BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 18 2008, pages 2030–2036
doi:10.1093/bioinformatics/btn351

Genetics and population analysis

Bayesian survival analysis in genetic association studies

Ioanna Tachmazidou[1,*], Toby Andrew[2], Claudio J. Verzilli[3], Michael R. Johnson[4] and Maria De Iorio[1]

[1]Department of Epidemiology and Public Health, Imperial College, London W2 1PG, [2]Twin Research Unit, King's College, London SE1 7 EH, [3]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT and [4]Division of Neurosciences and Mental Health, Imperial College, London SW7 2AZ, UK

Received on April 9, 2008; revised on June 19, 2008; accepted on July 08, 2008

Advance Access publication July 9, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Large-scale genetic association studies are carried out with the hope of discovering single nucleotide polymorphisms involved in the etiology of complex diseases. There are several existing methods in the literature for performing this kind of analysis for case-control studies, but less work has been done for prospective cohort studies. We present a Bayesian method for linking markers to censored survival outcome by clustering haplotypes using gene trees. Coalescent-based approaches are promising for LD mapping, as the coalescent offers a good approximation to the evolutionary history of mutations.

Results: We compare the performance of the proposed method in simulation studies to the univariate Cox regression and to dimension reduction methods, and we observe that it performs similarly in localizing the causal site, while offering a clear advantage in terms of false positive associations. Moreover, it offers computational advantages. Applying our method to a real prospective study, we observe potential association between candidate ABC transporter genes and epilepsy treatment outcomes.

Availability: R codes are available upon request.

Contact: ioanna.tachmazidou@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Much of the current focus in human genetics is on disentangling the genetic contribution to complex diseases via genetic association studies. Numerous methods have been proposed for the analysis of genetic data from case-control studies, but very little is available for the analysis of time-to-event outcomes, such as patients' overall survival time or time to cancer recurrence.

The most popular approach to modelling survival data is the Cox proportional-hazards regression (Cox, 1972). However, in the context of genetic association studies, Cox regression faces the same problems as common regression which are related mainly to the size of datasets currently being collected and the collinearity between markers that may exist due to linkage disequilibrium (LD). The simplest approach would be to use univariate Cox models to assess the association between each marker and outcome separately. However, univariate analyses can be inefficient as they do not account for the aforementioned statistical correlation or LD between markers, as opposed to multi-marker approaches.

In this article, we propose to tackle these problems (high-dimensionality and multi-collinearity) by clustering haplotypes with similar hazard risks. The proposed method is an extension of approach described in Tachmazidou *et al*. (2007), which deals with case-control data. Here, we assume a parametric model for survival times and search for genetic variants, mostly single nucleotide polymorphisms (SNPs), that show important associations with the survival times. In particular, we scan the chromosomal region of interest for sub-regions of no obligate recombination, or parallel and back mutations. Each sub-region can be represented by a unique evolutionary tree, called gene tree or perfect phylogeny (PP) (Griffiths, 2001), whose topology approximates the mutational history of the haplotypes therein. Coalescent approaches are promising for LD mapping, as the coalescent is likely to provide a better approximation to the evolutionary history of mutations compared to empirical clustering methods. We use a Markov chain Monte Carlo (MCMC) algorithm to iteratively sample from the PPs that make up our genetic region, and we cluster the haplotypes according to the relative ages of the markers in the sampled PP. The main idea behind our clustering metric is that ancestrally similar haplotypes are likely to have similar hazard risks. After convergence, we obtain the posterior probability of each SNP being a cluster centre, and we treat this as the posterior density of the location of a putative causal variant, since high values correspond to markers where haplotypes are best separated, suggesting the presence in the region of a variant influencing the risk of the clinical event. The proposed method is fast and can handle large datasets with many markers and/or patients. Its performance is compared in simulation studies to the univariate Cox regression, and to the dimension reduction methods of Li and Gui (2004), implemented in the software PCRCox, and Bair and Tibshirani (2004) and Bair *et al*. (2006), implemented in the software SUPERPC.

Li and Gui (2004) propose a partial Cox regression (PCR) method that constructs uncorrelated components via repeated least square fitting of residuals and Cox regression fitting. From the resulting

*To whom correspondence should be addressed.

PCR components, the first k most important are determined by univariate Cox regression. Li and Gui (2004) also suggest that using PC analysis to find the non-trivial principle components and then fitting only these using their method, results in a more parsimonious model.

Bair and Tibshirani (2004); Bair *et al*. (2006) propose a semi-supervised form of PC analysis, called Supervised Principle Components (SPC). SPC initially computes univariate Cox regression coefficients, and retains those variables whose coefficients exceed in absolute value some threshold, estimated by cross-validation. Using the reduced dataset, it computes the first few principle components and provides important scores for the initial variables.

In our simulation studies, we consider different scenarios varying in genetic relative risk, minor allele frequency of the causal allele, sample size and censoring. The proposed method yields similar localization performance to the other methods considered, while showing a clear advantage in terms of false positive associations. We applied our approach to data from the SANAD (a study of Standard and New Antiepileptic Drugs) UK prospective study (www.liv.ac.uk/neuroscience/sanad) and found potential associations between candidate ATP-binding cassette (ABC) transporter genes and epilepsy treatment outcomes.

2 METHODS

2.1 Perfect phylogenies

Over genomic regions characterized by strong LD, where there is no evidence of recombination and recurrent point mutations, haplotypes are said to have evolved according to a PP. Haplotypes within these regions can be represented by a unique gene tree that describes their mutational history.

For example, Table 1 represents the incidence matrix for a set of 4000 haplotypes. Columns correspond to six diallelic SNPs and rows are the unique haplotypes in the dataset. Alleles are coded as 0 for the major allele and 1 for the minor allele. A rooted PP assumption poses the constraint that, for any two SNPs in the incidence matrix, not all three combinations (01, 10, 11) exist. In contrast, recombination and back or parallel mutation lead to the possible existence of all three combinations.

When the PP assumption is valid, we can use Gusfield's algorithm (Gusfield, 1991) to construct the gene tree compatible with the data. Figure 1 shows the gene tree for the haplotypes in Table 1. The nodes in the tree correspond to mutations and the gene tree is rooted at the haplotype with all major alleles. Mutations are ordered on the tree according to their relative age. If the causal mutation is embedded between SNPs 2 and 3 say, all descendant haplotypes will inherit it and will therefore have a more recent shared ancestry than the other haplotypes. Thus, the survival associated with the 567 haplotypes that correspond to the last two branches of the tree will tend to be similar. However, the phenotype–haplotype relationship may become more fuzzy due to dominance, epistasis or the effect of environmental factors.

Therefore, the use of the PP model implies little or no recombination in genomic segments and the 'infinitely many sites' assumption from population genetics. We propose to split the chromosomal region of interest into consecutive PPs with window boundaries deterministically defined by the locations where the PP assumption breaks. Details of how this is implemented are given in Tachmazidou *et al*. (2007). Once the set of windows and corresponding trees have been identified, we use a Bayesian partition model to search through trees to identify those, if any, where the corresponding set of haplotypes appear to form clusters that discriminate high from low-hazard risk, thus possibly harbouring a causal variant.

Table 1. Incidence matrix for six distinct haplotypes and their frequencies, consisting of six SNPs (S_1–S_6)

Frequency	S_1	S_2	S_3	S_4	S_5	S_6
204	1	1	0	0	0	0
932	1	1	0	0	1	1
233	1	1	0	0	1	0
2064	0	0	0	0	0	0
565	0	1	1	1	0	0
2	0	1	1	0	0	0

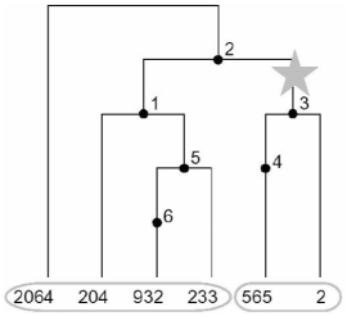

Fig. 1. The gene tree consistent with the haplotypes in Table 1. Labels 1–6 refer to mutations S_1–S_6. At the bottom of each branch, we report the multiplicity of each observed haplotype in the sample.

2.2 Haplotype clustering

We use a Bayesian partition model to split the haplotype space into an a priori unknown number of disjoint clusters on the basis of haplotype similarity. Our similarity measure has an evolutionary interpretation, with sequences sharing a cluster depending on the time to their most recent common mutation. In other words, our distance metric is based on the order with which the mutations have arisen in the haplotype sample. This information is provided by the topology of the gene tree. Therefore, any SNP set selected as cluster centres can be time ordered and we assign haplotypes to clusters according to the relative ages of these centres.

Suppose, for example that SNPs 1, 3 and 5 of Figure 1 are selected as cluster centres. SNP 5 is younger than SNP 1, and SNP 3 is on a different branch, implying that a haplotype carrying mutation 3 cannot carry mutation 1 or 5. Starting with SNP 5, we assign the haplotypes that correspond to the third and fourth branch of the tree as members of this cluster. Only the second haplotype is assigned to the cluster with SNP 1 as centre because, although the third and fourth haplotypes carry mutation 1, they have been already allocated to a cluster. The last two haplotypes are allocated to a separate cluster with centre SNP 3, and the remaining haplotype is assigned to a hypothetical 'null' cluster, which can be interpreted as a baseline risk group. Therefore, every haplotype is deterministically allocated to the cluster with the closest centre. Haplotypes within each cluster are assumed to have similar survival probabilities and risks.

2.3 Modelling approach

Let us assume that the haplotype data can be split into n_{tr} PP or gene trees, that tree T is selected as harbouring the causal mutation and that the haplotype space is currently partitioned into $n_c = n_{clust} + 1$ clusters (n_c includes the 'null' cluster, while n_{clust} is the number of SNPs selected as cluster centres). Conditionally on tree T, an indicator vector $\boldsymbol{\gamma}$ represents the partition, with $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{n_{SNP_T}})$, $\gamma_k \in (0, 1)$, $k = 1, \ldots, n_{SNP_T}$, such that $\gamma_k = 1$ if the k-th

SNP is selected as cluster centre and $\gamma_k = 0$ otherwise, where n_{SNP_T} is the number of SNPs in T.

For a failure time X, right-censored data can be represented by pairs of random variables (t, δ). When the lifetime X is observed, $\delta = 1$ and t is equal to X, whilst for right-censored data, $\delta = 0$ and t is equal to the censoring time C_r, i.e. $t = \min(X, C_r)$. Therefore, $\delta_{ij} \in \{0, 1\}$ is the censoring status indicator of survival time t_{ij} for haplotype $i = 1, \ldots, n_j$ in cluster $j = 1, \ldots, n_c$. The vector of responses for cluster j is denoted by $\boldsymbol{D}_j = \{(t_{1j}, \delta_{1j}), \ldots, (t_{n_j j}, \delta_{n_j j})\}'$ and let $\boldsymbol{D} = \{\boldsymbol{D}_j, j = 1, \ldots, n_c\}$. Each t_{ij} is assumed to have an exponential distribution with cluster-specific parameter θ_j. Thus, the likelihood of the data is:

$$L = \prod_{j=1}^{n_c} \theta_j^{\sum_{i=1}^{n_j} \delta_{ij}} \exp\left(-\theta_j \sum_{i=1}^{n_j} t_{ij}\right),$$

where $\sum_{i=1}^{n_j} t_{ij}$ is the total time in the study for all n_j haplotypes in cluster j, and $\sum_{i=1}^{n_j} \delta_{ij}$ is the observed number of events (e.g. deaths) in cluster j, with $j = 1, \ldots, n_c$. Parameters θ_j are given independent Gamma(ν_0, ν_1) priors. The marginal probability of the data is available analytically and given by

$$p(\boldsymbol{D}|\boldsymbol{\gamma}, \nu_0, \nu_1) = \left[\frac{\nu_1^{\nu_0}}{\Gamma(\nu_0)}\right]^{n_c} \prod_{j=1}^{n_c} \frac{\Gamma\left(\nu_0 + \sum_{i=1}^{n_j} \delta_{ij}\right)}{(\nu_1 + \sum_{i=1}^{n_j} t_{ij})^{\nu_0 + \sum_{i=1}^{n_j} \delta_{ij}}} \quad (1)$$

where Γ denotes the Gamma function.

A priori each tree is equally likely to contain the putative mutation. Conditionally on tree T, we impose a binomial prior on the number of cluster centres n_{clust}. Given n_{clust}, any cluster configuration is equally likely a priori. Then the joint prior distribution of tree T and partition $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} p(T, \boldsymbol{\gamma}) &= p(T)p(\boldsymbol{\gamma}|T) = p(T)p(n_{clust}|T)p(\boldsymbol{\gamma}|n_{clust}, T) \\ &= \frac{1}{n_{tr}} \binom{n_{SNP_T}}{n_{clust}} q^{n_{clust}} (1-q)^{n_{SNP_T} - n_{clust}} \frac{1}{\binom{n_{SNP_T}}{n_{clust}}}, \end{aligned} \quad (2)$$

where q is the success probability of the Binomial distribution, and can be chosen to penalize or favour big number of clusters in such a way to reflect the investigator's prior beliefs.

The posterior conditional distribution of θ_j is also available in closed form

$$\theta_j|\boldsymbol{D}_j, \boldsymbol{\gamma}, \nu_0, \nu_1 \sim \text{Gamma}\left(\nu_0 + \sum_{i=1}^{n_j} \delta_{ij}, \nu_1 + \sum_{i=1}^{n_j} t_{ij}\right), \forall j = 1, \ldots, n_c.$$

Upon convergence, we obtain a posterior sample of partitions and an estimate of the posterior probability that the causal mutation is embedded within each of the trees. We then calculate the Bayes factor in favour of association at each marker site as the ratio of the posterior to prior odds (Kass and Raftery, 1995), where the prior odds of each SNP being a cluster centre is evaluated by simulation using Equation (2).

2.4 Details of the MCMC algorithm

We use an MCMC algorithm to sample from the posterior distribution over the space of possible partitions. At each MCMC iteration, we perform two M-H steps:

- Change partition: sample a new partition from the posterior distribution of the cluster centres without changing the current gene tree.
- Update tree: sample a new tree and a new partition from their joint posterior distribution.

In particular, we use a Metropolis–Hastings (M–H) step to sample from the conditional distribution of partition $\boldsymbol{\gamma}$ given the data and tree T. We consider two possible moves in the partition space: adding (birth) or deleting (death) a cluster centre. At each move, we randomly select SNP i and we propose $\gamma_i^\star = 1$, if the current $\gamma_i = 0$ or $\gamma_i^\star = 0$ otherwise. Thus, the proposal distribution $q(\boldsymbol{\gamma}^\star|\boldsymbol{\gamma})$ is simply $1/n_{SNP_T}$. Given the cluster centres, the observed haplotypes are deterministically allocated to the haplotype clusters according to our similarity metric. The logarithm of the acceptance probability for the first M-H sampler ($\mathtt{a}_1$) simplifies to the logarithm of the Bayes factor (BF) in favour of $\boldsymbol{\gamma}^\star$ over $\boldsymbol{\gamma}$, i.e. $p(\boldsymbol{D}|\boldsymbol{\gamma}^\star)/p(\boldsymbol{D}|\boldsymbol{\gamma})$ where the marginal probability is calculated using Equation (1) $\pm \log((1-q)/q)$ depending if the death or birth of a cluster centre is proposed respectively, i.e.

$$\log(\mathtt{a}_1) = \log\left(\text{BF}(\boldsymbol{\gamma}^\star, \boldsymbol{\gamma})\right) \pm \log\left[\frac{1-q}{q}\right] = \log\left[\frac{p(\boldsymbol{D}|\boldsymbol{\gamma}^\star)}{p(\boldsymbol{D}|\boldsymbol{\gamma})}\right] \pm \log\left[\frac{1-q}{q}\right].$$

We use a second M-H step to sample a tree from the n_{tr} possible trees with probability $1/n_{tr}$ and a new partition given the tree. Each SNP in the tree has a 0.5 probability of being proposed as a cluster centre. Therefore, the proposal probability of the tree and the partition space is equal to $1/(n_{tr} \times 2^{n_{SNP_T}})$, and the joint prior distribution of a gene tree T and a partition $\boldsymbol{\gamma}$ is given by Equation (2). The logarithm of the acceptance probability for the second M-H sampler ($\mathtt{a}_2$) simplifies to the logarithm of the Bayes factor in favour of the proposed partition over the current partition plus the logarithm of the prior ratio and the logarithm of the proposal ratio, i.e.

$$\log(\mathtt{a}_2) = \log\left(\text{BF}(\boldsymbol{\gamma}^\star, \boldsymbol{\gamma})\right) + \text{K} = \log\left[\frac{p(\boldsymbol{D}|\boldsymbol{\gamma}^\star)}{p(\boldsymbol{D}|\boldsymbol{\gamma})}\right] + \text{K}, \text{ with}$$

$\text{K} = (n_{SNP_T}^{prop} - n_{SNP_T}^{curr})\log(2) + (n_{clust}^{prop} - n_{clust}^{curr})\log(q) + (n_{SNP_T}^{prop} - n_{clust}^{prop} - n_{SNP_T}^{curr} + n_{clust}^{curr})\log(1-q)$, where $n_{SNP_T}^{prop}$ and $n_{SNP_T}^{curr}$ is the number of SNPs in the proposed and current tree respectively, and n_{clust}^{prop} and n_{clust}^{curr} is the number of clusters in the proposed and current tree, respectively.

3 RESULTS

3.1 Simulation study

The performance of the proposed method was compared to that of the univariate Cox regression, the PCRCox and SUPERPC software. PCRCox and SUPERPC were mainly developed for analysing gene expression data. Here, we explore their performance when searching for marker-survival phenotype associations.

We simulated a population of 20 000 haplotypes over 1 Mb chromosomal region with recombination hotspots using the FREGENE software (Hoggart *et al.*, 2007) under the default parameters. From the simulated data, we retained markers with minor allele frequency (MAF) $\geq 1\%$. From these markers, we selected 1000 SNPs with probability proportional to $p(1-p)$, where p is the allele frequency of a marker in the sample. In this way, we obtain an average spacing of 1 kb. We then selected at random a SNP with allele frequency between $p - 0.005$ and $p + 0.005$, where p was in a range between 0.02 and 0.3, to be the causal locus.

We sampled with replacement pairs of haplotypes to form an individual's genotype and generated their survival times from a Gompertz(α, β) distribution (with $\alpha = 1, \beta = 1$). The Gompertz distribution (Gompertz, 1825) is a popular probability model for human mortality and it is a proportional hazards model like the Cox model.

Assuming additive genotype effect, the hazard function is given by $h(t|G_i) = \beta e^{\alpha t} e^{\rho G_i}$, where G_i is the number of copies of the causal allele, $G_i = 0, 1, 2$, and ρ is the logarithm of the hazard relative risk of the heterozygote, HRR(Aa). We varied the relative risk between 1.2 and 3. The censoring times were generated from an exponential distribution with mean equal to $t^\star = 1/\lambda$. To evaluate $t^\star$, we assumed that 5% of the whole population experiences the disease by time $t^\star$, i.e. $P(t \leq t^\star) = P(t \leq t^\star|G_0)(1-p)^2 + P(t \leq t^\star|G_1)2p(1-p) + P(t \leq t^\star|G_2)p^2 = 5\%$, where $P(t \leq t^\star)$ is the cumulative distribution function of the survival times, and p is the allele frequency of the causal variant. We considered

a censoring level of 30, 50, 70 and 90%, and sample sizes of 400, 1000, 2000 and 4000 individuals, which at the 90% censoring level yielded 40, 100, 200 and 400 disease cases, respectively.

In all analyses, we removed the causal allele from the dataset and, using the technique described in the 'Perfect phylogenies' section, we constructed the perfect phylogenies in the dataset. The average number of gene trees was 200 and the average number of SNPs in a gene tree was 4.

The MCMC algorithm was run for 100 000 iterations with a burn-in of 10 000 iterations for 50 datasets under different combinations of the simulation parameters. For a dataset of 1000 markers and 4000 haplotypes the proposed method took 14 min to construct the phylogenies and 13 min to run the algorithm on an Intel Xeon 3.40 GHz processor with 2 Gb of memory. The computing time for PCRCox and SUPERPC was 4 and 59 min respectively. An R (www.r-project.org) package called BETA-Surv (Bayesian Evolutionary Tree-based Association analysis for Survival) implementing the proposed method is available on request from the first author.

In the simulated and real data examples, we set the prior hyperparameters as follows: we assume a Gamma$(0.1, 0.1)$ prior on θ_j, $j = 1, \ldots, n_c$, and a Binomial$(n_{SNP_T}, p = 0.98)$ for n_{clust} given tree T. Such a high success probability of the Binomial reflects an a priori belief that each haplotype in the tree has its own risk.

3.2 Model performance

We investigate the performance of BETA-Surv, univariate Cox, PCRCox and SUPERPC in terms of localization, power and false positive rates. To run PCRCox, we first performed PC analysis and we only used the significant principle components (i.e. those whose SD is bigger than 10^{-10} in absolute value), as suggested in the software documentation. We were also advised (Jiang Gui, personal communication) that the first five principle components were usually sufficient. From these, we estimated regression coefficients of the original variables. SUPERPC returns importance scores for each of the significant variables, which we use to estimate associations. In order to estimate the best threshold, SUPERPC computes a cross-validated likelihood ratio (LR) statistic using the first, the first two or the first three principle components. Because for most of the simulated datasets none of the three LR tests were significant, we calculated scores using three different thresholds (i.e. using the first, the first two or the first three principle components), and we took the mean value of the scores for each significant SNP. Results reported for each simulation scenario are averages over 50 replicates. However, SUPERPC spuriously failed to produce results on approximately half of the datasets under each scenario. Therefore, for this method results are over approximately 25 replicates.

To measure each method's accuracy in estimating the position of an untyped causal allele, we report the probability that the identified location is within some distance from the true location. The position of the susceptibility variant is estimated by the physical location of the SNP with the maximum Bayes factor for the proposed approach and with the minimum P-value for univariate Cox regression. For PCRCox and SUPERPC the causal location is estimated by the SNP with the highest absolute regression coefficient or score, respectively.

To determine power, we define a window of 100 kb either side of the causative allele and calculate the proportion of the 50 replicates having a significant association within the window. The significance of a signal is assessed via the following rules. For BETA-Surv, a SNP is considered to have positive signal when its Bayes factor in favour of association is ≤ 3 (Kass and Raftery, 1995). For univariate Cox regression, we use two different significance thresholds, i.e. a P-value ≤ 0.05, and the Bonferroni-adjusted value. For PCRCox and SUPERPC, a SNP is regarded as a positive hit if its regression coefficient or score, in absolute value, is larger than or equal to the upper 90% quantile of the absolute value of the regression coefficients or scores, respectively. We also define as false positives those markers with smaller P-values, larger Bayes factors, or smaller absolute regression coefficients or scores lying outside a window of 100 kb either side of the causal variant.

3.3 Results across a range of simulation scenarios

Figure 2 shows the localization performance of the different methods. Results are averages over 50 simulated datasets under a default scenario, with 2000 sample size, 1.6 HRR(Aa), MAF of causal allele 5% and 50% censoring. Overall, BETA-Surv and univariate Cox perform better than PCRCox and SUPERPC. For BETA-Surv and univariate Cox, there were no significant differences in the distribution of distances for reasonable location errors, with univariate Cox showing advantage for distances >100 kb.

Figure 3 reports the power and false positive rates of the methods for the 50 datasets simulated under the default scenario. Univariate Cox and PCRCox have the highest power, but the worst performance in false positives. BETA-Surv and SUPERPC have similar power, but BETA-Surv has the lowest false positive rate, which is almost as low as Bonferroni-corrected univariate Cox.

In Supplementary Figure 1, we plot the power and false positive rates of the different methods across the various simulation scenarios and over the 50 replicates. In each plot, we vary a simulation parameter along the x-axis whilst assuming default values for the remaining ones. Uncorrected single locus test and PCRCox are the most powerful approaches, having however the worst performance in terms of false positives. The false positive rates for BETA-Surv are very low under all simulation scenarios. The same conclusions apply when we vary the censoring level. The choice of a 100 kb window is sensible but arbitrary; a 200 kb window was also investigated and

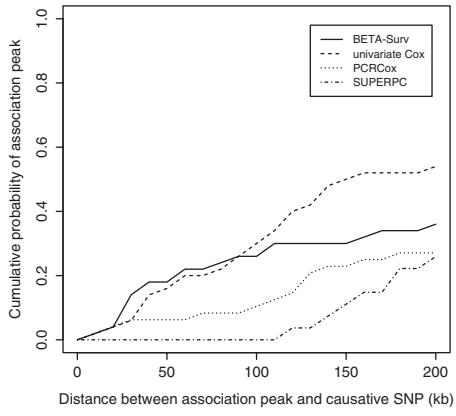

Fig. 2. Distribution of distances between the association peak and the causal SNP. Analysis of 50 datasets simulated with sample size equal to 2000, 1.6 HRR(Aa), MAF of causal allele 5%, and 50% censoring.

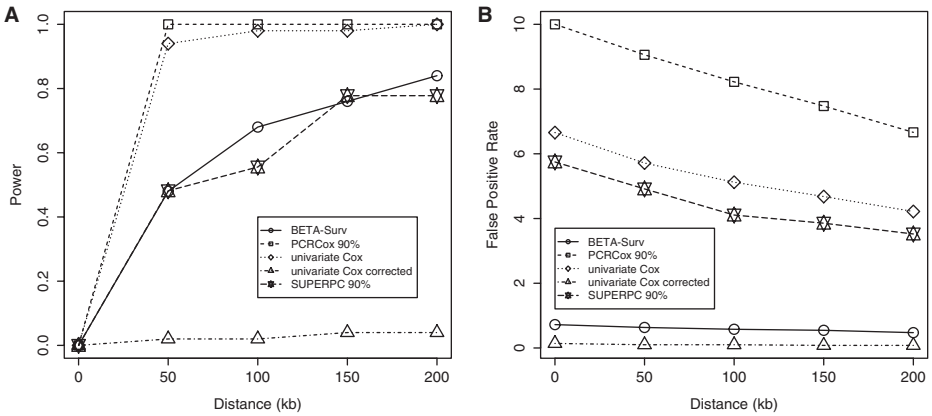

Fig. 3. Probability of detecting true associations (**A**) and false positive rate (**B**). True positives are significant SNPs within some physical distance with the causal allele, otherwise they are classified as false positives. Analysis of 50 datasets simulated with sample size equal to 2000, 1.6 HRR(Aa), MAF of causal allele 5%, and 50% censoring.

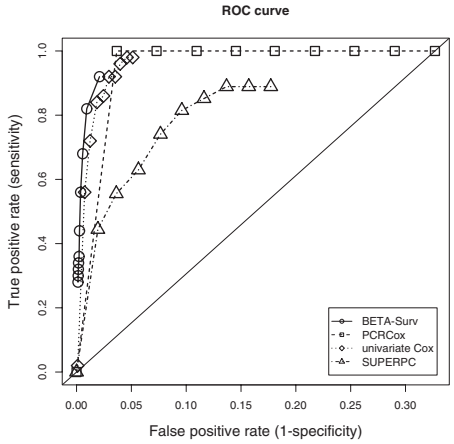

Fig. 4. Power versus false positive rate for different significance thresholds. Analysis of 50 datasets simulated with sample size equal to 2000, 1.6 HRR(Aa), MAF of causal allele 5%, and 50% censoring.

did not alter conclusions about both power or false positives. Also, using the upper 95% quantile of the absolute standardized regression coefficients or scores for PCRCox and SUPERPC, instead of the 90%, results in lower power and lower false positive rates for these methods, but BETA-Surv has still by far the best performance in terms of false positives.

To show the trade-off between power and false positive rate for the different methods using varying significance thresholds, we plot ROC curves for the different simulation scenarios (Fig. 4 and Supplementary Figs 2–5). The threshold for univariate Cox regression ranges between Bonferroni correction and 0.05, for BETA-Surv from 1.1 to 10, and for PCRCox and SUPERPC from 60% to 100%. The graphs show that BETA-Surv and univariate Cox regression are the best performing methods. Compared to univariate Cox regression, BETA-Surv has a small advantage in power when the association signal is weak, while it performs similarly for stronger signals. However, it has lower false positive rate in every simulation scenario. Overall, SUPERPC has the worst performance in power, whereas PCRCox has the worst performance in false positive associations.

Finally, we constructed 50 datasets under a null model of no disease association and we calculated the false positive rate. For the univariate analysis, this was 5.39% (P-value ≤ 0.05) and 0 when using Bonferroni correction, while BETA-Surv resulted in a false positive rate of 0.634% (Bayes factor ≥ 3). PCRCox and SUPERPC had a 10% and 4.80% false positive rate using the upper 90% quantile, and 5% and 2.46% using the upper 95% quantile, respectively.

3.4 Prospective epilepsy study

The ABC proteins are a superfamily of efflux pumps that extract several classes of drugs from the eukaryotic cell. The ABC transporters are currently the focus of a major effort to determine their role in mediating drug resistance in a variety of human diseases including cancer (Ambudkar *et al*., 2003), HIV (Sankatsing *et al*., 2004) and epilepsy (Schmidt and Loscher, 2005).

Retrospective studies have reported associations between epilepsy treatment outcome and drug transporter genes (Siddiqui *et al*., 2003; Zimprich *et al*., 2004), using individual SNPs or 3-SNP haplotypes. However, results have been contradictory (Sills *et al*., 2005; Tan *et al*., 2004). Leschziner *et al*. (2006b) analyzed data from the UK SANAD prospective study for the gene complex ABCB1/ABCB4. They used single SNP log-rank tests, 3-SNP haplotype analyses and Cox multiple regression with stepwise selection on a subset of the genotypes (due to the problems of SNP collinearity and over-fitting), and observed no significant genetic association.

Here, we use our proposed method to simultaneously analyze five drug transporters genes. For a prospective cohort of 503 epilepsy patients from the SANAD study, 500 potential SNPs were genotyped across five ABC transporter genes (ABCB1/ABCB4, ABCC1, ABCC2 and ABCC5 located in 7q21.12, 16p13.1, 10q24 and 3q27 respectively). Details of genotyping, SNP identification and LD structure are given in Leschziner *et al*. (2006a). Of the 500 loci identified, only 317 were polymorphic with $\sim$60% of SNPs with MAF $\geq$5%. SNPs with $\leq$1% MAF, showing evidence

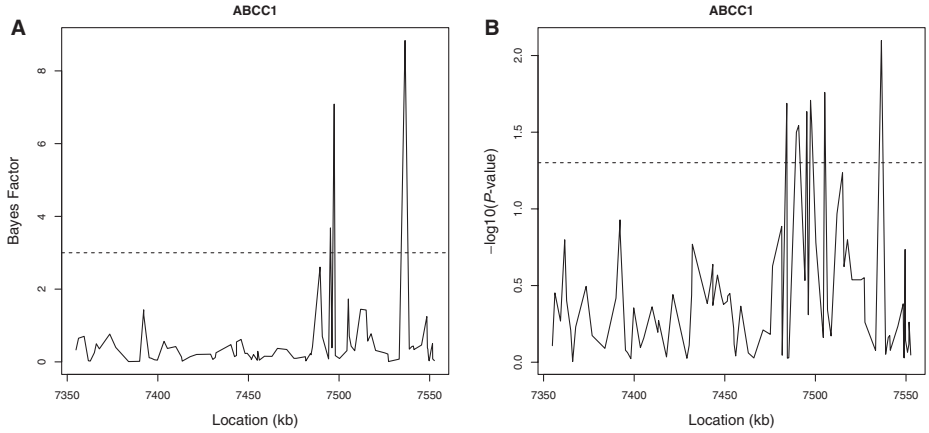

Fig. 5. Results for gene ABCC1 from BETA-Surv and univariate Cox regression for the SANAD epilepsy cohort with outcome 'time to 12 month remission'. Bayes factor in favour of association at each marker from BETA-Surv (**A**), and *P*-values from Cox regression (**B**) averaged over the 10 datasets.

Table 2. Summary of ABC genotype SANAD data

Gene	size (kb)	SNPs	Mono	MAF<1%	HWD	Miss>10%	Used
ABCB1/4	283	186	73	22	3	5	85
ABCC1	193	162	50	12	5	3	92
ABCC2	69	76	32	19	2	3	24
ABCC5	98	76	28	8	4	2	34
Total	643	500	183	61	14	13	235

of Hardy Weinberg disequilibrium (HWD $\chi^2_1 \geq 12$) or with $\geq$10% missing per SNP were discarded, leaving a total of 235 SNPs for analysis (Table 2). Treatment outcomes were prospectively recorded for patients commencing anti-epileptic drug therapy. Here we concentrate on two outcomes: time to 12 month remission (512 observed and 494 censored events), and time to withdrawal from drug due to unacceptable adverse side-effects (194 observed and 812 censored events).

We used PHASE (Stephens *et al*., 2001) to phase the genotypes. Each haplotype pair was chosen at random according to its posterior probability. To account for phase uncertainty, we repeated the above procedure 10 times and we analyzed each of the 10 datasets separately. Gene region ABCB1/ABCB4 consists of 54 trees, and genes ABCC1, ABCC2 and ABCC5 of 67, 14 and 27 trees, respectively averaged over the 10 datasets.

For time to 12 month remission, the proposed method yielded no evidence of association for ABCB1/ABCB4 and ABCC5 regions for any of the datasets. For gene ABCC1, markerwise Bayes factors from BETA-Surv and (−log)*P*-values from univariate Cox regression, averaged over the 10 datasets, are given in Figure 5. Positive hits from BETA-Surv were observed at 7495.248, 7497.311 and 7536.426 kb with average Bayes factors 3.68, 7.09 and 8.84, respectively. For gene ABCC2, both BETA-Surv and Cox regression identified the variant at position 20312.532 kb (average Bayes factor 4.08 and *P*-value 0.03).

For time to withdrawal due to adverse side-effects, the proposed method yielded no evidence of association for ABCB1/ABCB4 and ABCC2 regions for any of the datasets. For gene ABCC1, BETA-Surv found positive association at 7548.444 kb with average Bayes factor 3.36, whereas Cox regression reported the SNP at 7549.729 kb with average *P*-value 0.026. For gene ABCC5, both BETA-Surv and Cox regression identified the variant at position 90194.67 kb in all 10 datasets (average Bayes factor 5.1 and *P*-value 0.03).

Generally, BETA-Surv and Cox regression yielded similar results. There is positive but not strong evidence of association between some ABC transporter genes and epilepsy treatment outcomes. To the best of our knowledge, there are no other reported analyses or associations between genes ABCC1, ABCC2 or ABCC5 and epilepsy.

4 DISCUSSION

We have presented a method to analyze genetic association studies with time-to-event outcomes. Cohort studies are a useful and increasingly common study design, but there is a noticeable lack of statistical methods for their analysis. The method presented here is best suited for densely genotyped candidate gene regions and can easily handle large number of individuals and markers. Compared to univariate Cox regression and multi-marker dimension reduction techniques, our method performs similarly in terms of localization, while offering clear advantages in terms of false positive associations. Moreover, it runs fast and it offers computational advantages especially over methods that rely on cross-validation to determine model parameters.

Here, we assume that survival times can be modelled parametrically by the exponential distribution within each cluster. The use of such a simple distribution may seem restrictive, but offers computational advantages over more complicated models. We have also used the Weibull distribution to model the survival outcome, which in simulation studies offered no significant additional advantages, while increasing the running times of the method.

The incorporation of environmental covariates in the model is an issue that has not been investigated in this work. One possible way of dealing with this, is by fitting a cluster-specific survival regression using the exponential distribution.

Finally, phase uncertainty in haplotype reconstruction from genotype data could be incorporated in the analysis by adding another step in the MCMC algorithm and sampling from the

different haplotype reconstructions at each MCMC step before performing the rest of the analysis for the chosen phase. However, this approach is likely to add significant computational burden. A simpler approach, and one we adopt in the application to the real data, consists in repeating the analysis for a number of different haplotype reconstructions and average the results.

ACKNOWLEDGEMENT

We wish to thank Clive Hoggart for the FREGENE population, and Guy Leschziner for collecting the drug transporter data.

Funding: This work was supported by the Wellcome Trust.

Conflict of Interest: none declared.

REFERENCES

Ambudkar,S.V. *et al*. (2003) P-glycoprotein: from genomics to mechanism. *Oncogene*, **22**, 7468–7485.

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods for predicting patient survival from gene expression data. *PLoS Biol*., **2**, 511–522.

Bair,E. *et al*. (2006) Prediction by supervised principal components. *JASA*, **101**, 119–137.

Cox,D. (1972) Regression models and life tables. *J. R. Stat. Soc. B*, **34**, 187–220.

Gompertz,B. (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philos. Trans. R. Soc. Lond.*, **115**, 513–585.

Griffiths,R.C. (2001) Ancestral inference from gene trees. In Donnelly,P. and Foley,R.A. (eds), *Genes, Fossils, and Behaviour: an Integrated Approach to Human Evolution.* NATO Science Series A, Life Sciences, IOS Press, Amsterdam, pp. 137–172.

Gusfield,D. (1991) Efficient algorithms for inferring evolutionary trees. *Networks*, **21**, 19–28.

Hoggart,C. *et al*. (2007) Sequence-level population simulations over large genomic regions. *Genetics*, **177**, 1725–1731.

Kass,R.E. and Raftery,A.E. (1995) Bayes factors. *JASA*, **90**, 773–795.

Leschziner,G. *et al*. (2006a) Exon sequencing and high resolution haplotype analysis of ABC transporter genes implicated in drug resistance. *Pharmacogenet. Genomics*, **16**, 439–450.

Leschziner,G. *et al*. (2006b) Clinical factors and ABCB1 polymorphisms in prediction of antiepileptic drug response: a prospective cohort study. *Lancet Neurol.*, **5**, 668–676.

Li,H. and Gui,J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, **20**, 208–215.

Sankatsing,S.U. *et al*. (2004) P glycoprotein in human immunodeficiency virus type 1 infection and therapy. *Antimicrob. Agents Chemother.*, **48**, 1073–1081.

Schmidt,D. and Loscher,W. (2005) Drug resistance in epilepsy: putative neurobiologic and clinical mechanisms. *Epilepsia*, **46**, 858–877.

Siddiqui,A. *et al*. (2003) Association of multidrug resistance in epilepsy with a polymorphism in the drug-transporter gene ABCB1. *N. Engl. J. Med.*, **348**, 1442–1448.

Sills,G.J. *et al*. (2005) Lack of association between the C3435T polymorphism in the human multidrug resistance (MDR1) gene and response to antiepileptic drug treatment. *Epilepsia*, **46**, 643–647.

Stephens,M. *et al*. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Tachmazidou,I. *et al*. (2007) Genetic association mapping via evolution-based clustering of haplotypes. *PLoS Genet.*, **3**, e111.

Tan,N.C. *et al*. (2004) Failure to confirm association of a polymorphism in ABCB1 with multidrug-resistant epilepsy. *Neurology*, **63**, 1090–1092.

Zimprich,F. *et al*. (2004) Association of an ABCB1 gene haplotype with pharmacoresistance in temporal lobe epilepsy. *Neurology*, **63**, 1087–1089.