32 Methods for SNP Regression Analysis in Clinical Studies

Selection, Shrinkage, and Logic

Michael LeBlanc, Bryan Goldman, and Charles Kooperberg

CONTENTS

32.1 INTRODUCTION

Investigations of the association of patient outcome with a few candidate single nucleotide polymorphisms (SNPs) or much larger numbers of SNPs have been undertaken in various therapeutic studies in oncology (e.g., Durie et al. 2009, Song et al. 2010). Since the genomic material often consists of germline DNA, not tumor DNA, the primary associations to therapeutic efficacy are typically not expected to be as strong as those seen for tumor gene expression. However, even with non-tumor

DNA, there could potentially be some strong correlations with disease symptoms at diagnosis, measures of drug metabolism and patient adverse events due to treatment.

While primarily outside the therapeutic setting, there have been many high-dimensional SNP studies which can be useful in defining good statistical strategies. For instance, there are an increasing number of validated associations seen from high throughput SNP studies including genome wide association studies (GWAS) (e.g., Hindorff et al. 2009, Peterson et al. 2010, Thomas et al. 2009). Often these are case-control studies and may include subject level meta-analyses on multiple cohorts to arrive at total numbers in the multiple thousands of cases and controls to achieve power to detect at least modest individual SNP associations with outcome. In addition, there has been some development of multi-SNP risk models from GWAS including Miyake et al. (2009), Zheng et al. (2008), Yang et al. (2010).

In most therapeutic clinical trials, the number of patients are typically only several hundreds, even when combining across studies. These sample sizes are modest enough to make it far more challenging to conduct well powered tests of association or risk modeling. Realistically, only large effects will be reliably identified in these moderately sized studies. However, given sufficient signal, SNP association studies are feasible; therefore, this chapter will consider model building strategies to construct prognostic or disease risk models, trading off variance control as well as interpretation. A good statistical strategy for risk modeling with GWAS data was outlined in Kooperberg et al. (2010) but we think it applies more broadly to smaller scale SNP analyses more typical of cancer therapeutic studies. Our proposal for a straight-forward statistical regression modeling approach can be outlined as follows: (1) data cleaning, (2) selection of a smaller number SNPs (if there are initially a large number under consideration), (3) modeling in some parsimonious fashion; shrinkage methods are one possibility or using a method that combines features in some logical fashion (such as logic regression or regression trees) and (4) a strategy to avoid drawing false positive associations or building models that are overly complex.

We note that there are many options for modeling in this context. Our focus is on SNP regression; we do not address alternatives here in terms of haplotype reconstruction, although algorithms have been developed for that purpose (for instance, see SNP-Haplotype Adaptive Regression [SHARE]) (Dai et al. 2009). Other than direct haplotype methods, there are methods to reduce dimensionality by using regularization to combine SNPs within gene as a component of the modeling procedure (Chen et al. 2010).

We illustrate the methods with a SNP data set from a clinical trial of multiple myeloma patients from the University of Arkansas generated as part of the Bank on a Cure project.

32.2 SIMPLE GENOTYPE DATA

Humans have two copies of each autosomal chromosome. The total length is about three billion base pairs. The most common variation between humans are variations in a single locus, known as SNPs. SNPs are typically coded as 0,1, or 2, the number of minor (variant) alleles at a particular locus. These data can be re-coded as a variable for the dominant effect by labeling 1 if $SNP = \{1,2\}$ or 0 otherwise and for the recessive effect as a 1 if $SNP = \{2\}$ or 0 otherwise. Often, given such a large number of SNPs, and hope for mostly cumulative association with subject outcome, the additive code {0,1,2} is used in many statistical testing and regression strategies. This coding is often the most powerful in detecting SNP disease associations.

Data quality is, of course, a fundamental issue in any analysis. However, in this chapter we will not address steps to assess the quality of the genotyping calls. These issues are platform dependent and checking quality would likely involve investigation of control and replication samples. In addition, in some cases, depending on the platform, it could involve returning to the images of relative intensities to re-evaluate the calls. Other quality control techniques involve inspection of QQ plots of the associations (where there are sufficient numbers of SNPs being investigated) to check for more global departures from the 45° line than what would be expected for the hypothesized scenarios where only a few SNPs are thought to be associated with patient outcome or toxicity.

Additional filtering of samples and SNPs for subsequent analysis also typically involves removal based on a sufficient number of observed called genotypes. For instance, often all samples with a call rate smaller than some value (say 97% for large arrays, but often this is set somewhat lower for smaller scale genotyping technologies) will be removed for consideration for further analysis. In addition SNPs that substantially fail the assumption of Hardy–Weinberg (HW) equilibrium, for instance, with a p-value of 10^{-3}–10^{-5}, are not considered. The extremeness of the p-value would need to depend on the number of SNPs under consideration. We note that typically the check of HW is done in control samples only; in our clinical trial settings all patients typically have disease. So while checking HW plays a role in data cleaning, it is not clear that HW equilibrium needs to be valid for all SNPs in the therapeutic "all case" setting. Another filtering option used primarily for power considerations, is to remove SNPs with a low minor allele frequency (say .05) prior to any formal model building.

32.3 EXAMPLE: MYELOMA SNP ANALYSIS

To demonstrate methods in this chapter, we use data based on patients with previously untreated multiple myeloma enrolled in the TT2 trial at the University of Arkansas between October 1998 and February 2004. Details of patient characteristics plus treatment and clinical outcomes have been reported previously (Durie et al. 2009). The multiple myeloma baseline evaluation included serum and urine protein electrophoresis, quantitative immunoglobulin measurements, total 24 h urine protein excretion and serum beta-2 microglobulin. The outcome was defined as extensive bone disease defined by x-ray criteria.

While the original study had data on 282 patients, we construct a larger simulated data set which we think is a more appropriate size for demonstrating regression modeling methods in this chapter. In addition to the observed data cases, 118 additional cases were drawn as a simple bootstrap sample to augment the real sample to yield a total of 400 real and simulated patients for analysis. Given the data set we used is partially simulated, the results presented in this chapter do not agree with the prior published results for this data set.

32.4 ASSESSING ASSOCIATIONS

Continuous, binary, or survival endpoints are potentially of interest in the SNP association studies in the context of cancer clinical trials. Therefore, to keep the discussion general, we present the models in terms of the regression component of each of the outcome models.

32.4.1 Univariate Associations

Most SNP association studies involve, at a minimum, a report of single SNP associations, potentially after adjusting for population heterogeneity; the adjustments may be based on reported race, genomic measurements of racial variation, and/or baseline clinical factors in a therapeutic study. As noted earlier, a single SNP has three levels, so common coding for assessment of association can be as linear $X = \{0,1,2\}$, dominant $X = 1$ if SNP is 1 or 2, 0 otherwise recessive $X = 1$ if $SNP = 2$ or 0 otherwise.

Consider a regression setting, where there are n observations on variables including non-genomic Z_1:$l = 1, \ldots, L$ and genomic X_k:$k = 1, \ldots, K$. To simplify presentation, assume only a single adjustment variable denoted as Z. Then testing individual SNPs can be reduced to assessing score or likelihood test statistics of $\beta_k = 0$ in the regression model for coded SNP k

$$\eta(\mathrm{X}, \mathrm{Z}) = \beta_0 + \gamma Z + \beta_k X_k.$$

Nominal p-values can be calculated for all $k = 1, \ldots, K$ tests of association. If the goal is to identify univariate associations, then strategies to control the error rates for false positives are of primary importance. The simplest way to control the family wise error rate (FWER) is to use a Bonferroni correction. However, it is often preferable with moderate numbers of SNPs (some chosen that may have relatively high correlation with each other—high linkage disequilibrium) to acknowledge the correlation structure. A simple way to incorporate the correlation structure in testings is by permutation sampling and to compare the observed statistics to those observed from a sample from the permutation distribution. Where the model includes adjustment variables, permutation of the score residuals and recalculating the test statistics is a more relevant null distribution. If the primary objective focuses on risk or prognostic modeling based on multiple SNPs, then the selection of a set of SNPs for further modeling does not require such a stringent selection of SNPs. One may select some limited number regardless of their overall significance. For instance, with a 3000 SNP study, one may select the SNPs with the top 1% or 5% of p-values to reduce overall variability of the subsequent modeling method. However, as described later, additional strategies for model selection (such as cross-validation) will ultimately be needed.

32.4.1.1 Example: Univariate Statistics

After filtering for low allele frequency and call rate, 1903 out of 3404 genotyped SNPs remained in the myeloma data set. We calculated univariate statistics for each of the SNPs displayed in Figure 32.1, testing the SNP associations with bone disease. While one could assess significance via permutation sampling here,

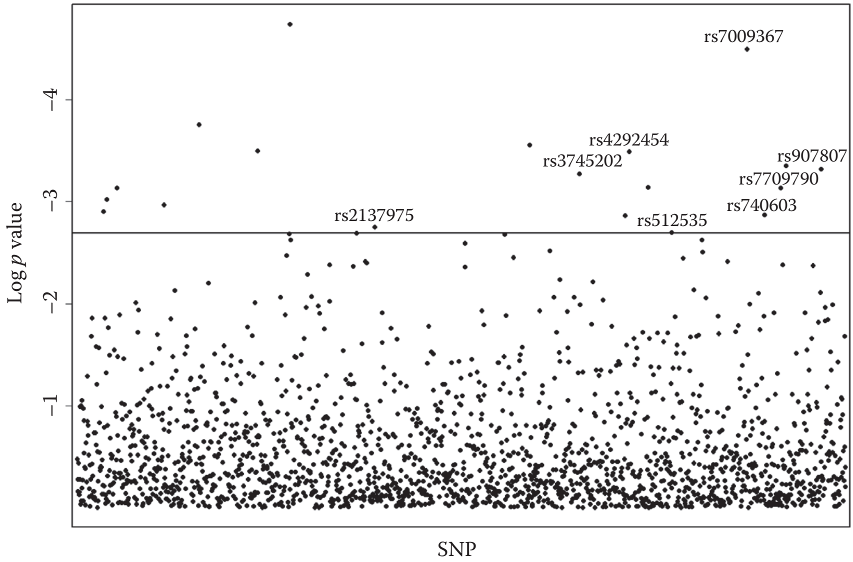

FIGURE 32.1 Univariate SNP p-values.

Bonferroni-corrected .05 level corresponds to the log10 p-value of −4.58 which only includes one SNP. The labels on the plot correspond to the SNPs selected as part of the regularized regression method (least absolute and selection operator [LASSO]) described in a later section. It also demonstrates that it may be useful to include slightly more variables (SNPs) as part of the modeling method, even if they do not achieve significance by multiple comparison adjusted methods.

32.5 MULTIVARIATE ASSOCIATIONS

Assume there is interest in combining genetic information across loci. The simplest model is some additive combination of the genotype data. For instance, for binary outcome data or case control studies, a logistic regression model can be used where the probability of disease or toxicity is $P(Y=1|Z, X)=\exp(\eta(X,Z))/(1+\exp(\eta(X,Z)))$ where a linear combination of individual SNPs which is the simplest way to combine information between SNPs adjusting for any baseline factors is

$$\eta(\mathbf{X},\mathbf{Z}) = \beta_0 + \gamma \sum_{l=1}^{L} \gamma_l Z_l + \sum_{k=1}^{K} \beta_k X_k. \tag{32.1}$$

32.5.1 Penalized Regression

Consider the likelihood-based regression setting. Assume there are n independent observations of the genotype and non-genotype data described earlier. Denote the likelihood function as $l(\cdot)$ which would be a binomial likelihood for binary outcomes, partial likelihood for survival endpoints or the typically normal likelihood for continuous outcome data. While step-wise model building is a reasonable strategy for building models (with the Akaike information criteria [AIC] Akaike [1974] or BIC to do the model selection), we believe penalized methods offer the advantage of reduced variability especially in moderate sample size settings. A popular regularization or penalization method is the LASSO (Tibshirani 1996) and its extensions (e.g., Hastie et al. 2009). There has been considerable subsequent work on rapid estimation methods. Suppressing the notation for the adjustment variables, which may or may not be penalized, the LASSO estimate $\tilde{\beta} = (\tilde{\beta}_1, \ldots \tilde{\beta}_m)'$ is defined as the maximizer of

$$g(\beta) = \sum_{i=1}^{n} l\left(Y_i, \sum \beta_k X_{ik}\right) - \lambda_1 \sum |\beta_k|^1,$$

where λ_1 is a non-negative penalty parameter. This estimator has the attractive property that as the penalty λ_1 increases, maximizing $g(\beta)$ with respect to β leads to some of the β_k set to zero. In addition, the variable selection and regression function estimates tend to have overall less variability than those obtained from forward or backward variable selection methods. Further variance reduction, at the cost of potentially less sparse solution involving more non-zero coefficients, is obtained by using a mixture of L_1 and L_2 penalty called the "elastic net" proposed by Zou and Hastie (2005). The elastic net can be expressed as an optimization problem with the objective function with both squared and absolute penalty terms

$$g(\beta) = \sum_{i=1}^{n} l\left(Y_i, \sum \beta_k X_{ik}\right) - \lambda_1 \sum |\beta_k|^1 - \lambda_2 \sum |\beta_k|^2 .$$

There is overall shrinking of the linear predictor and setting of some of the coefficients to zero in the model as the penalty parameters λ_1 and λ_2 are increased. Flexible software that fits continuous, binary, and survival data is available in *R* statistical language (GLMNET). In this section, we have described these methods in terms of the original predictors X_k; we could generalize to sets of regression splines or even more complex multivariable basis functions, $B_j(X)$, $j = 1, \ldots, p$ as described in the next section.

For the case of LASSO, the models are indexed by a single parameter λ_1 or for the elastic net by two parameters λ_1 and λ_2. To objectively choose these tuning parameters, one can either use a separate data set or use a resampling technique such as K-fold cross validation. For K-fold cross validation, the data are divided in approximately K groups (for instance, $K = 5$ or 10), and fractions $(K - 1)/K$ are used to construct the models and index the sequence of models by the tuning parameters, and the log-likelihood is evaluated for each model on the remaining $1/K$ of the data, called the test data. The analysis is repeated for each of the K subsets of the data, and test sample log-likelihoods are averaged over the K subsets. Tuning parameters (λ_1 and λ_2 in the case of the elastic net) are chosen that lead to maximum likelihood solutions. It is important that all of the variable selection aspects of the modeling be

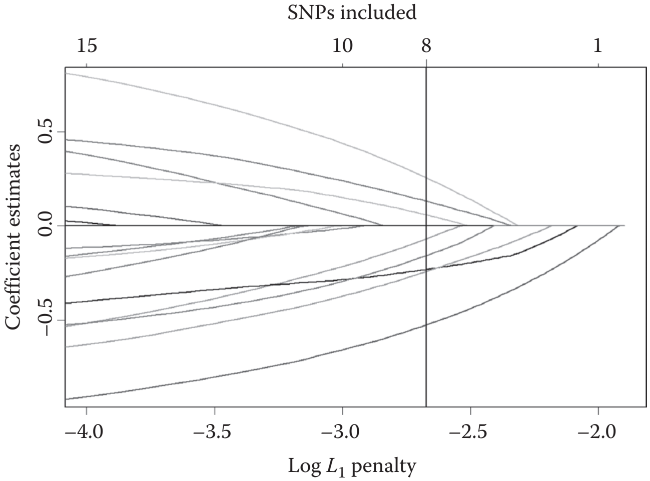

FIGURE 32.2 L_1 penalized regression coefficient paths.

included in the cross-validation loop. For instance, if the initial filter is to use only the top q most significant variables in the penalized regression algorithm, that should be part of the cross-validation loop.

32.5.1.1 Example: L_1 Penalized Regression Bone Disease Model

We chose to use L_1 regression to construct a multi-SNP model of bone disease. Based on prior methodological work we filtered the number of SNPs under consideration prior to using LASSO. For this example we chose only the top 20 SNPs (the top 1% SNPs) with the smallest p-values. We left the SNP coding as ordinal {0,1,2}. We applied five fold cross validation over both the variable selection as well as the model building and the final model chosen included eight SNPs. We acknowledge the number of SNPs filtered is another tuning parameter which could also be estimated using cross-validation. The coefficient profile is presented in Figure 32.2. There are eight SNPs remaining in the model; the labels of these SNPs are included in Figure 32.1. Cross-validated log-likelihood is presented in Figure 32.3.

32.5.1.2 Example: L_1 Penalized Regression Simulation

A concern with the relatively small sample sizes with SNP studies as part of therapeutic cancer studies is that only large effects would be seen. We conducted a small simulation study to investigate this issue. SNP data were generated by resampling from the "observed 400" patient cohort and the disease response was simulated from a single SNP regression model out of the total of 1903 SNPs under consideration. For an odds ratio of 1.75, 91% of the LASSO models (with complexity selected by five fold cross validation) included the true SNP. On average, 5.2 SNPs were selected, indicating at least some tendency for over-fitting. For an odds ratio of 2.0 the probability of selecting the correct SNP increased to 98.6% with a similar level of overfitting. This indicates the potential for identifying moderately strong associations from clinical SNP data.

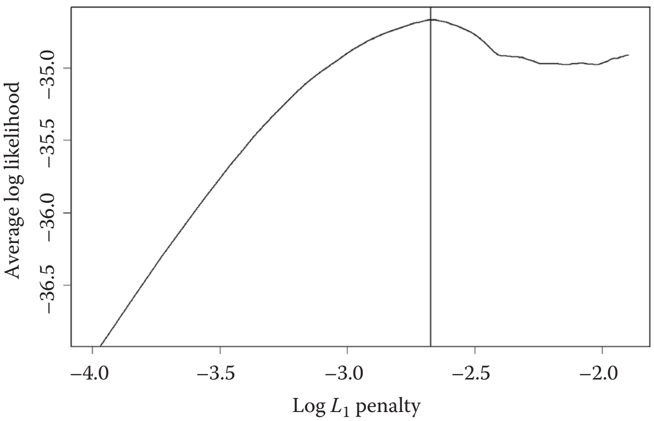

FIGURE 32.3 Cross-validated log-likelihood for L_1 regression.

32.5.2 Logic Regression

One way to extend the linear model, described earlier, is to consider more complex SNP combinations. For instance,

$$\eta(\mathbf{X}) = \beta_0 + \sum_h \beta_h B_h(X),$$

where the basis functions $B_h(X)$ could represent nonlinear functions of several of the X_k. A specific example of this model is regression trees (Breiman et al. 1984) and extended to survival data, for instance, by Segal (1988) and LeBlanc and Crowley (1993). In that case, the basis functions $B_h(X)$ are products of subset functions of the individual variables, of SNPs. In tree-terminology they would be terminal nodes of a regression tree. Trees are discussed in detail in another chapter and hence are not developed here. Trees have been used in the analysis of SNP data both as single trees (Durie et al. 2009), and as ensembles, such as random forests (Ishwaran et al. 2008).

In this section we describe a method which can be viewed as one that uses alternative interpretable basis functions based on logical or Boolean rules. The method, called "logic regression" is a methodology that is particularly suited for situations where the data are binary or in the case of SNP data "almost binary" and are binary if they are first coded as binary or recessive or dominant codes for each SNP (Ruczinski et al. 2003, Kooperberg and Ruczinski 2005). The resulting model, again suppressing the adjustment variables, can be expressed as basis functions or as Boolean combination of binary predictors X_j, $j = 1, \ldots, p$ such as

$$B_h(X) = \left[\left(X_1 \text{ OR } X_4^c\right) \text{AND } X_2\right],$$

where the X_j are binary coding of the SNP data as either dominant or recessive effects and hence X_j are assumed to be either 0 or 1. X_j^c is the complement function, so $X_j^c = 1 - X_j$. Additional adjustment covariates Z can be included in the model as for the other models described earlier.

Logic regression is usually implemented as a stochastic simulated annealing algorithm which selects those logic terms $B_h(\cdot)$ which maximize the log-likelihood corresponding to the model. Given the potential complexity or adaptivity in fitting each basis function, the number of logic terms m is set to be some small constant (between 1 and 3).

There is a tree-based representation of any logic term, which allows an easy specification for the stochastic optimization algorithm. At each step of the simulated annealing algorithm, one logic tree can be replaced by another logic tree through simple change operations on the tree. These operations are demonstrated in Figure 32.4.

As is true for other simulated annealing algorithms, if the likelihood of the new model is larger than for the current model, then the new model is chosen; if the current model has a smaller likelihood than the new model, then new model is chosen with a probability that is a function of the difference between the current and new model log-likelihood. The probability of choosing the new model is related to the current step number of the algorithm. At early steps of the procedure most of new models are accepted, while after many steps, only improved models are chosen with high probability.

Similar to penalized regression methods, the model complexity (which we have measured as the number of leaves or terms in the logic model) should be selected in such a fashion that acknowledges the significant adaptivity of the logic regression algorithm. Logic regression allows both permutation tests to assess overall association as well as K-fold cross validation.

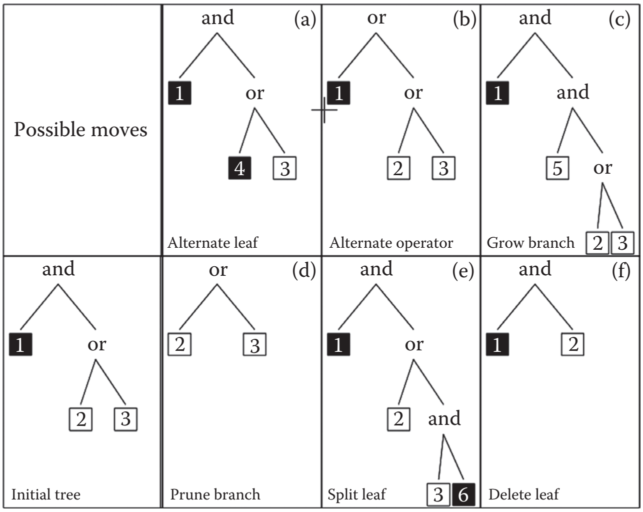

FIGURE 32.4 Logic regression optimization operations.

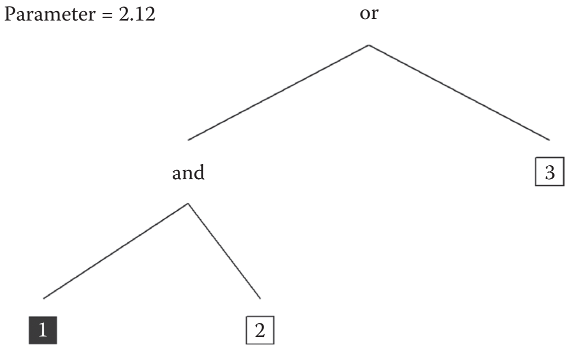

FIGURE 32.5 Selected logic regression tree with 3 SNPs (1) rs4292454, (2) rs3745202, and (3) rs7843746.

32.5.2.1 Example Revisited: Logic Regression

We return to our example, but now with the goal of constructing a simple logic regression model of Boolean combinations of SNPs. After conducting full cross validation, involving both filtering to the 20 SNPs and application of logic regression, there was no evidence of improvement in prediction over the null model. As expected in the relatively modest sample size setting, the more adaptive and discrete logic regression method suffers somewhat from additional variance, even if the interpretation of a small number of SNPs would be desirable. However, for demonstration purposes we show the three-leaf tree which on the full data set had a deviance of 288 compared to the null model deviance of 321. The model is presented in Figure 32.5. The estimated odds ratio was 2.12 and the logic representation was

$$L = (\text{NOT dominant rs4292454}) \text{ AND } (\text{dominant rs3745202}) \\ \text{OR } (\text{dominant rs7843746})$$

which represented patients at a higher risk of extensive bone disease.

32.6 ASSESSING TREATMENT–GENE INTERACTIONS

In the prior section, we demonstrated logic regression which can be viewed as a procedure which builds models within a special class of interactions. However, there is now an increasing literature for efficiently assessing more simple gene × gene or gene × baseline clinical factor, or gene × treatment interactions. It has been shown that significant gains in power can be obtained in many situations by only considering SNPs with the most significant marginal association prior to testing the interaction. It has been shown that if the test of interaction is independent of the marginal test, then one needs to adjust only for the number of interactions tested rather than the total number of marginal tests conducted (Kooperberg and LeBlanc 2008, Dai et al. 2011). Penalized regression strategies that directly incorporate interactions were considered by Park and Hastie (2008). Several strategies for evaluating and utilizing interactions in genomic studies were described in Kooperberg et al. (2009).

However, potentially the most important class of interactions of interest in therapeutic studies is the interaction of a SNP with the assigned treatment group. A question may arise: Does the impact of treatment (say on toxicity) depend on a gene? A simple case of one treatment and one SNP in a multiplicative model can be represented as

$$\eta(X, Z) = \beta_0 + \gamma Z + \beta_k X_k + \delta Z X_k, \qquad (32.2)$$

where

$Z = \{0,1\}$ indicates treatment

X_k represents the specific genetic variable

To assess potential interactions, one can test all K gene by treatment interactions. Another strategy, as noted earlier, is to first test all the gene main effects, and then only test a subset of the interactions corresponding to the top most significant gene main effects. It can be shown that the second stage testing is asymptotically independent of the first stage; so if only M interaction tests are considered at the second stage, then significance only need to be adjusted by the factor of M tests. This can lead to increased power to find interactions in many settings. Of course, if the interaction is "pure" or hidden entirely in the marginal test, then it will be missed at second stage testing (Dai et al. 2011).

If a true interaction is not the primary goal, but rather interest focuses on any gene association that may be modified by baseline clinical factors, then more general weighted tests can be used. For instance, LeBlanc and Kooperberg (2009) constructed adaptively weighted test statistics that can be substantially more powerful than the single tests if interactions are truly present in the data so that within a subset of patients the genetic association is substantially stronger.

32.7 DISCUSSION

In this chapter our goal was not to provide an exhaustive review of methods for prognostic or risk modeling with SNP data, but rather to focus on two techniques which have been used and/or developed for SNP data which represent smooth prediction and non-smooth interpretation-based strategies: penalized regression and logic regression. Obviously alternatives to linear penalized models could include regression trees demonstrated in the analysis of clinical SNP data by Durie et al. (2009) or ensembles of trees such as random forests. Other methods have been proposed, including multifactor dimensionality reduction (MDR, Richie et al. 2001) which focuses on low-dimensional combinations of SNPs.

We have not addressed sensible ways to combine SNPs that may be close together; for instance, some sets of SNPs may be thought to correspond to a haplotype block. In that setting, prior to doing some of the modeling proposals we have made in this chapter, one may first want to use a haplotype reconstruction method (e.g., Li et al. 2006). After appropriately acknowledging the haplotype uncertainty, one can use regression methods to predict the outcome. An adaptive technique that does SNP selection and haplotype regression, "SNP and HAplotype REgression" was developed by Dai et al. (2009). An alternative approach is to group using regularized methods on SNPs localized in a region; see, for instance, Chen et al. (2010).

An important issue with the analysis of SNP data in the context of moderate-size clinical studies is an assessment of power to detect meaningful associations. For instance, unlike some large case control studies including meta-analyses used for GWAS studies, our experience is that SNP association studies in therapeutic settings often consist of small numbers of cases. Therefore, statistical methods which control the variability and don't over-fit the data are critically important. In addition, where power to detect reasonable sized effects is limited, combining across clinical studies, where scientifically sensible, may be a useful strategy.

Software for logic regression and GLMNET are currently available at CRAN.

ACKNOWLEDGMENTS

This work was supported by the U.S. NIH through R01 CA90998, R01 HG006124, and P01 CA53996. The authors thank Dr. Brian Durie and the International Myeloma Foundation and the Bank on a Cure project for use of the data. They acknowledge helpful interactions with Dr. Antje Hoering for information relating to the SNP data.

REFERENCES

Akaike HA. 1974. New look at model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. 2010. Insights into colon cancer etiology using a regularized approach to gene set analysis of GWAS data. *American Journal of Human Genetics*, 86:860–871.

Dai JY, Kooperberg C, LeBlanc M, Prentice R. 2011. On two-stage hypothesis testing procedures via asymptotically independent statistics. Submitted to *Biometrika.*

Dai JY, LeBlanc M, Smith NL, Psaty B, Kooperberg C. 2009. SHARE: An adaptive algorithm to select the most informative set of SNPs for candidate genetic association. *Biostatistics*, 10(4):680–693.

Durie BG, Van Ness B, Ramos C, Stephens O, Haznadar M, Hoering A, Haessler J et al. 2009. Genetic polymorphisms of EPHX1, Gsk3beta, TNFSF8 and myeloma cell DKK-1 expression linked to bone disease in myeloma. *Leukemia*, 10:1913–1919.

Hastie T, Tibshirani R, Friedman J. 2009. *Elements of Statistical Learning*, 2nd edn., Springer, New York.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences USA*, 106:9362–9367.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. 2008. Random survival forests. *Annals of Applied Statistics*, 2:841–860.

Kooperberg C, LeBlanc M. 2008. Increasing the power of identifying gene X gene interactions in genome-wide association studies. *Genetic Epidemiology*, 32:255–263.

Kooperberg C, LeBlanc M, Dai JY, Rajapakse I. 2009. Structures and assumptions: Strategies to harness gene x gene and gene x environment interactions in GWAS. *Statistical Science*, 24(4):472–488.

Kooperberg C, LeBlanc M, Obenchain V. 2010. Risk prediction using genome-wide association studies. *Genetic Epidemiology*, 34(7):643–652.

Kooperberg C, Ruczinski I. 2005. Identifying interacting SNPs using Monte Carlo logic regression. *Genetic Epidemiology*, 28:157–170.

LeBlanc M, Crowley J. 1993. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88:457–467.

LeBlanc M, Kooperberg C. 2009. Adaptively weighted association statistics. *Genetic Epidemiology*, 33(5):442–452.

Li Y, Ding J, Abecasis GR. 2006. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics*, 79:S2290.

Miyake K, Yang W, Hara K, Yasuda K, Horikawa Y, Osawa H, Furuta H et al. 2009. Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. *American Journal of Human Genetics*, 54:236–241.

Park MY, Hastie T. 2008. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9:30–50.

Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, Arslan AA et al. 2010. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33., *Nature Genetics*, 42(3):224–228.

Richie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, 69:138–147.

Ruczinski I, Kooperberg C, LeBlanc M. 2003. Logic regression. *Journal of Graphical and Computational Statistics*, 12:475–511.

Segal MR. 1988. Regression trees for censored data. *Biometrics*, 44:35–48.

Song Y, Barlow WE, Albain KS, Choi JY, Zhao H, Livingston RB, Davis W et al. 2010. Gene polymorphisms in cyclophosphamide metabolism pathway, treatment-related toxicity, and disease-free survival in SWOG 8897 clinical trial for breast cancer. *Clinical Cancer Research*, 16:6169–6176.

Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE et al. 2009. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nature Genetics*, 41(5):579–584.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*: *Series B*, 58:267–288.

Yang et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–569.

Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, Adami HO et al. 2008. Cumulative association of five genetic variants with prostate cancer. *New England Journal of Medicine*, 358:910–919.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*: *Series B*, 67:301–320.