From phenotype to genotype: a Bayesian solution

M. J. Denwood, A. E. Mather, D. T. Haydon, L. Matthews, D. J. Mellor and S. W. J. Reid*

Boyd Orr Centre for Population and Ecosystem Health, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G61 1QH, UK

The study of biological systems commonly depends on inferring the state of a ‘hidden’ variable, such as an underlying genotype, from that of an ‘observed’ variable, such as an expressed phenotype. However, this cannot be achieved using traditional quantitative methods when more than one genetic mechanism exists for a single observable phenotype. Using a novel latent class Bayesian model, it is possible to infer the prevalence of different genetic elements in a population given a sample of phenotypes. As an exemplar, data comprising phenotypic resistance to six antimicrobials obtained from passive surveillance of Salmonella Typhimurium DT104 are analysed to infer the prevalence of individual resistance genes, as well as the prevalence of a genomic island known as SGI1 and its variants. Three competing models are fitted to the data and distinguished between using posterior predictive p-values to assess their ability to predict the observed number of unique phenotypes. The results suggest that several SGI1 variants circulate in a few fixed forms through the population from which our data were derived. The methods presented could be applied to other types of phenotypic data, and represent a useful and generic mechanism of inferring the genetic population structure of organisms.

Keywords: phenotype to genotype; Bayesian; mixture model; latent class model; antimicrobial resistance; posterior predictive p-value

1. INTRODUCTION

In statistical terms, the interpretation of data obtained from many biological systems commonly depends on inferring the state of a hidden variable, such as an underlying genotype, from that of an observed variable, such as an expressed phenotype. Latent class analysis using multilevel modelling is a well-established quantitative tool for obtaining information about a hidden variable state given an observed variable state [1]. However, in most cases, assessing the presence of genetic elements encoding phenotypes also requires the use of mixture models because expressed phenotypes usually have more than one possible genetic mechanism. Fitting latent class mixture models to such data presents a difficulty when using the standard method for Bayesian model selection, the deviance information criterion (DIC) [2], because the effective number of parameters in the model, pD, is undefined when using latent Bernoulli parameters [3,4]. For this reason, calculation of DIC in the OpenBUGS software is disabled when using mixture models specified using a latent Bernoulli parameter [4]. Bayesian model selection using Bayes factors is theoretically possible for mixture models, but calculation of the complex integrals required is impractical for application to high-dimensional models such as these [5]. Posterior predictive p-values, which assess how extreme the data are given the model, provide an alternative way of assessing model fit that is independent of the likelihood [6–10]. This is usually performed by assessing the ability to predict the aspect of the data to which the model was fitted. However, an alternative is to use aspects of the data that are not directly modelled. Assessing model fit in this way allows a comparison of models that is not influenced by the effective number of parameters or degrees of freedom in the model [11], and is therefore suitable for use with mixture models. Such methods of assessing model fit also allow the possibility of all models being rejected. This is an advantage over the use of Bayes factors or likelihood-based model comparisons, which can only compare the performance of each model relative to the competing models.

Here, we use passive surveillance data for antimicrobial resistance in Salmonella Typhimurium DT104 (hereafter, DT104) to illustrate our approach. Antimicrobial resistance in pathogenic bacteria of humans and animals constitutes a major, and growing, obstacle in the effective treatment of common infections, and is an increasing source of concern within both the medical and veterinary professions [12,13]. Effective control of antimicrobial resistance requires an understanding of its spread through human and animal populations, for which knowledge of the genetic mechanisms underlying the observed resistance phenotype is essential. Resistance may be characterized phenotypically or genotypically, but current surveillance systems typically produce phenotypic resistance profiles, consisting of a string of logical values representing the observed resistance classification to a series of antimicrobials. The inability to infer the genetic basis of the observed resistance profiles limits our understanding of resistance spread. While genetic sequencing readily provides this information, it can be expensive and time-consuming to obtain for large numbers of...
samples. High-throughput sequencing may overcome some of these limitations, but this technique is not yet widely used in surveillance and cannot be used to obtain genetic information for historic data where samples are no longer viable. It is possible to infer the possible genetic mechanisms of resistance of an individual isolate using techniques such as interpretive reading [14,15] or ‘black box’ systems such as VITEK [16], but these techniques are non-quantitative and are not able to incorporate data from multiple isolates to strengthen inference. The numerous resistance genes reported in the literature demonstrate the various potential genetic mechanisms of resistance, but the relative prevalences of these underlying genotypes cannot be estimated easily where more than one genetic mechanism exists for a single observable phenotype. The ability to infer genotypic resistance information from the widely available phenotypic data would therefore provide an important advance in understanding the prevalence, evolution and spread of resistance determinants within bacterial pathogens, and aid in the monitoring of the development of antimicrobial resistance.

Resistance to the antimicrobials ampicillin (A), chloramphenicol (C), spectinomycin (Sp), streptomycin (St), sulphonamides (Sx) and tetracycline (T) in DT104 is commonly chromosomally encoded, the genes located in a 43 kb region termed Salmonella genomic island 1 (SGI1) [17]. Variants of this SGI1 in DT104, with different corresponding resistance profiles, have been reported [18–26]. However, while SGI1 is known to be common, it is possible that phenotypic resistance to these antimicrobials is due to other genes, or the same genes outside of SGI1. The true genetic mechanisms of resistance therefore constitute a hidden variable in a latent class model, and the phenotypic resistance profile for each isolate constitutes the observed variable. As previously discussed, differentiating these multiple potential genetic sources of a phenotypically characterized resistance profile also requires the use of mixture models. Based on prior information from the literature on the known SGI1 variants, we derive three alternative models representing each of the biologically plausible mechanisms of SGI1 operation, each with the same representation of genetic elements outside of SGI1. Model 1 allows observed resistances to be attributed to the full SGI1, containing genes for all six resistances examined (A, C, Sp, St, Sx and T); this is by far the most common ‘variant’ of the genomic island reported. Model 2 allows all six SGI1 variants that have been reported in DT104 to occur, including the absence of SGI1. For Model 3, the presence of individual genetic elements within SGI1 was not fixed, reflecting the possibility that genetic elements conferring resistance are mobile within the genomic island. Using posterior predictive p-values allows us to compare the observed number of unique phenotypes and the distribution of possible numbers of unique phenotypes given the model assumptions and posterior parameter distributions. Based on the ability to predict this aspect of the data despite the fact that it is not directly modelled, we use this metric to select the most biologically plausible model from the three candidates. These results provide new information on the prevalence of underlying mechanisms of the observed resistances that could not be obtained from phenotypic data using traditional methods.

2. MATERIAL AND METHODS
(a) Data collection and microbiological techniques
The data comprised 2761 human S. Typhimurium DT104 isolates submitted from 1990 to 2004 to the Scottish Salmonella, Shigella and Clostridium difficile Reference Laboratory (SSSCDRL), and are part of an ongoing passive surveillance programme for Salmonella. Diagnostic medical laboratories forwarded isolates identified as Salmonella to the SSSCDRL for confirmation and further testing; these were from domestic cases with no history of foreign travel. Serotyping of the isolates was accomplished according to internationally standardized methods based on the Kauffman–White scheme [27], and phage typing was carried out according to standardized protocols [28,29]. Antimicrobial susceptibility was assessed using a modified breakpoint method, involving solid agar plates containing a pre-determined concentration of antimicrobial. Categories of resistance to each antimicrobial were non-resistant or resistant.

(b) Competing biological models
The data were analysed using three Markov chain Monte Carlo model formulations. Briefly, the first two formulations are based on the premise that SGI1 or SGI1 variants are circulating in the population in a few fixed forms; Model 1 including only the most common SGI1 variant in addition to the absence of SGI1, and Model 2 permitting all six variants reported in DT104 including the absence of SGI1; variants found in other Salmonella serovars, or other bacteria, were excluded. Variants containing resistance genes to antimicrobials other than the six antimicrobials in the full SGI1 (A, C, Sp, St, Sx and T) were represented only by resistances to the six antimicrobials of interest, with any other resistances being disregarded. Model 3 is conceptually different in that random re-assortment of genetic elements within SGI1 is allowed, as follows. Some resistances occur together: A and Sx are both found on the same integron, with the gene coding for resistance to Sp&St found on a separate integron. However, T and C are not located on the same genetic element as any other resistance gene within SGI1, giving four separate genetic elements within SGI1. The model therefore assumed there to be independent probabilities for the presence of the genetic elements encoding resistance for A&Sx, Sp&St, T and C, producing a total of 16 combinations. These combinations include all but one of the SGI1 variants incorporated into Model 2: the Sp/St/Sx variant is not possible under the assumptions of Model 3 as it requires a functional sul1 gene that is believed to be non-functional in the full SGI1 [17].

Each formulation allows each observed phenotypic resistance, in each isolate, to occur either as a result of an SGI1 variant, or as a result of an individual resistance element elsewhere on the genome. The latter sources of resistance are characterized in each model in terms of the combined probability that one or more such genetic elements exist, and the probability that the phenotype would be expressed given that the genetic element exists. Each resistance element outside of SGI1 is assumed to arise independently of the presence or absence of other phenotypic resistances. A diagrammatic representation of each model is given in figure 1, and mathematical representation in the electronic supplementary material, appendix A. The JAGS code (in the BUGS syntax) for all three models discussed is given in the electronic supplementary material, appendix B.

All three models allow population-level inference to be made on the probabilities of phenotypic expression of SGI1-linked resistances, and combined probabilities of expression
and occurrence of resistances unlinked to SGI1, for each of the six antimicrobials. Models 1 and 2 also allow the prevalence of the full SGI1 variant (that containing resistance genes to all six antimicrobials) to be estimated along with the proportion of DT104 isolates containing no SGI1 variant, with Model 2 additionally allowing the prevalence of another four SGI1 variants to be estimated (identified in table 1). Model 3 also estimates the proportion of DT104 isolates with and without SGI1, but allows re-arrangement of the genetic elements conferring resistance to A&Sx, Sp&St, C and T within SGI1. Separate probabilities for the presence of these four genetic elements, given the presence of SGI1, therefore allows 16 (24) possible SGI1 variants to be modelled.

(e) MCMC simulation
Each model was run using minimally informative Beta(1,1) priors for all probability variables. The probability vector for the multinomial trial with the second model was similarly given a minimally informative Dirichlet(1,1,1,1,1,1) prior. All models were run using six chains with over-dispersed initial values, and run until convergence, as demonstrated by a Gelman–Rubin statistic of less than 1.05. Necessary sample length was calculated using the Raftery and Lewis diagnostic, and the final output of each model was examined for true convergence manually using trace plots. All models were run using JAGS [30] from R [31], automated using the runjags R package [32]. Code for the three models in the BUGS language can be found in the electronic supplementary material, appendix B.

(d) Calculation of posterior predictive p-values
Posterior predictive p-values were calculated using the output of each MCMC model thinned to 1000 iterations. The combinations of parameter values at each iteration were used to obtain 1000 bootstrapped datasets equal in size to the observed sample size. Rather than using an aspect of the data that had been directly modelled, such as the prevalence of resistance to individual antimicrobials, the distribution of the number of unique profiles per bootstrapped dataset (using parameter values from a single iteration) was compared with the observed number of unique profiles. A p-value was then obtained via Monte Carlo integration, reflecting the probability that a dataset as extreme as (or more extreme than) the observed dataset would be obtained using the parameter values at that iteration. This process was repeated for the combination of parameter values at each iteration, and the resultant 1000 p-values along with an ‘overall’ p-value were generated for each model type.

3. RESULTS
Three competing biological models, differing in included SGI1 variants, were fitted to the observed phenotypic data. Posterior predictive p-values were used to compare the abilities of each model to replicate the observed number of distinct resistance phenotypes. The distribution of posterior predictive p-values and histogram of the number of unique profiles produced from the output of each model is shown in figure 2. The overall p-values were 0.006, 0.379 and 0.002 for Models 1, 2 and 3, respectively, indicating that the observed data were extreme for the assumptions underlying Models 1 and 3, but not so for Model 2. The uneven distribution of p-values seen in figure 2 is because of the discrete nature of the observed number of unique profiles, resulting in derived p-values that appear to ‘jump’ between values corresponding to movement of the mean number

Table 1. The presence (+) or absence (−) of six antimicrobial resistance genes (ampicillin, chloramphenicol, spectinomycin, streptomycin, sulphonamides and tetracycline) in six SGI1 variants reported in S. Typhimurium DT104 [18–26], as used in Model 2.

<table>
<thead>
<tr>
<th>Variant</th>
<th>A</th>
<th>C</th>
<th>Sp</th>
<th>St</th>
<th>Sx</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>variant 1</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>variant 2</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>variant 3</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>variant 4</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>variant 5</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>variant 6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
of bootstrapped profiles through each integer threshold relative to the observed number. This statistical artefact does not introduce any bias in the \( p \)-value estimates, and therefore does not affect the validity of the model selection process.

The median and 95 per cent confidence intervals (CI) for each parameter estimated by the best model, Model 2, are shown in Table 2. Variant type 5 (the full SGI1; that with all resistance genes present) was the most prevalent, with types 1 and 2 less common, and types 3 and 4 sufficiently rare that they may not occur in the data. Approximately 4 per cent of the profiles were predicted not to contain any SGI1 variant. The 95 per cent CI for the probability of phenotypic expression of SGI1 genes (denoted ‘\( E \)’) all included 1, reflecting the high prevalence of profiles with phenotypic resistance to all six antimicrobials. The estimates of the probabilities of expression for resistance genes foreign to SGI1 (denoted ‘\( G \)’) were much lower because they also incorporate the probability that a resistance gene foreign to SGI1 occurs, whereas the probabilities of expression for SGI1 genes do not incorporate the prevalence of the genetic elements. The prevalence of the full SGI1 was similar regardless of which model was used, as can be seen from Table 3.

### 4. DISCUSSION

Latent class analysis is a useful way of providing inference on a hidden variable of interest from a related observable feature of the system. To analyse data in this way, an appropriate model must first be selected and then fitted to the data. Posterior checks of model fitness such as the use of posterior predictive \( p \)-values allow the fit of a model to be assessed without depending on likelihood or deviance-based tests. Although usually used to check the model fit to an aspect of the data that is explicitly modelled, such methods also allow the model fit to different aspects of the data to be assessed [11]. This allows additional information to be used that could not be incorporated into the likelihood for the model: in our case, the number of unique resistance profiles to the six

![Figure 2](http://rspb.royalsocietypublishing.org/)
The model fit as reflected by posterior predictive p-values gives a clear indication that Model 2 provides a better representation of the biological system than Models 1 and 3. This suggests that although most of the isolates contain the most prevalent genomic island (the full SGI1), it is not sufficient, at least in terms of the number of unique profiles observed, to consider only this genomic island when modelling the data. The poor fit of Model 3 also suggests that the individual genetic elements within SGI1 are not subject to random re-assembly; rather, the SGI1 variants circulate in the population in a few fixed forms. This conclusion is consistent with the current thinking on the molecular biological aspects of DT104 [17,33]. Although Models 1 and 3 did not reflect the observed data in terms of the number of unique profiles generated, the inferred prevalence of the main genomic island (the full SGI1) was similar in all three models. This is because the improvement in fit provided by Model 2 was largely the result of a reduced number of rare profiles arising from the model; the distribution of the common profile was very similar in all the models because these prevalences heavily influenced the model likelihoods. Indeed, it is for this reason that the number of unique profiles was chosen as the measure on which to base the posterior predictive p-values; a diversity measure that is heavily weighted by the abundance of the most common profile (the reciprocal Berger–Parker measure) was not able to distinguish the three models. Similarly, comparison of the mean and variance of the Hamming distance between bootstrapped isolates was not significantly different to the observed value for any of the models. In contrast, use of the Shannon entropy measure, which more evenly weights the number of unique profiles and profile abundance, gave the same inference regarding fit of the three models as was obtained using the number of unique profiles. However, it should be noted that the utility of posterior predictive p-values is critically affected by the choice of parameter on which to base the estimate, so careful consideration to the most appropriate measure should always be given. The similarity in predicted prevalence of the full SGI1 variant implies that all models adequately reflected certain features of the data and therefore may have produced sufficiently accurate results for some applications. However, the p-values indicate that Model 2 provides the best representation of the biological system in terms of the number of unique profiles observed. This is despite the greater flexibility of Model 3: all but one of the SGI1 variants included in Model 2 is also allowed in Model 3, with an additional 11 possible variants also included in the more flexible model. Based on the results of Model 2, the population from which our data were derived has a 90 per cent prevalence of the most important genomic island variant (the full SGI1), with a 4 per cent prevalence of a variant containing resistances to Sp, St and Sx (SGI1-C), 2 per cent prevalence of a variant containing resistances to A and Sx (SGI1-B), and a 4 per cent prevalence of isolates with no SGI1. The remaining two variants identified in the literature had a vanishingly small prevalence in our data. Although not a central conclusion of this paper, it is also interesting to note that the probability of phenotypic expression of an SGI1-mediated resistance was very close to one for each antimicrobial. This finding is in line with the current thinking about expression of SGI1-mediated resistance genes, which lends further credence to the biological plausibility of Model 2 [34].

Surveillance of antimicrobial resistance in bacterial pathogens is currently limited by the difficulty in inferring the underlying genetic mechanisms of resistance from observed phenotypic data. We acknowledge that genotypic characterization of submitted samples is the gold standard approach, but it is also expensive to perform for large numbers of submitted samples and cannot be used to infer the prevalence of genetic mechanisms of resistance when performed on small numbers of samples. While high-throughput sequencing may provide this information, this technique is not yet widely used in surveillance and is unlikely to be used in a retrospective fashion. It would also need to be applied to very large numbers of isolates, given the very small prevalence of the rarer phenotypes and genotypes, to be comparable to our approach. The current methods of inferring genotypic data from phenotype are intended for analysis of single isolates [14,15], do not quantify the uncertainty involved with predicting the genotype this way and are not a replacement for genomic characterization [15]. Multivariate data analysis has been suggested for analysis of such data [35], but not with the intention to provide inference on the prevalence of genetic elements within the population. The method presented here provides a viable solution that is well suited to analysis of data from S. Typhimurium DT104, which is a clonal organism with well-characterized chromosomally encoded multidrug resistances. Our method will be similarly applicable to datasets on other organisms where there is significant variation between phenotypes and where associations between individual resistances are reasonably well defined, but would be less useful in other settings. The model also requires similar prior information on the possible genetic mechanisms of observed phenotype in order to estimate the prevalence of these mechanisms within the population. For these applications, and others involving latent class mixture models, assessment of model fit using posterior predictive p-values provides an elegant method of model selection that is unaffected by the effective number of parameters or degrees of freedom in the model.

Table 3. The median and 95 per cent CI estimates for the prevalence of the main SGI1 variant using the three models from the analysis of the 2761 human S. Typhimurium DT104 resistance profiles.

<table>
<thead>
<tr>
<th>Model</th>
<th>lower CI</th>
<th>median</th>
<th>upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.889</td>
<td>0.901</td>
<td>0.912</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.887</td>
<td>0.899</td>
<td>0.910</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.891</td>
<td>0.901</td>
<td>0.913</td>
</tr>
</tbody>
</table>
There is a clear application for the methods presented here in the analysis of antimicrobial resistance data. However, the potential value of inferring genotypic data from phenotype has also been recognized in applications such as disease traits in humans [36]. Our methods represent a useful and a generic mechanism of inferring the genetic population structure of organisms, and could also be applied to other types of phenotypic data such as these. Given the comparatively high costs of collecting genetic information compared with that associated with obtaining phenotypic data, this has the potential to provide a valuable and cost-effective inference in a variety of applicable fields.

The authors thank the Scottish Salmonella, Shigella and Clostridium difficile Reference Laboratory and Health Protection Scotland for providing data, and Patrick Boerlin for helpful comments and advice. A.E.M. is supported by the William Stewart Fellowship, University of Glasgow, and M.J.D. is supported by the Scottish Government’s Centre of Excellence in Epidemiology, Population Health and Disease Control (EPIC). The authors are grateful to the anonymous referees for helpful comments and suggestions.

REFERENCES


32 Denwood, M. J. 2008 Runjags: run Bayesian MCMC models in the BUGS syntax from within R. R package version 0.9.2-5. See http://cran.r-project.org/web/packages/runjags.


