Integration of association statistics over genomic regions using Bayesian adaptive regression splines

Xiaohua Zhang,[1] *Kathryn Roeder,*[1]* *Garrick Wallstrom*[1] *and Bernie Devlin*[2]

[1]Department of Statistics, Carnegie Mellon University, Pittsburg, PA 15213, USA
[2]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
**Correspondence to*: Tel: +412 268 2513; Fax: +412 268 7828; E-mail: roeder@stat.cmu.edu

Date received (in revised form): 8th September 2003

Abstract

In the search for genetic determinants of complex disease, two approaches to association analysis are most often employed, testing single loci or testing a small group of loci jointly via haplotypes for their relationship to disease status. It is still debatable which of these approaches is more favourable, and under what conditions. The former has the advantage of simplicity but suffers severely when alleles at the tested loci are not in linkage disequilibrium (LD) with liability alleles; the latter should capture more of the signal encoded in LD, but is far from simple. The complexity of haplotype analysis could be especially troublesome for association scans over large genomic regions, which, in fact, is becoming the standard design. For these reasons, the authors have been evaluating statistical methods that bridge the gap between single-locus and haplotype-based tests. In this article, they present one such method, which uses non-parametric regression techniques embodied by Bayesian adaptive regression splines (BARS). For a set of markers falling within a common genomic region and a corresponding set of single-locus association statistics, the BARS procedure integrates these results into a single test by examining the class of smooth curves consistent with the data. The non-parametric BARS procedure generally finds no signal when no liability allele exists in the tested region (ie it achieves the specified size of the test) and it is sensitive enough to pick up signals when a liability allele is present. The BARS procedure provides a robust and potentially powerful alternative to classical tests of association, diminishes the multiple testing problem inherent in those tests and can be applied to a wide range of data types, including genotype frequencies estimated from pooled samples.

Keywords: *association study, adaptive regression splines, complex disease, genome scan, linkage disequilibrium (LD), non-parametric regression*

Introduction

The hunt is on for genetic variants that increase the risk for complex diseases, such as type 2 diabetes and schizophrenia. Methods to detect these liability alleles, however, are at a crossroads. Most tests of association between disease status and marker alleles have targeted one or a few markers within a candidate gene. With the advent of large-scale single nucleotide polymorphism (SNP) discovery and relatively inexpensive genotyping, the trend is to target large genomic regions surrounding selected genes, substantially larger regions defined by linkage signals,[1] or even the entire genome.[2] For human populations, linkage disequilibrium (LD) typically extends only over a narrow region surrounding a liability locus.[3,4] Thus, it might require tens of markers to evaluate the region around a gene for association, a much larger number of markers to interrogate a linkage region and orders of magnitude more markers to scan the genome.[5,6] As the cost of genotyping plummets, however, massive genotyping to accomplish fine-scale screening is no longer unfathomable.

For the data analyst, the challenge presented by such massive datasets should not be underestimated. Even the scale of the problem remains nebulous.[7] Any way you look at it, however, the problem is large. Imagine performing a genome scan with $N = 300,000$ SNPs. One could perform N single-locus tests, and make appropriate correction for multiple testing. The concern raised by this simple approach is that the sample size is more than an order of magnitude smaller than the number of SNPs in the genome, even ignoring other genetic variation that could have an impact on liability to disease. Moreover, while in expectation LD between liability alleles and marker alleles declines smoothly with distance under some simple models of evolution, in fact the pattern of pairwise LD is known to be highly variable in the human genome, so much so that it often appears erratic.[3,4]

The nature of pairwise LD has inspired the investigation of higher level LD structure, such as that embodied by haplotypes. Results from the genomic analysis of haplotypes do indeed look promising, in that LD at higher levels of dependence is much more predictable.[8–11] From this observation sprang the HapMap project, which has as its goal to define the haplotype structure of the human genome and to identify the SNPs needed to 'tag' haplotypes. Whether higher-level LD will turn out to be sufficiently predictable to streamline the discovery process for liability alleles is unclear,[12] and it is expected that it will probably depend on the nature of the population sampled. Even in the best of circumstances, however, there remains an abyss between theory and practice: different analytical methods lead to different fine-scale haplotype structure in the genome. This can be taken to mean that higher-level LD is by no means absolute, and thus a multitude of different analyses will be required to ensure adequate testing for association. Adding to the complexity, it is not even clear if haplotype-based tests of association are more powerful than a series of single-locus tests. Not surprisingly, it appears that the answer depends strongly on the local patterns of LD.[13–16]

Another wrinkle to the problem is the type of genotyping performed on the sample. Obviously, molecular haplotyping of some kind provides the maximum amount of information about the LD in a region, per subject, but the molecular methods can be expensive. When individuals within families are genotyped at multiple loci, haplotype structure often can be inferred without error, but collection of the sample can be expensive. Usually less expensive are samples consisting of unrelated individuals, but then some information about haplotype structure is lost (albeit less than one might think: cf. Schaid[17] and Douglas *et al.*[18]) Pooled genotyping, however, offers the most economical approach for obtaining genotypes but the accuracy of haplotype reconstruction fades quickly as the number of samples comprising the pool increases.

The situation for the gene hunter is therefore perplexing. Single-locus tests suffer from correction for multiple testing, and cannot be guaranteed to be effective, even as the sample size tends to infinity, because the tested marker alleles might not be in LD with critical liability alleles. Haplotype-based tests capture more of the LD structure of a genomic region, and thus could be more efficient than single-locus tests, but the question of which haplotypes to test raises the spectre of very large corrections for multiple testing when large genomic regions are evaluated.

A single best recipe for hunting liability alleles is unlikely to exist. In some circumstances, it may be best to combine information over single markers in some computationally efficient way, to discover target regions. Once identified in a preliminary manner, those regions of the genome that appear to harbour liability alleles would be ideal for more refined fine-scale haplotype tests. In this paper, methods to combine information over individual markers are explored. The authors' analyses exploit the fact that LD between a liability allele and marker alleles is expected to decline with distance. Thus, it might be reasonable to fit a smooth function to the data, looking for regions with a consistent overall pattern of LD supporting the existence of a liability allele in the region.

Smoothing the pattern of LD in a target region has been successfully applied in the context of fine mapping.[19–21] While the various approaches differ in the extent to which they incorporate parametric modelling assumptions, most of them constrain the problem substantially by assuming, *a priori*, that a liability allele is present in the assessed interval. When the primary objective is testing for the presence of a liability allele, however, a more flexible approach is required. In regions where no liability alleles are present, the pattern of observed LD is expected to exhibit no signal; however, due to sampling error, population substructure and evolutionary forces, there will be random patterns in the observed LD signal. To model such data, non-parametric curve fitting approaches were investigated. Specifically, for a sample of m markers with physical locations, $x_1, x_2, \ldots, x_m$ and measured LD $y_1, y_2, \ldots, y_m$, the observed LD were fitted to an arbitrary smooth curve $g(\cdot)$, which allows for additional noise, $e_1, e_2, \ldots, e_m$:

$$y_i = g(x_i) + e_i$$

In particular, contrary to many fine-mapping methods, this approach does not force the fitted function to be unimodal. Next, the authors constructed a test based on an estimate of $g(\cdot)$ that utilised all of the LD measures in the region, to determine if there is evidence for one or more liability alleles in the region.

Materials and methods

The authors' objective was to develop a method for combining single-marker measures of association across markers in a chromosomal region to test for the presence of liability alleles. Non-parametric regression methods, which do not require an inferential model, seemed ideal for the task. In theory, any summary statistic might be used in the non-parametric regression. For example, from a series of transmission disequilibrium tests (TDTs) tests,[22] one might use the $-\log_{10}$ (p-value)s or the odds ratios. From a case-control sample, a statistic measuring differentiation between cases and controls at each marker can be used. It is important, however, that the statistics exhibit a pattern of association that, on average, is inflated in the vicinity of the liability allele.

The authors focussed on statistics of association for a case-control sample, in particular measures of LD between liability and marker alleles. Although some LD measures can be shown to be superior to others for fine-mapping simple Mendelian diseases,[23] none of them routinely outperforms the others in practice.[20] In this article, the authors have chosen to use two LD measures, δ and Nei's G_{ST}.[24,25] δ has proven to be useful for mapping mutations inducing Mendelian diseases,[26,27] and is a simple function of the recombination fraction between a disease and marker locus.[23,28] G_{ST} is a natural measure for multiallelic loci and measures the probability that an allele drawn from the case population differs in state from an allele drawn from the control population (see Appendix A1 for formulae for these measures).

Variance in LD measures is induced by two sources, the process of evolutionary drift over generations (evolutionary error) and the effect of taking a sample from the current population (sampling error). While it is difficult to estimate the evolutionary error, the sampling error can be easily computed. Under certain evolutionary conditions, it can be shown that the former quantity is approximately proportional to the latter.[28] This follows because both sampling and evolutionary error are primarily functions of the allele frequency and the sample/population size, respectively; in particular, for neutral alleles, evolutionary error is largely the result of repeated sampling error over the generations. This assumption is utilised here, in the BARS procedure.

In Appendix A1, formulae are provided for the sampling errors of δ and G_{ST}. In addition to sampling error, measures of LD obtained from the same general vicinity on a chromosome are likely to be correlated, even after one factors in the expected exponential decay described previously. Unlike evolutionary and sampling error, however, there is no direct statistical model from which to estimate the correlation between LDs sampled in a restricted region.

Non-parametric regression

There are many approaches to non-parametric regression such as a simple *running-mean*, which was used to fine-map hereditary haemochromatosis,[29] and the more complex *splines*.[30] Although all non-parametric regression methods assume a flexible form for the function $g(\cdot)$, methods vary in how smoothly they interpolate the neighbouring observations in a manner that avoids over-fitting; these approaches have been reviewed by Green and Silverman.[31]

Spline methods are based on the same principle as polynomial regression: a basis is chosen and then one proceeds to fit the curve using least squares regression. Unlike polynomial regression, however, the B-spline basis is chosen to facilitate fitting the curve primarily using the neighbouring observations. The interval of interest is divided into a set of ordered points, called knots, from which to build the basis function. Between each consecutive pair of knots, a cubic polynomial is fitted to the observations. To produce a smooth curve overall, the fitted cubic functions are forced by constraints to connect smoothly at each of the knots. Two extra terms are included in the basis, to constrain the behaviour of the fitted curve outside the range of the data. Consequently, altogether a model with k free-knots has dimension $k+2$.

Recently, a promising non-parametric regression approach known as Bayesian adaptive regression splines (BARS) was developed.[32] In contrast to *smoothing splines*, which place a knot at every data point (x_i), BARS uses a free-knot basis. Specifically, this approach estimates the best locations for placement of a minimum number of knots for the spline. The fewer the number of knots locally, the smoother the fitted curve. By estimating the optimal location of the knots, free-knot spline methods can adapt to local changes in smoothness. Consequently, BARS is highly flexible and has the capacity to adjust the smoothness of the fitted curve automatically to the local smoothness of the underlying function.

To illustrate various non-parametric regression approaches, the authors display the hereditary haemochromatosis data from Feder *et al.*[29] LD is measured using G_{ST} and the pattern is fitted using (i) a simple cubic polynomial, (ii) a smoothing spline and (iii) the BARS method (Figure 1).

With BARS, to estimate $g(\cdot)$, a free-knot spline approach is used, with k knots located at undetermined positions $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_k)$ within the specified interval of interest. The authors use $b_j(x_i)$ to denote the (i, j)th element in the matrix $\mathbf{B}$. As with polynomial regression models, it is assumed that the function $g(\cdot)$ can be expressed as a linear combination of the terms in the basis, with a vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{k+2})^T : g(x) = \sum_{j=1}^{k+2} b_j(x)\beta_j$. Or, to express this concept in matrix terms, with $\mathbf{g} = (g(x_1), g(x_2), \ldots, g(x_m))^T$, $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$ and $\mathbf{e} = (e_1, \ldots, e_m)^T$

$$\mathbf{g} = \mathbf{B}\boldsymbol{\beta} \quad \text{and} \quad \mathbf{y} = \mathbf{B}\boldsymbol{\beta} + \mathbf{e}$$

As in most regression models, with BARS it is assumed that the residual errors are independent and identically distributed (IID) normal random variables with unknown variance σ^2. To complete the BARS model specification, priors must be chosen for the unspecified parameters $(k, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma)$. The priors recommended in DiMatteo *et al.* were selected to be

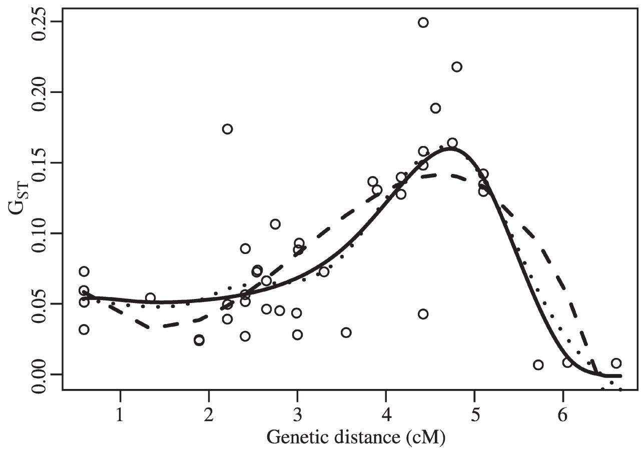

Figure 1. Non-parametric regression curves fitted to the hereditary haemochromatosis data from Feder *et al.*[28] using G_{ST} as a measure of LD. Lines are obtained by using cubic polynomial regression (dashed), smoothing spline regression (dotted) and the BARS method (solid).

essentially non-informative and hence have little influence on the resulting fitted curves.[32]

For the LD application, unequal variances are anticipated, due to varying allele frequencies across the loci, as well as correlated residuals due to the evolutionary process. (See, for example, Devlin *et al.*[28] or Lazzeroni.[19]) To apply the BARS modelling approach to LD data, the authors incorporated a more complex model for the error structure. First, they allow the residual errors to have non-constant variance; let $e_i = \delta_i \epsilon_i$, and assume that each ϵ_i is normally distributed with mean zero and variance σ^2. The constant terms, δ_i, $i = 1, 2, \ldots, m$, are taken to be proportional to the standard deviations of the y_i. Secondly, they model the correlation between error terms using an exponential decay function. To differentiate the two approaches they label them IID BARS and non-IID BARS. To choose δ_i^2 in practice, one could use a function of the statistical variances (v_i) computed for the LD measure being utilised. The authors follow DiMatteo *et al.* in choice of priors.[32] For details, see Appendix A2. To fit the model, a reversible-jump Markov chain Monte Carlo (MCMC) algorithm can be used;[33] see Zhang for details.[34]

The *credible interval* for a feature of the curve, say the mode, $\mathscr{M}$, is the Bayesian counterpart to a confidence interval. Let C_α denote the $(1 - \alpha)$ credible interval. It has the property $Pr(\mathscr{M} \in C_\alpha | \mathbf{y}) = 1 - \alpha$. A principal advantage of taking a Bayesian approach to inference is that a credible interval of any feature of the curve can be computed directly without requiring any approximations. (See Appendix A2 for details.) Because the confidence and credible interval concepts are essentially indistinguishable for this application, the credible interval for $\mathscr{M}$ will hereafter be referred to as the confidence interval.

The width of the confidence interval for the mode indicates how strongly the data support the location of the peak in the fitted curve. For instance, contrast results for G_{ST} and the homozygosity measure used in Feder *et al.* to map the causal variant (Figure 2).[29] Both curves place the mode similarly, but the associated confidence intervals show differing levels of precision in the estimators.

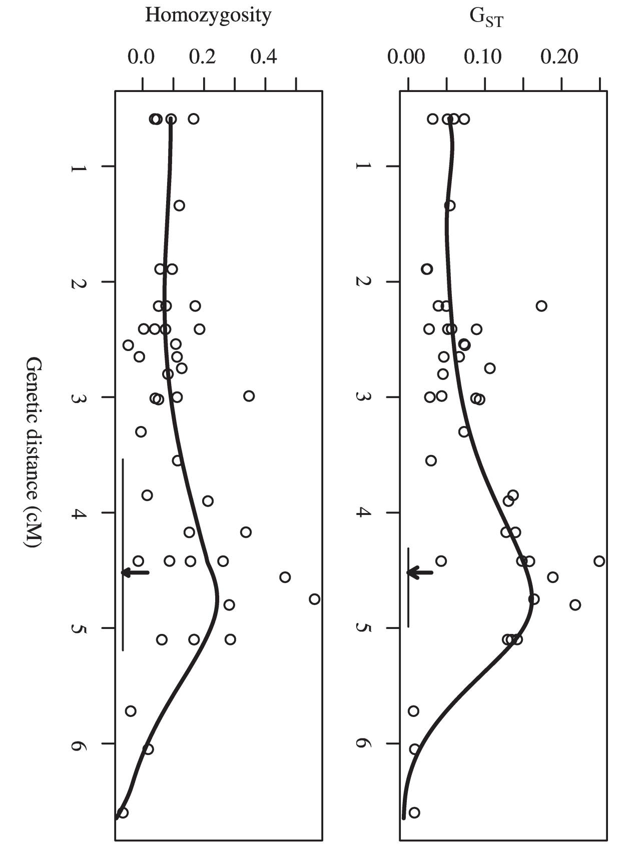

Figure 2. G_{ST} and a homozygosity measure fitted to the hereditary haemochromatosis data from Feder *et al.*[29] The arrow indicates the location of the mutation, and the corresponding line segment indicates the confidence interval obtained for the mode.

The BARS procedure

Theory suggests that LD should be greatest in the immediate vicinity of a liability allele. Consequently, the authors' interest lies in discovering the mode of $g(\cdot)$. $\mathscr{M}$ is considered to be a reasonable estimator of the location of a liability locus, if any are present in the region. If none are present, then no notable signals are expected in the LD pattern. Specifically, there is expected to be a lack of a definitive mode to the function— this is the basis of the BARS test.

To formulate the BARS test, the authors further develop the insight that if there are no signals from liability alleles in the region, then the confidence interval should encompass the entire region of interest. Ψ is defined as an indicator variable that takes the value 1 if there is a liability allele in the region of interest (Δ) and 0 otherwise. The aim is to test $H_0 : \Psi = 0$ versus $H_a : \Psi = 1$ and to control $Pr(\text{reject } H_0 | \Psi = 0)$. The test is based on the assumptions that, under the null hypothesis, $g(\cdot)$ is essentially constant for all x in the interval under investigation; ie the mode of the function is the entire interval, hence the confidence interval for the mode should include the entire interval. In practice, Δ is defined as the interval defined by the sampled grid points, less a negligible factor(η) to allow for edge effects in the spline fitting procedure: $\Delta = [x_1 + \eta, x_m - \eta]$.

It is assumed that a confidence interval is generated for the mode of $g(\cdot)$, as described previously, and $H_0 : \Psi = 0$ is rejected when C_α is a strict subset of Δ. Alternatively, if C_α encompasses Δ, the null hypothesis $\Psi = 0$ is not rejected. For illustration, see Figure 3, which shows a sample of five realisations of curves obtained by the MCMC algorithm as it moves through the parameter space selecting curves consistent with the data. Because there is no clear mode in these data, the modes of the five curves vary broadly across the interval.

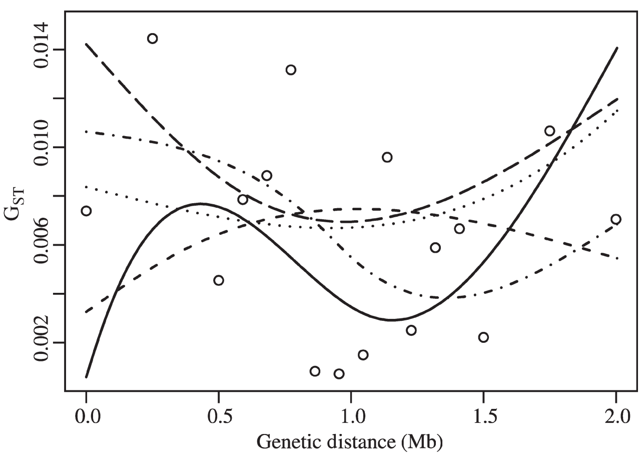

Figure 3. Simulated linkage disequilibrium values measured using G_{ST} when there is no liability allele in the region. The curves depict five of the many obtained via the MCMC algorithm.

The fitted regression curve for these data would be the average of R curves like these.

Unlike the typical multiple testing problem, it seems that the BARS test controls for the experiment-wide error rate automatically. That is, $\Pr(\Delta \subset C_\alpha | \Psi = 0) = 1 - \alpha$; ie the curve-fitting approach controls the overall probability of a false positive (designated α_e) at α without requiring any corrections for multiple testing.

Standard tests of association applied to individual markers sequentially can only directly control the false-positive rate for each marker, designated α_i however. If the m tests were independent, then $\alpha_e = 1 - (1 - \alpha_i)^m$, but obviously the test statistics are positively correlated. It follows that $\alpha_e \leq 1 - (1 - \alpha_i)^m$. If, using the Bonferroni criterion, one sets $\alpha_i = \frac{\alpha}{m}$, then it follows that $\alpha_e \leq \alpha$, but the exact value of α_e will not be known. Because α_e is less than the pre-selected α, then the power of the overall test can be low. At the other extreme, if $\alpha_i = \alpha$, a choice often made in practice, a high false-positive rate is the likely result.

The advantages of the smoothing approach are two-fold: first, α_e can be directly controlled; secondly, this procedure is less sensitive to errors in the data (see Mitchell *et al.*[35]). For example, suppose the LD for one marker is extremely high, but spurious. The LDs for the other markers in the neighbourhood of this marker will be likely to be less impressive. Curve-fitting methods combine the information of LDs along all the markers in the neighbourhood. Hence, this high LD will not have much effect in the authors' non-parametric LD method due to the smoothness of the fitted curve, while it may result in a false-positive association using standard multiple testing methods.

For comparison with the BARS procedure, the authors also investigated an alternative procedure for computing a confidence interval for the mode. A popular approach to non-parametric regression is smoothing splines, with the smoothing parameter chosen by generalised cross validation.[36] A percentile bootstrap approach can be used to produce a confidence interval.[37]

Results

As a preliminary proof of concept, the authors applied the BARS procedure to three classic data sets often used to illustrate the performance of fine mapping techniques: (1) 101 hereditary haemochromatosis patients and 64 controls measured at 43 single tandem repeat (STR) markers spanning a 6.5 Mb region;[38] (2) 94 cystic fibrosis disease haplotypes and 92 normal haplotypes measured at 23 bi-allelic markers spanning a 1.77 Mb region;[39] and (3) a sample of haplotypes with and without the Huntington disease mutation measured for 27 restriction fragment length polymorphisms (RFLPs).[40] Applying the G_{ST} measure to all three data sets yielded excellent results. The resulting confidence interval for the mode spanned the disease mutation in each data set. In addition, because the resulting confidence intervals spanned a fraction of the region of interest (0.66, 0.59 and 1.1 MB, respectively), the BARS procedure definitively indicated the presence of a disease mutation within the region. In addition, the BARS procedure was applied using the homozygosity measure for data set (1) and the δ measure for data sets (2–3) with similar results.

To explore the performance of the BARS procedure in more depth, the authors used an evolutionary simulation study to investigate the properties of the confidence interval for the mode of the fitted curve. In particular, they examined two features: (i) the false-positive rate when there was no liability allele in the region; and (ii) the coverage of the confidence interval obtained using G_{ST} for k-allelic markers when there was a liability allele present. They examined coverage because a procedure that has poor coverage properties will be likely to have a high false-positive rate in the proposed test. They also compared the coverage of the confidence interval obtained by smoothing splines with that of the BARS approach.

The case and control populations were simulated using an evolutionary simulation program that mimics features of natural populations by using direct simulation techniques; see Lam *et al.* for details.[41] Recombinations and mutations were permitted in each generation. Diploid individuals paired at random in their generation, mated and produced a random number of children. Each population was founded by 1,000 individuals and remained at that size for 50 generations to create random LD among alleles on normal chromosomes. After 50 generations, a disease mutation was introduced on one chromosome and the population grew exponentially for 200 generations, to a final size of 50,000 individuals.

Sixteen STR markers were simulated, covering a 2 Mb critical region, with spacings between markers of 0.25 Mb for the outer two gaps flanking a core region with 11 gaps of 0.09 Mb. The disease mutation was located in the middle of the region. The mutation rate was 0.001. The recombination process was a no-interference Poisson model based on the assumption that 1 cM = 1 Mb.

From each population, samples of 'disease' and 'normal' chromosomes were chosen for analysis. The authors first investigated the performance of the proposed BARS test under the null hypothesis. To do so, they obtained 100 cases and 100 controls by randomly subsampling from the samples of normal chromosomes just described. Six hundred such populations were generated and the authors investigated the size of the test with $\alpha = 0.05$ and 0.01. Using the IID BARS model, they found that the false-positive rate was quite close to the nominal rate: 0.045 and 0.010, respectively.

Next 200 data sets were sampled, with 100 cases and 100 controls each drawn from diseased and normal populations generated as described above. With these data, the authors evaluated the coverage of confidence intervals for the mode using smoothing splines, IID BARS and non-IID BARS methods. The results (Table 1) show that the coverage obtained for both of the BARS methods were almost exactly on target. The length of the intervals using non-IID BARS were slightly longer, as expected. Nevertheless, modelling correlated errors and non-constant variance had only a small effect on the performance of the BARS procedure, at least for these simulations.

From the size of the standard errors of the estimated modal quantity (Table 1) it was also concluded that the authors' test statistic had good power to detect the presence of a liability allele. The interval of interest was 2 Mb long, while the average 95 per cent confidence interval was 1.4 Mb. From this, it was concluded that most confidence intervals did not include the entire interval and hence would have rejected the null hypothesis.

In contrast to the BARS procedures, the coverage of the 95 per cent confidence interval obtained using smoothing splines with the smoothing parameter selected using generalised cross validation was surprisingly low (Table 1). Clearly, a test statistic based on this non-parametric regression procedure would not have good properties.

Table 1. The coverage of a liability locus by the 95% confidence interval using G_{ST} in 200 simulated datasets

	Smoothing splines	Bars	
		IID	non-IID
Coverage	0.540	0.940	0.950
SD	0.123	0.338	0.363

Note: SD is the standard deviation of the location of the mode.

Next, the authors investigated the behaviour of the BARS procedure under conditions designed to mimic the type of data likely to be encountered when studying a complex disease. To generate an additive model, they set the penetrance parameters f_j, $j = 0, 1, 2$ for j copies of the disease allele so that $f_2 = 2f_1 - f_0$, where f_0 is the probability that an affected individual has zero copies of the liability allele at the locus of interest. They set the prevalence $K = 0.005$ to model a relatively uncommon disorder, such as autism. To model a liability allele with a moderate effect, they set the attributable fraction, defined as $1 - f_0/K$, at 0.2. Given the relative frequency of the liability allele in the population, p, the genetic model was then complete. Two distinct models were obtained by choosing $p = 0.2$ and 0.4. To generate cases and controls, haplotypes were drawn from the simulated populations described previously. To produce genotypes for affected individuals, j haplotypes that bear liability alleles were drawn at random (using the implied probability distribution $\Pr(j|\text{case})$, and $2 - j$ haplotypes that did not bear liability alleles. Genotypes for control individuals were generated similarly.

For these models, the authors assessed the power to detect the presence of a liability allele using a sample size of 1,000 cases and controls and $\alpha = 0.05$. They found power of 62 per cent and 61 per cent and average length of a confidence interval of 1.61 and 1.62 for the models with $p = 0.2$ or 0.4, respectively. While only covering a minuscule portion of the space of potential genetic models, it is worth noting that both of these choices yielded a small genotype relative risk (f_1/f_0): 1.75 and 1.25 respectively. Thus, the authors' simulations suggested that the BARS procedure has promise. For these simulations, single-locus tests would require the Bonferroni correction for 16 markers (ie $\alpha = 0.0016$), or some other adjustment, and the power of the single-locus tests would be further eroded because alleles at these multiallelic markers are only in LD with the disease allele (the causal variant was not recorded). In reality, even more markers are likely to be tested, and if these do not include the causal variant(s), the power of single-locus tests will surely be low.

Discussion

In this paper, the authors have explored the use of non-parametric regression methods to integrate information about genetic association over multiple markers in a circumscribed genomic region. Motivating this exploration was the expected shift from association analysis targeting one or a few SNPs within a candidate gene to large scale association analysis, in which a dense set of SNPs distributed over substantial genomic regions, or perhaps the entire genome, can be queried. The analytical challenges in such data can be daunting and, for this reason, the authors hoped to develop a quick and facile screening tool to identify regions of the genome worthy of deeper genetic analysis.

In this spirit, they explored a particular non-parametric regression method called the BARS procedure, and contrasted it with a related method, smoothing splines. Their results suggested that BARS has promise as a quick screening tool. It successfully combined information for markers across a chromosomal region naturally by tracing the pattern of association. Furthermore, unlike the approach using a smoothing spline, the confidence intervals constructed with the BARS procedure achieved the proper coverage level. Incorporating correlated errors or non-constant variance for the measures of LD in the BARS procedure improved the coverage in some cases, but the amount of improvement was not substantial. Therefore, a simple and computationally efficient form of BARS could be applied to data in practice.

Despite these promising results, the BARS procedure requires further validation. At the present time, there is little agreement in the literature about whether single-locus or haplotype-based tests of association are more powerful. The authors believe that the diversity of opinions and results stems from the fact that the space of alternative hypotheses is huge, and that portions of this space favour single-locus tests while other portions favour haplotype-based tests. They conjecture that yet other portions of the space will favour the BARS procedure, namely regions in which there are association signals from multiple tested markers. It is also likely that the BARS procedure will often perform well when haplotype-based tests are most powerful. It is also worth noting that the BARS procedure can be applied to data that are obtained at considerably less cost (pooled genotypes) and hence it might be the most cost-effective method of analysis, even when it is not most powerful for a given sample size.

The principal assumption underlying the BARS approach is that, if a liability allele exists in the region under study, then the pattern of LD exhibited by the pairwise measures in the immediate vicinity of the liability allele exhibit, on average, higher LD than in the region overall. The complementary assumption for the procedure concerns the pattern of LD when no liability allele exists in the region under study. In this setting, it is assumed that the pattern of LD does not exhibit a distinct mode. Finally, for small samples, it is expected that the BARS procedure may fail to detect a mode in the pattern of statistics, even if one exists. The BARS method does not require the stronger assumption, often made for fine mapping procedures, that the pattern of LD is unimodal — declining smoothly as a function of the distance from the causal variant.

For a simple Mendelian disorder, the assumptions of the BARS procedure hold for most measures of LD.[23] By contrast, even for simple genetic disorders, the stronger assumption made by most fine mapping methods is not met for many measures of association. For instance, suppose the LD measures are pairwise test statistics for association. It is well known that the power of a test of association is a function of the allele frequency distribution at the chosen marker. Thus, two markers, both located in the immediate proximity of the causal variant, are likely to have a different power to detect the association. Consequently, even if the true pattern of LD is declining smoothly as a function of distance, the pattern of the test statistics will be somewhat irregular. For complex disorders, the situation is even less predictable. Nevertheless, the BARS procedure can handle a considerable amount of irregularity in the pattern of the LD signals. Ultimately, all that is required is that there exists a cluster of markers in the region under investigation that exhibit higher LD, on average, than the full set of markers.

Recently, the effect of haplotype blocks on measures of LD has been a topic of keen interest. For instance, in an attempt to fully incorporate the spatial effect of ancestral recombinant events on the LD pattern, Conti and Witte developed a hierarchical model for fine mapping that models both the smooth decay of LD over distance, together with the plateaux of constant LD predicted within a haplotype block.[21] By contrast, the BARS procedure does not seek to capture the added information potentially available in haplotype blocks. In this sense, the BARS procedure may be less powerful than one that does model this feature; however, the BARS procedure is valid whether the region under investigation possesses haplotype blocks or not. Consequently, the BARS procedure could be more robust, and even more powerful, than a method that seeks to test for association using knowledge of haplotype blocks when they are not present.

The other implicit assumptions of the model are that: (i) the measurement error is normally distributed, (ii) the variance of the LD measure is proportional to the sampling error and (iii) the correlation between neighbouring measures decays exponentially as a function of distance between the measured markers. None of these assumptions is likely to strictly hold in practice. Nevertheless, the authors' investigations suggest that these assumptions are not critical to the performance of the procedure.[34]

Translating these results to the analysis of large genomic regions also requires further exploration of how to divide the region, or even the entire genome, into segments that maximise the power of the BARS procedure. As the results indicate, large gaps between denser sets of markers should be treated as boundaries. The authors likewise suspect that one might want to partition the region into functional units, such as on the basis of plausible candidate genes or clusters of genes. Quite possibly, one might want to employ more than one tiling of BARS tests over a region.

All of these open questions can be answered by theoretical and empirical analyses. These results suggest the non-parametric BARS procedure has much potential as a tool to screen genomic regions for liability alleles because of its good statistical properties. Over the next decade, it will be interesting to see which methods prove most successful in the hunt for liability alleles.

Acknowledgments

This research was supported by National Institute of Health grant MH57881 and National Science Foundation grant DMS-9803433.

Appendix A1: Measure of association

Assume that a sample of n_1 cases and n_2 controls has been obtained ($n_1 + n_2 = n$). Let $p_i = \frac{n_i}{n}$ and $p_{ij} = \frac{n_{ij}}{n}$ ($i = 1, 2, \ldots, k$; $j = 1, 2$), where n_{ij} is the count of the ith allele in sample j and $n = \sum_{j=1}^{2}\sum_{i=1}^{k} n_{ij}$ (Table 2). Consider the case (or control) group alone as a population. The frequency of the i^{th} allele in the case (control) group is $p_{i|1}(p_{i|2})$, where $p_{i|j} = \frac{p_{ij}}{p_j}$ ($j = 1, 2$) and $\boldsymbol{\pi}_{i|j} = \frac{\boldsymbol{\pi}_{ij}}{\boldsymbol{\pi}_j}$.

The measure δ (or p_{excess}), is equivalent in some settings to the population attributable risk:[22]

$$\delta = \frac{p_{a|1} - p_{a|2}}{1 - p_{a|2}},$$

where a is the allele most associated with the liability allele. The approximate variance of δ is

$$\mathrm{Var}[\delta] = \frac{1 - p_{a|1}}{(1 - p_{a|2})^2}\left(\frac{1}{n_1}\cdot p_{a|1} + \frac{1}{n_2}\cdot\frac{p_{a|2}(1 - p_{a|1})}{1 - p_{a|2}}\right)$$

Nei's G_{ST} for cases and controls and a single locus with k alleles is defined as follows. Define gene diversity, H_{jl}, between groups j and l to be the probability that two alleles are different in structure when they are randomly drawn, respectively, from groups j and l ($j, l = 1, 2$), namely, $H_{jl} = 1 - \sum_{i=1}^{k} p_{i|j}\cdot p_{i|l}$. Define the net gene diversity, D_{jl}, between groups j and l to be the difference in gene diversity between groups j and l and the average of gene diversities within cases and within controls, namely $D_{jl} = H_{jl} - \frac{H_{jj}+H_{ll}}{2}$. Then, the average gene diversity H_S within cases and controls, the net gene diversity D and G_{ST} is

$$H_S = \frac{1}{2}(H_{11} + H_{22}) = 1 - \frac{1}{2}\left(\sum_{i=1}^{k} p_{i|1}^2 + \sum_{i=1}^{k} p_{i|2}^2\right),$$

$$D = D_{12} = H_{12} - H_S = \frac{1}{2}\sum_{i=1}^{k}(p_{i|1} - p_{i|2})^2,$$

$$G_{ST} = \frac{D}{2H_s + D} = \frac{\frac{1}{2}\sum_{i=1}^{k}(p_{i|1} - p_{i|2})^2}{2 - \frac{1}{2}\sum_{i=1}^{k}(p_{i|1} + p_{i|2})^2}.$$

Table 2. Sample frequencies of alleles obtained at a particular locus in a case-control study

	Cases	Controls	Marginal
Allele 1	n_{11}	n_{12}	
Allele 2	n_{21}	n_{22}	
⋮	⋮	⋮	
Allele k	n_{k1}	n_{k2}	
Marginal	n_1	n_2	n

Note: k may differ across loci.

For a bi-allelic locus,

$$G_{ST} = \frac{(p_{1|1} - p_{1|2})^2}{1 - (p_{1|2} - p_{2|1})^2}$$

Using the Delta Method, we obtain an approximate estimate of the variance of G_{ST}.[33] Let

$$D_{ij}^0 = (-1)^j\left(p_{i|2} - p_{i|1} + \sum_{r=1}^{k} p_{r|j}(p_{r|1} - p_{r|2})\right).$$

$$Hs_{ij}^0 = -p_{i|j} + \sum_{r=1}^{k} p_{r|j}^2,$$

and

$$\omega_{ij} = D_{ij}^0\cdot Hs - Hs_{ij}^0\cdot D.$$

It follows that

$$\mathrm{Var}[G_{ST}] = \frac{4}{(2H_s + D)^4}\left\{\sum_{j=1}^{2}\sum_{i=1}^{k}\frac{p_{i|j}}{n_j}\omega_{ij}^2 - \frac{1}{n}\left(\sum_{j=1}^{2}\sum_{i=1}^{k} p_{i|j}\omega_{ij}\right)^2\right\}.$$

Appendix A2: Technical details of the model

A model is required for the variance-covariance matrix of $(e_1, e_2, \ldots, e_m)$. It is assumed that e_i has variance equal to $\delta_i^2\sigma^2$, where δ_i is prespecified. The information about the form of the variances is recorded in a weight matrix $\mathbf{W}$, which is diagonal with $\mathbf{W}_{ii} = 1/\delta_i^2$. Secondly, the correlation between error terms is modelled using an exponential decay function: $\mathrm{Corr}(\epsilon_i,\epsilon_j) = \Sigma_{ij} = \exp\{-\gamma|x_i - x_j|^a\}$, in which the parameters (γ, a) are unspecified. Putting these two features together, and expressing the resulting covariances in matrix form, produces the following model, $\mathrm{Cov}(\mathbf{e}) = \boldsymbol{\Phi} = \sigma^2\mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{-1/2}$.

To choose δ_i^2 in practice, one could use the statistical variances (v_i) computed for the LD measure being utilised, or to obtain a more robust estimator one could use a linear combination between this estimated quantity and the average variance ($\bar{v}$) of the set of measured LDs: $(1 - q)v_i + q\bar{v}$. In their simulations, the authors used the latter procedure with $q = 0.5$.

Following DiMatteo *et al.*,[31] priors were chosen in an analogous manner: $\boldsymbol{\beta} \sim$ Normal with mean 0 and variance $m\sigma^2(\mathbf{B}^T\mathbf{\Phi}^{-1}\mathbf{B})^{-1}$; σ has prior proportional to $1/\sigma$; the spacings between knots were assumed to be uniformly distributed; $k \sim$ Poisson[5]; γ has prior proportional to $\gamma_0^{-1}\exp\{\gamma/\gamma_0\}$, with $\gamma_0 = 3$; and a has prior proportional to $a_0^{-1}\exp\{a/a_0\}$, with $a_0 = 0.6$.

To fit the model, a reversible-jump Markov chain Monte Carlo (MCMC) algorithm can be used.[32] This algorithm is suitable because the dimension of the model, which is a function of the number of free knots, k, is a free parameter. The reversible-jump algorithm allows the Markov chain to move from one dimension to another, and consequently the number of knots and the number of associated coefficients in the regression equation can change. The implementation of this algorithm for the non-IID BARS model is similar to the algorithm presented in DiMatteo *et al.*;[31] see Zhang for details.[33] It should be noted, however, that the algorithm may encounter numerical difficulty if most of the response variables (y_i) are near zero. For this reason, adding an arbitrary constant (0.01) to each response value before analysing the observations is suggested.

The credible interval of any features of the BARS curve can be computed directly using the MCMC algorithm. To compute the credible interval for the mode, one simply records the mode for each of R cycles of the MCMC algorithm recorded after the Markov chain has converged. Typically, for a model as complex as this, R should be at least as large as 10,000, with an initial burn-in period of 5,000 iterations. These outcomes are ordered from smallest to largest: $\mathscr{M}_{(1)} \leq \mathscr{M}_{(2)} \leq \ldots \leq \mathscr{M}_{(R)}$. The $(1-\alpha) \times 100\%$ credible interval is defined as the interval spanning from the $\alpha/2 \times 100th$ to the $(1-\alpha/2) \times 100th$ percentiles of the sampled modes's distribution, obtained from the MCMC algorithm.

Electronic-database information

A program written in C, with an R wrapper to perform the IID-BARS calculations, will soon be posted at http://www.stat.cmu.edu/~ roeder/

Zhang X (2002), will also be posted electronically at http://www.stat.cmu.edu/

References

1. Rioux, J.D., Daly, M.J., Silverberg, M.S. *et al.* (2001), 'Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease', *Nat. Genet.* Vol. 29, pp. 223–228.
2. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* Vol. 273, pp. 1516–1517.
3. Jorde, L.B. (1995), 'Linkage disequilibrium as a gene-mapping tool', *Am. J. Hum. Genet.* Vol. 56, pp. 11–14.
4. Jorde, L.B. (2000), 'Linkage disequilibrium and the search for complex disease genes', *Genome Res.* Vol. 10, pp. 1435–1444.
5. Kruglyak, L. (1999), 'Prospects for whole-genome linkage disequilibrium mapping of common disease genes', *Nat. Genet.* Vol. 22, pp. 139–144.
6. Gabriel, S.B., Schaffner, S.F., Nguyen, H. *et al.* (2002), 'The structure of haplotype blocks in the human genome', *Science* Vol. 296, pp. 2225–2229.
7. Carlson, C.S., Eberle, M.A., Rieder, M.J. *et al.* (2003), 'Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans', *Nat. Genet.* Vol. 33, pp. 518–521.
8. Reich, D.E., Cargill, M., Bolk, S. *et al.* (2001), 'Linkage disequilibrium in the human genome', *Nature* Vol. 411, pp. 199–204.
9. Abecasis, G.R., Noguchi, E. and Heinzmann, A. (2001), 'Extent and distribution of linkage disequilibrium in three genomic regions', *Am. J. Hum. Genet.* Vol. 68, pp. 191–197.
10. Daly, M.J., Rioux, J.D., Schaffner, S.F. *et al.* (2001), 'High-resolution haplotype structure in the human genome', *Nat. Genet.* Vol. 29, pp. 229–232.
11. Maniatis, N., Collins, A., Xu, C.F. *et al.* (2002), 'The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis', *Proc. Natl. Acad. Sci. USA*, Vol. 99, pp. 2228–2233.
12. Botstein, D. and Risch, N. (2003), 'Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease', *Nat. Genet.* Vol. 33(Suppl.), pp. 228–237.
13. Akey, J., Jin, L. and Xiong, M. (2001), 'Haplotypes vs single marker linkage disequilibrium tests: What do we gain?', *Eur. J. Hum. Genet.* Vol. 9, pp. 291–300.
14. Long, A.D. and Langley, C.H. (1999), 'The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits', *Genome Res.* Vol. 98, pp. 720–731.
15. Kaplan, N. and Morris, R. (2001), 'Prospects for association-based fine mapping of a susceptibility gene for a complex disease', *Theor. Popul. Biol.* Vol. 60, pp. 181–191.
16. Zhang, K., Calabrese, P., Nordborg, M. and Sun, F. (2002), 'Haplotype block structure and its applications to association studies: Power and study designs', *Am. J. Hum. Genet.* Vol. 71, pp. 1386–1394.
17. Schaid, D.J. (2002), 'Relative efficiency of ambiguous vs. directly measured haplotype frequencies', *Genet. Epidemiol.* Vol. 23, pp. 426–443.
18. Douglas, J.A., Boehnke, M., Gillanders, E. *et al.* (2001), 'Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies', *Nat. Genet.* Vol. 28, pp. 361–364.
19. Lazzeroni, L.C. (1998), 'Linkage disequilibrium and gene mapping: an empirical least-squares approach', *Am. J. Hum. Genet.* Vol. 62, pp. 159–170.
20. Cordell, H.J. and Elston, R.C. (1999), 'Fieller's theorem and linkage disequilibrium mapping', *Genet. Epidemiol.* Vol. 17, pp. 237–252.
21. Conti, D.V. and Witte, J.S. (2003), 'Hierarchical modeling of linkage disequilibrium: Genetic structure and spatial relations', *Am. J. Hum. Genet.* Vol. 72, pp. 351–363.
22. Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993), 'Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *Am. J. Hum. Genet.* Vol. 52, pp. 506–516.
23. Devlin, B. and Risch, N. (1995), 'A comparison of linkage disequilibrium measures for fine-scale mapping', *Genomics* Vol. 29, pp. 311–322.
24. Nei, M. (1973), 'Analysis of gene diversity in subdivided populations', *Proc. Natl. Acad. Sci. USA* Vol. 70, pp. 3321–3323.
25. Nei, M. (1987), *Molecular Evolutionary Genetics*, Columbia University Press, New York.
26. Hastbacka, J., de la Chapelle, A., Kaitila, I. *et al.* (1992), 'Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland', *Nat. Genet.* Vol. 2, pp. 204–211.
27. Hastbacka, J., de la Chapelle, A., Mahtani, M.M. *et al.* (1994), 'The diastrophic dysplasia gene encodes a novel sulfate transporter: Positional cloning by fine-structure linkage disequilibrium mapping', *Cell* Vol. 78, pp. 1073–1087.

28. Devlin, B., Risch, N. and Roeder, K. (1996), 'Disequilibrium mapping: Composite likelihood for pairwise disequilibrium', *Genomics* Vol. 36, pp. 1–16.
29. Feder, J.N., Gnirke, A., Thomas, W. *et al.* (1996), 'A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis', *Nat. Genet.* Vol. 13, pp. 399–408.
30. Wahba, G. (1990), *Spline Models for Observational Data*, Society for the Industrial and Applied Mathematics, Philadelphia, USA.
31. Green, P. and Silverman, B. (1994), 'Nonparametric Regression and Generalized Linear Models', Chapman & Hall, London.
32. DiMatteo, I., Genovese, C.R. and Kass, R.E. (2001), 'Bayesian curve fitting with free-knot splines', *Biometrika* Vol. 88, pp. 1055–1071.
33. Green, P. (1995), 'Reversible jump Markov chain Monte Carlo computations and Bayesian model determination', *Biometrika* Vol. 82, pp. 711–732.
34. Zhang, X. (2002), 'Statistical Methods for Discovering Disease Susceptibility Genes in Human Populations', Carnegie Mellon University, Ph.D. Thesis.
35. Mitchell, A.A., Cutler, D.J. and Chakravarti, A. (2003), 'Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test', *Am. J. Hum. Genet.* Vol. 72, pp. 598–610.
36. Venables, W.N. and Ripley, B.D. (1997), 'Modern Applied Statistics with S-Plus', Springer, New York, pp. 323–327.
37. Davison, A. and Hinkley, D. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, UK, p. 202.
38. Thomas, W., Fullan, A., Loeb, D.B. *et al.* (1998), 'A haplotype and linkage disequilibrium analysis of the hereditary hemochromatosis gene region', *Hum. Genet.* Vol. 102, pp. 517–525.
39. Kerem, B., Rommens, J.M., Buchanan, J.A. *et al.* (1989), 'Identification of the cystic fibrosis gene: Genetic analysis', *Science* Vol. 245, pp. 1073–1080.
40. MacDonald, M.E., Lin, C., Srinidhi, L. *et al.* (1991), 'Complex patterns of linkage disequilibrium in the Huntington disease region', *Am. J. Hum. Genet.* Vol. 49, pp. 723–734.
41. Lam, J.C., Roeder, K. and Devlin, B. (2000), 'Haplotype fine mapping by evolutionary trees', *Am. J. Hum. Genet.* Vol. 66, pp. 659–673.