i/b_1^c$ )	I	19.49	44.19	–	–
	II	18.09	43.67	–	–
IM( $a_1^i/B_1^c$ )	I	30.93	58.19	–	–
	II	30.96	58.96	–	–
IM( $a_1^i/b_1^c$ )	I	17.18	41.92	–	–
	II	16.35	41.58	–	–
IM( $A_2^c/b_2^i$ )	I	36.12	65.27	–	–
	II	16.91	40.98	–	–
IM( $a_2^c/B_2^i$ )	I	42.00	72.44	–	–
	II	18.65	43.41	–	–
IM( $a_2^c/b_2^i$ )	I	45.53	77.41	–	–
	II	17.43	42.12	–	–
CAM <sub>1</sub> ( $a_2/b_1$ )	–	23.65	52.09	–	–
CAM <sub>2</sub> ( $a_2/b_1$ )	–	18.83	43.12	–	–
CAM <sub>3</sub> ( $a_2/b_1$ )	–	15.62	40.38	–	–
CAM <sub>4</sub> ( $a_2/b_1$ )	–	42.57	72.55	–	–

<sup>a</sup> Definition of the recessive allele (type I, type II or undefined).

at the combined genotype. It is easy to show that IAMs and CAMs considering  $a_2$  and  $b_1$  as phenotype-conferring alleles at each locus are the only models in their respective class that agree with this observation. The remaining IAMs and CAMs (with at least one of the other alleles,  $a_1$  and  $b_2$ , as phenotype-conferring alleles) would not fit the data well, being then removed from the analysis. To define the most plausible inhibiting models, one needs to specify a phenotype-conferring at one locus and an inhibiting allele at the other locus. With respect to phenotype-conferring allele, it can be either  $a_2$  or  $b_1$ , as considered above for independent and cumulative action models. Conversely, the putative inhibiting allele can be either  $a_1$  or  $b_2$  at each locus. Since the phenotype-conferring and inhibiting alleles are assumed to be at different loci, there are only two possible combinations for this pair of alleles in the inhibiting models: ( $a_1$ ,  $b_1$ ) and ( $a_2$ ,  $b_2$ ). Finally, we consider all possible combinations of dominant and recessive (phenotype-conferring and inhibiting) alleles at each locus. At the end, we only need to compare 25 two-locus allelic penetrance models (Table 5).

We first compare the models with PMP and DIC measures (Table 5). In this regard, the best models are IAM( $A_2/b_1$ ) with  $b_1$  as either a type I or a type II recessive allele, and IAM( $a_2/b_1$ ) with both  $a_2$  and  $b_1$  as type II recessive alleles. Then, SLNCPO and PPP are computed for these three models (Table 5). To calculate PPP, we used BIC-MC estimator because it seems reliable, computationally stable, and easily calculated in WinBUGS. Both measures agree that IAM( $A_2/b_1$ ) with  $b_1$  as a type I recessive allele is the best model for the data. Since the same model is also the best according to PMP and DIC, we can state that infection susceptibility seems under control of an independent action between a dominant allele of the resistant strain at locus A and a type I recessive allele of the susceptible strain at locus B. Previously, IAM( $A_2/b_1$ ) with  $b_1$  as a type II recessive allele was considered to be the best model for the data (Sepúlveda et al., 2007). Therefore, this analysis improves on previous results by considering a type I recessive allele at locus B.

Table 6 shows relevant parametric estimates of IAM( $A_2/b_1$ ). It is worth noting that some genotypic penetrances have the same posterior estimates due to the parametric structure of the models (see, for example, penetrances of genotypes  $a_1a_1/b_1b_2$  and  $a_1a_1/b_2b_2$  that are only parameterized by external penetrance). Fig. 3 shows prior and posterior densities for two genotypic penetrances, which suggests a fair update of their prior distributions by experimental data. It is worth noting that, although uniform prior distributions had been considered to the allelic and external penetrances, the respective prior distributions for genotypic penetrances cannot be regarded as non-informative. The estimates for allelic penetrances pointed out that the alleles  $A_2$  and  $b_1$  have a moderate probability of being expressed at the level of the phenotype. The HPD credible interval for external penetrance ranges from 6% to 28%, which suggests the presence of other loci in the genetic background and/or external factors with a minor role in listeria infection susceptibility. The posterior means for the genotypic penetrance are close to the observed values, except for the genotype  $a_1a_1/b_1b_1$ . The same remarks can be drawn from the posterior credible region at 95% for the genotypic penetrances, where the observed values are within the respective HPD credible intervals at 99.44%, except for the above-mentioned genotype. In this regard, a better fit can be obtained from the same model IAM( $A_2/b_1$ ), but considering  $b_1$  as a type II recessive allele, instead of type I (results not shown). In fact, the respective credible region of genotypic penetrances includes all observed values. However, this is achieved by an increase

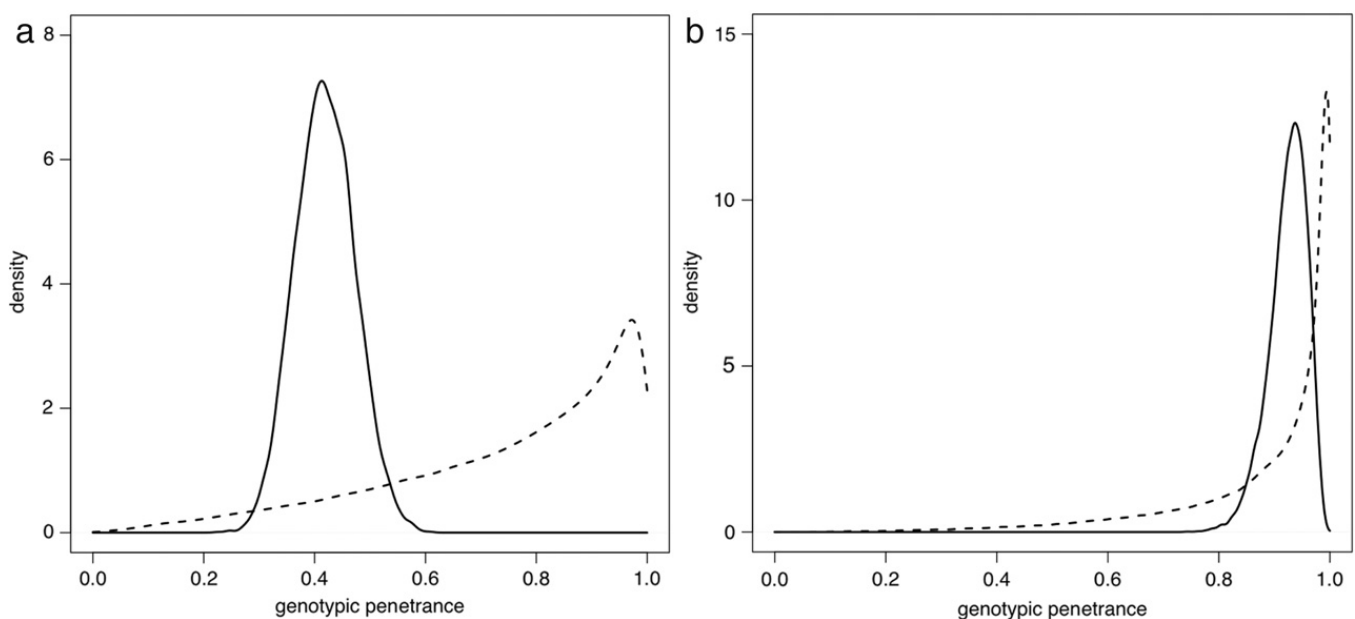
**Table 6**

Relevant posterior estimates according to  $\text{IAM}(A_2/b_1)$  with  $b_1$  as a type I recessive allele. Posterior medians are similar to posterior means up to 1% (results not shown).

Parameters	Mean	SD	HPD CI
Penetrance of allele $A_2$	0.30	0.07	(0.16; 0.45) <sup>a</sup>
Penetrance of allele $b_1$	0.58	0.10	(0.38; 0.76) <sup>a</sup>
External penetrance	0.16	0.06	(0.06; 0.28) <sup>a</sup>
Genotypic penetrances			
$a_1a_1/b_1b_1$	0.84	0.07	(0.61; 0.98) <sup>b</sup>
$a_1a_1/b_1b_2$	0.16	0.06	(0.03; 0.34) <sup>b</sup>
$a_1a_1/b_2b_2$	0.16	0.06	(0.03; 0.34) <sup>b</sup>
$a_1A_2/b_1b_1$	0.89	0.05	(0.74; 0.98) <sup>b</sup>
$a_1A_2/b_1b_2$	0.42	0.06	(0.29; 0.57) <sup>b</sup>
$a_1A_2/b_2b_2$	0.42	0.06	(0.29; 0.57) <sup>b</sup>
$A_2A_2/b_1b_1$	0.93	0.04	(0.81; 0.99) <sup>b</sup>
$A_2A_2/b_1b_2$	0.59	0.08	(0.39; 0.78) <sup>b</sup>
$A_2A_2/b_2b_2$	0.59	0.08	(0.39; 0.78) <sup>b</sup>

<sup>a</sup> HPD credible intervals at 95%.

<sup>b</sup> HPD credible intervals at 99.44%.



**Fig. 3.** Two examples of prior (dashed line) and posterior (solid line) densities of genotypic penetrances according to  $\text{IAM}(A_2/b_1)$  with  $b_1$  as a type I recessive allele: (a) penetrance of genotype  $a_1A_2/b_1b_2$  and (b) penetrance of genotype  $A_2A_2/b_1b_1$ .

in the amplitude of the respective HPD credible intervals due to an additional parameter in the model. Therefore, for sake of simplicity,  $\text{IAM}(A_2/b_1)$  with  $b_1$  as a type I recessive allele is, in our opinion, the best model to explain the data.

## 5. Concluding remarks

The present work considered the mathematical formulation of two definitions of a recessive allele under the allelic penetrance approach. In both examples, we have found evidence for type I recessive alleles, and none of type II. This might be due to both dominant allele model and the type I recessive allele model being two extreme cases of the type II recessive allele model. On one hand, when  $\theta_b = 1$  in Eq. (5), the type II recessive allele model leads to a parametric structure indistinguishable from the type I recessive allele model. On the other hand, when  $\theta_b = 0$  in Eq. (5), type II recessive allele model converts into the dominant allele model. Therefore, type II recessive allele model would only be the best to fit the data when  $\theta_b$  shows an intermediate value.

Bayesian methods are known to be useful to incorporate prior information coherently into data analysis. However, the allelic penetrance approach is still a novel genetic-based statistical framework to analyze complex binary traits and, therefore, it is currently difficult to elicit prior information from experts regarding the allelic and external penetrances. Inevitably, the subsequent Bayesian analysis would be based on non-informative settings such as those used here. Nevertheless, it is worth noting that the priors induced on the genotypic penetrances are far from being non-informative, as illustrated in Fig. 2. Therefore, it is important for geneticists to become familiar with allelic penetrance approach and its new concepts, though this will only be achieved with its wide application to experimental data. This would certainly allow enough prior

beliefs on genotypic penetrances to be acquired, making it possible to accommodate them in appropriate prior distributions. This might open the way to get corresponding priors for the original parameters of some interaction models by following an analogous procedure to that used in Paulino et al. (2003).

Here we applied MCMC methods via WinBUGS to simulate posterior distributions. However, data augmentation algorithm (Tanner and Wong, 1987) might alternatively be used on the basis that the observed data can be regarded as incomplete under the allelic penetrance models, because one cannot know whether a particular allele is being expressed in each individual.

Finally, our models heavily rely on the notion of allelic penetrance, which embodies the probability of an allele being expressed the level of the phenotype. Although this stochastic concept is certainly present at a cellular level, such as in the expression of T cell receptor genes (Sepúlveda et al., 2005) and cytokine genes (Paixão et al., 2007), the same might not be true at the organism level. Epigenetic mechanisms may support the notion of allelic penetrance in some cases (Rakyan et al., 2002; van Vliet et al., 2007; Strickland and Richardson, 2008), but they cannot be invoked in general. Therefore, other biological mechanisms are yet needed to justify a broader application of the allelic penetrance approach to the analysis of complex binary traits.

## Acknowledgements

We thank Victor Boyartchuk for providing the *Listeria* data set, Jorge Carneiro and Henrique Teotônio for valuable discussions, Rui Gardner and Eurico de Sepúlveda for reviewing the paper. The first two authors acknowledge financial support from Fundação para a Ciência e Tecnologia (fellowship SFRH/BD/19810 and through Center for Mathematics and Applications, IST).

## References

- Alper, C.A., Awdeh, Z., 2000. Incomplete penetrance of MHC susceptibility genes: Prospective analysis of polygenic MHC-determined traits. *Tissue Antigens* 56, 199–206.
- Alper, C.A., Marcus-Bagley, D., Awdeh, Z., Kruskall, M.S., Eisenbarth, G.S., Brink, S.J., Katz, A.J., Stein, R., Bing, D.H., Yunis, E.J., Schur, P.H., 2000. Prospective analysis suggests susceptibility genes for deficiencies of IgA and several other immunoglobulins on the [HLA-B8, SC01, DR3] conserved extended haplotype. *Tissue Antigens* 56, 207–216.
- Boyartchuk, V.L., Broman, K.W., Mosher, R.E., D'Orazio, S.E., Starnbach, M.N., Dietrich, W.F., 2001. Multigenic control of *Listeria monocytogenes* susceptibility in mice. *Nature Gen.* 27, 259–260.
- Chen, M.-H., Shao, Q.-M., 1999. Monte Carlo estimation of Bayesian credible and HPD intervals. *J. Comp. Graph. Stat.* 8, 69–92.
- 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, H., Todd, J., Hill, N., Lord, C., Lyons, P., Peterson, L., Wicker, L., Clayton, D., 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.
- Di Serio, C., Vicard, P., 2005. Graphical chain models for the analysis of complex genetic diseases: An application to hypertension. *Stat. Model.* 5, 119–143.
- Gelfand, A.E., 1996. Model determination using sampling-based methods. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 145–161.
- Gilks, W., 1992. Derivative-free adaptive rejection sampling for Gibbs sampling. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics 4*. Oxford University Press, Oxford, 641–665.
- Griffiths, A.J., Miller, J.H., Suzuki, D.T., Lewontin, R.C., Gelbart, W.M., 2000. *An Introduction to Genetic Analysis*. W. H. Freeman and Company, New York.
- Haegert, D.G., Galutira, D., Murray, T.J., O'Connor, P., Gadag, V., 2003. Identical twins discordant for multiple sclerosis have a shift in their T-cell receptor repertoires. *Clin. Exp. Immunol.* 134, 532–537.
- Hohler, T., Hug, R., Schneider, P.M., Krummenauer, F., Grienberg-Lerche, C., Granfors, K., Marker-Hermann, E., 1999. Ankylosing spondylitis in monozygotic twins: Studies on immunological parameters. *Ann. Rheum. Dis.* 58, 435–440.
- Houwing-Duistermaat, J.J., Bijkerk, C., Hsu, L., Stijnen, T., Slagboom, E.P., van Duijn, C.M., 2003. A unified approach to modelling linkage to quantitative and qualitative traits. *Ann. Hum. Genet.* 67, 457–463.
- Lalucque, H., Silar, P., 2004. Incomplete penetrance and variable expressivity of growth defect as a consequence of knocking out two K<sup>+</sup> transporters in the euscomycete fungus *Podospora anserina*. *Genetics* 166, 125–133.
- Léon, K., Faro, J., Lage, A., Carneiro, J., 2004. Inverse correlation between the incidences of autoimmune disease and infection predicted by a model of T cell mediated tolerance. *J. Autoimmun.* 22, 31–42.
- Lynch, M., Walsh, B., 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc, Sunderland, USA.
- Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Roy. Stat. Soc. B* 56, 3–48.
- Paixão, T., Carvalho, T.P., Calado, D.P., Carneiro, J., 2007. Quantitative insights into stochastic monoallelic expression of cytokine genes. *Immunol. Cell Biol.* 85, 315–322.
- Paulino, C.D., Soares, P., Neuhaus, J., 2003. Binomial regression with misclassification. *Biometrics* 59, 670–675.
- Raftery, A.E., Newton, M.A., Satagopan, J.M., Krivitsky, P.N., 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (Eds.), *Bayesian Statistics 8*. Oxford University Press, Oxford, pp. 371–416.
- Rakyan, V., Blewitt, M.E., Druker, R., Preis, J.L., 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.
- Sepúlveda, N., Statistical models for the joint action of two loci in complex binary traits (in portuguese). Master Thesis (Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, 2004).
- Sepúlveda, N., Boucontet, L., Pereira, P., Carneiro, J., 2005. Stochastic modelling of T cell receptor  $\gamma$  gene rearrangement. *J. Theor. Biol.* 234, 153–165.
- Sepúlveda, N., Paulino, C.D., Carneiro, J., Penha-Gonçalves, C., 2007. Allelic penetrance approach as a tool to model two-locus interaction in complex binary traits. *Heredity* 99, 173–184.
- Smith, B., Bayesian output analysis program (BOA) version 1.1.6 user's manual. (Department of Biostatistics, College of Public Health, University of Iowa, 2007).
- Spiegelhalter, D., Best, N., Carlin, B., van der Linden, A., 2002. Bayesian measures of model complexity and fit (with discussion). *J. Roy. Stat. Soc. B* 64, 583–640.
- Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., WinBUGS user manual, Version 1.4. (MRC Biostatistics Unit, Institute of Public Health & Department of Epidemiology and Public Health, Imperial College School of Medicine, 2003).



- Stewart, J., 2002. Towards the genetic analysis of multifactorial diseases: The estimation of allele frequency and epistasis. *Hum. Hered.* 54, 118–131.
- Strickland, F.M., Richardson, B.C., 2008. Epigenetics in human autoimmunity. *Autoimmunity* 41, 278–286.
- Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Stat. Assoc.* 82, 528–550.
- van Vliet, J., Oates, N.A., Whitelaw, E., 2007. Epigenetic mechanisms in the context of complex diseases. *Cell. Mol. Life Sci.* 64, 1531–1538.
- Vieland, V., Huang, J., 2003. Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected-sib-pairs. *Am. J. Hum. Genet.* 73, 223–232.
- Yi, N., Xu, S., 1999. A random model approach to mapping quantitative trait loci for complex binary traits in outbred populations. *Genetics* 153, 1029–1040.
- Zipris, D., Crow, A.R., Delovitch, T.L., 1991. Altered thymic and peripheral T-lymphocyte repertoire preceding onset of diabetes in NOD. *Diabetes* 40, 429–435.