Identification of non-Hodgkin’s lymphoma prognosis signatures using the CTGDR method

Shuangge Ma1,∗, Yawei Zhang1, Jian Huang2, Xuesong Han1, Theodore Holford1, Qing Lan3, Nathaniel Rothman3, Peter Boyle4 and Tongzhang Zheng1

1School of Public Health, Yale University, New Haven, CT 06510, 2Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, 3Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD 20892, USA and 4International Prevention Research Institute, Lyon, France

Received on May 1, 2009; revised on September 28, 2009; accepted on October 16, 2009

Advance Access publication October 22, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Although NHL (non-Hodgkin’s lymphoma) is the fifth leading cause of cancer incidence and mortality in the USA, it remains poorly understood and is largely incurable. Biomedical studies have shown that genomic variations, measured with SNPs (single nucleotide polymorphisms) in genes, may have independent predictive power for disease-free survival in NHL patients beyond clinical measurements.

Results: We apply the CTGDR (clustering threshold gradient directed regularization) method to genetic association studies using SNPs, analyze data from an association study of NHL and identify prognosis signatures to diffuse large B cell lymphoma (DLBCL) and follicular lymphoma (FL), the two most common subtypes of NHL. With the CTGDR method, we are able to account for the joint effects of multiple SNPs, whereas most existing studies are single-marker based. In addition, we are able to account for the ‘gene and SNP-within-gene’ hierarchical structure and identify not only predictive genes but also predictive SNPs within identified genes. In contrast, existing studies are limited to either gene or SNP identification, but not both. We propose using resampling methods to evaluate the predictive power and reproducibility of identified genes and SNPs. Simulation study and data analysis suggest satisfactory performance of the CTGDR method.

Contact: shuangge.ma@yale.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

NHL (non-Hodgkin’s lymphoma) represents a heterogeneous group of lymphocytic disorders ranging in aggressiveness from very indolent cellular proliferation to highly aggressive and rapidly proliferative processes. Although it is the fifth leading cause of cancer incidence and mortality in the USA, NHL remains poorly understood and is largely incurable. Recent molecular studies suggest that genomic variations, which can be measured with SNPs (single nucleotide polymorphisms) in genes, may have independent predictive power for prognosis of NHL beyond clinical measurements (Cerhan et al., 2007; Wang et al., 2009; Warzocha et al., 1998). A genetic association study was conducted to identify genomic risk factors with predictive power for NHL prognosis (Zhang et al., 2004, 2005).

Existing association studies of NHL and other cancer prognosis are mostly single-marker based and consist of the following steps. First, for each SNP, a statistical model with survival as response and ‘SNP + clinical risk factors’ as covariates is constructed. Second, for each SNP, its significance level, measured with the P-value from (for example) a likelihood ratio test, is computed. Third, the multiple comparison adjustment is carried out, and SNPs with P-values below a certain threshold are declared as being ‘significantly associated with survival’. Although different studies may differ in terms of statistical models, covariate sets and multiple comparison adjustment methods, a common feature of these studies is that the analysis is carried out one marker at a time. Prognosis of lymphoma is affected by the joint effects of mutations or defects of multiple genes (Knudsen, 2006). Single-marker-based studies analyze each SNP/gene separately. Such analysis may miss genes with weak marginal but important joint effects on prognosis.

A recent study that investigates the joint effects of multiple SNPs is Wu et al. (2009). Their approach consists of the following steps: (i) conduct prescreening and select a relatively small number of SNPs for downstream analysis; (ii) for SNPs passed the first step screening, model their joint effects using a logistic regression model; and (iii) use the Lasso, a penalization approach, for estimation and SNP selection. The ‘logistic regression + Lasso’ approach has been well developed and extensively used in microarray gene expression studies (Ma and Huang, 2005; Ma et al., 2007). Wu et al. (2009) show convincingly that this approach is also applicable to association studies.

In association studies, the marker data has a two-level hierarchical structure: the gene level and the SNP-within-gene level. Compared with SNP-based analysis, gene-based analysis can lead to results that are more interpretable and more reproducible. Thus, genes, instead of SNPs, have been more commonly taken as the functional units. Hence, with the NHL association study, we are interested in identifying predictive genes. In addition, for a specific gene, different SNPs correspond to different segments of the gene. It is of equal interest to identify predictive SNPs within the selected
genes. Therefore, identification of predictive genomic markers in association studies can be considered as a two-level selection problem: selection of predictive genes and selection of predictive SNPs within genes. The CTGDR (clustering threshold gradient directed regularization) approach, which was developed by Ma and Huang (2007) in the context of gene expression analysis, seems a natural choice for such a purpose. The CTGDR can account for the hierarchical structure in covariates and conduct two-level selection.

In this article, we use the CTGDR method to analyze a NHL association study and construct prognosis signatures. This study advances from published ones on the following aspects. First, compared with single-marker analysis, we study the joint effects of multiple SNPs and genes. Such joint analysis can provide additional insights beyond single-marker analysis. Second, compared with Wu et al. (2009), a different regularization method is used for selection of predictive SNPs. The CTGDR approach can accommodate the two-level ‘gene and SNP within gene’ structure, which cannot be achieved with the Lasso. To the best of our knowledge, our method is the first of its kind that attempts to simultaneously identify genes and SNPs within genes in the joint modeling of association data. Third, the data structure considered is significantly different from that in Ma and Huang (2007). Here, we study SNP markers, which are categorical with at most three levels representing three genotypes, whereas gene expressions are considered as continuous measurements. In addition, with gene expression data, there are a relatively small number of clusters, whereas with association data, the number of clusters is the number of genes, which can be considerably large. Fourth, modifications are made to the CTGDR algorithm, in order to better accommodate association data. Lastly, we provide detailed analysis of a NHL association study and construct prognosis signatures for diffuse large B cell lymphoma (DLBCL) and follicular lymphoma (FL), which may provide valuable insights into the genomic factors that differentiate prognosis among NHL patients.

2 ASSOCIATION STUDY OF NHL PROGNOSIS
A genetic association study was conducted to identify genomic variations with predictive power for NHL prognosis (Zhang et al., 2004, 2005). Here, when referring to ‘prognosis’, we limit ourselves to disease-free survival. Overall and other types of survival may have different patterns and different genomic basis, and should be studied separately. We provide details on the prognostic cohort and DNA extraction and genotyping in the Supplementary Material. In this study, a candidate gene approach was adopted. A total of 1462 tag SNPs from 210 genes involved in immune response pathways were selected. In addition, 302 SNPs in 143 candidate genes previously genotyped by Taqman assay (Lan et al., 2006, 2007) were also included. There are a total of 1764 SNPs, representing 333 genes. The number of SNPs per gene ranges from 1 to 45, with median 3 and mean 5.3.

The prognostic cohort is composed of 546 NHL patients. Among them, SNP genotypes for 469 are available for downstream analysis. Among those 469 patients, 150 have DLBCL, 112 have FL, 51 have CLL/SLL, 52 have MZBL, 35 have T-cell lymphoma and 89 have other subtypes. Different subtypes of lymphoma have significantly different prognosis patterns and different genomic basis. We focus on DLBCL and FL due to sample size considerations.

We remove SNPs with >20% missing measurements and patients with >20% SNPs missing. The missingness has been caused by technical reasons, and removal of missing measurements will not lead to biased results. For DLBCL, 133 patients pass the screening. The genotypes of 1730 SNPs are available, representing 331 genes. For FL, 97 patients pass the screening. The genotypes of 1729 SNPs are available, representing 331 genes. For SNPs with missing genotypes, we impute their values using the 10-nearest neighbors approach.

The following clinical risk factors are also measured: age (continuous), education (categorical: level 1 = high school or less; level 2 = some college; level 3 = college graduate or more), tumor stage (categorical: levels 1–5), B-symptom presence (categorical: No; Yes; Unknown) and initial treatment (categorical: none; surgery; radiation; chemotherapy; other). Following creation of the International Prognostic Index, we dichotomize the age variable and create the binary indicator I(age ≥ 60). For categorical risk factors, we use binary dummy variables to represent their levels. Thus, there are a total of 13 covariates representing five clinical risk factors.

When investigating the prognosis of DLBCL or FL, we model the joint effects of clinical and genomic risk factors. Thus, there are a total of 336 ‘clusters’ of covariates. Within each cluster, there are one or multiple covariates. With clinical risk factors, one cluster corresponds to the different levels of a covariate. For example, the ‘education’ cluster has two covariates, which are the binary indicators representing levels 2 and 3 of education. With genomic risk factors, one cluster corresponds to one gene, and covariates within a cluster are SNPs in this gene.

3 REGULARIZED ESTIMATION AND SELECTION
3.1 Data and model setup
Let $Z$ be the covariates that include both clinical and genomic components. Assume there are $M$ covariate clusters and $m_j$ covariates within cluster $j = 1, \ldots, M$. Let $d$ be the length of $Z$, i.e. $d = \sum_{j=1}^m m_j$. Let $Z^j$ be the $j$-th covariate cluster and $Z^d$ be its $j$-th component. When $Z^d = (Z_{1,j}, \ldots, Z_{d,j})$ corresponds to a clinical risk factor, say tumor stage, $Z_{d,j}$ denotes the $j$-th dummy variable representing stage $j$. When $Z^d$ corresponds to a gene, $Z_{d,j}$ denotes measurement of the $j$-th SNP on this specific gene.

Let $U$ and $V$ be the time to collapse and censoring, respectively. We observe $(T = \min(U, V), \Delta = I(U \leq V), Z)$. We assume the Cox proportional hazards model, where the conditional hazard function $\lambda(s|Z) = \lambda_0(s) \exp(\beta' Z)$. Here, $\lambda_0(s)$ is the unknown baseline hazard, $\beta$ is the length, $d$ regression coefficient and $\beta'$ is the transpose of $\beta$. Denote $\beta^j$ as the component of $\beta$ that corresponds to $Z^d$. Assume $n$ i.i.d. observations $X_j = (T_j, Z_j), j = 1, \ldots, n$. The log-partial likelihood function is $\log L(\beta) = \sum_{j=1}^n h[Z_j, \log(\sum_{k\neq j} \exp(\beta' Z_k))]$, where $r_j = I(T_j \geq T_j)$ is the risk set at time $T_j$.

3.2 CTGDR method
We use the CTGDR method for selection of predictive genes and predictive SNPs within those selected genes. We make modifications to better accommodate the unique structure of association data.

3.2.1 Algorithm Let $\Delta v$ be the small positive increment. In the implementation, we set $\Delta v = 10^{-3}$. Denote $0 \leq t_1, t_2 \leq 1$.
as the cluster-level and within-cluster-covariate-level thresholds, respectively. The CTGDR algorithm consists of the following steps.

1. Initialize \( \beta = 0 \).
2. With the current estimate of \( \beta \), compute the length \( d \) vector of gradient \( g = \frac{\partial}{\partial \beta} \). Denote \( g^j/d \) as the component of \( g \) that corresponds to \( \beta^j \).
3. For covariate cluster \( i = 1, \ldots, M \), compute the length \( m_i \) within-cluster-covariate-level thresholding vector \( f^i_1 \), where its \( j \)-th component is \( f^i_1 j = I(\|g^j/d\| \geq \tau_2 \times \max_k \|g^j/d\|) \).
4. Compute the length \( M \) vector of cluster-level gradient \( G \), where its \( i \)-th component is \( G_i = \sum_{j=1}^{m_i} f^i_j |g^j/d| \) (\( i = 1, \ldots, M \)).
5. Compute the cluster-level thresholding vector \( f_2 \), where its \( i \)-th component is \( f_2 i = I(G_i \geq \tau_1 \times \max_i G_i) i = 1, \ldots, M \).
6. Update \( \beta^j = \beta^j + \Delta_1 \times g^j/d \times f^i_1 \times f_2 i \).
7. Steps 2–6 are repeated \( k \) times, where \( k \) is determined using cross-validation.

Without the thresholding steps, CTGDR becomes the simple gradient-searching maximization approach. The thresholding in Step 3 compares a covariate with other covariates within the same cluster, and the thresholding in Step 5 compares a cluster with other clusters. In Step 6, the CTGDR algorithm only updates estimates for those important clusters and important covariates within those clusters. Thus, selection is achieved using thresholding. We refer to Ma and Huang (2007) for detailed discussions of this algorithm.

The cluster-level thresholding (Steps 4 and 5) has been modified from Ma and Huang (2007). This modification consequently affects Step 6 and the final estimate. More specifically, in Ma and Huang (2007), \( G_i = \sum_{j=1}^{m_i} |g^j/d| \), as opposed to \( G_i = \sum_{j=1}^{m_i} f^i_j |g^j/d| \) in this study. That is, in Ma and Huang (2007), all gradients within a cluster are used to compute the cluster-level gradient. In contrast, in this study, only gradients for selected covariates are used. Consider, for example, a covariate cluster with 10 covariates. Among them, only five have predictive power. When investigating the predictive power of this cluster, we should focus on the five predictive covariates.

To further elaborate this point, consider a hypothetical scenario, where we add 10 white noises to the cluster. Intuitively, adding noises should not change the predictive power of this cluster. With the modified algorithm, under both scenarios (with 10 or 20 covariates), the cluster-level gradients are computed using the five predictive covariates. In contrast, with the original algorithm, under the two scenarios, 10 and 20 covariates are used to compute two possibly different gradients. Thus, this modification is reasonable and necessary.

### 3.2.2 Tuning parameter selection

The CTGDR approach has three tuning parameters \( k, \tau_1, \tau_2 \), which are selected using V-fold cross-validation as follows. (i) Randomly partition the data into \( V \) non-overlapping subsets with equal sizes. (ii) Remove subset \( \nu \) from the rest \( V-1 \) subsets. (iii) For subjects in the removed subset, compute the predictive scores \( \beta Z \). Dichotomize those scores at the median and create two risk groups. Compute the log-rank statistic, which measures the difference of survival between those two risk groups.

(iv) Repeat Steps (ii) and (iii) over all \( V \) subsets. Select \( k, \tau_1, \tau_2 \) that maximize the sum of the log-rank statistics. For \( \tau_1, \tau_2 \), we search over the discrete grid of \( 1, 0.95, \ldots, 0.05, 0 \).

The procedure described above differs from that in Ma and Huang (2007) in that the logrank statistic is used instead of the log-likelihood function. In this study, we are interested in whether the genomic and clinical risk factors can predict the relative risk of lymphoma survival. The log-rank statistic provides a direct measure of the relative difference in risk and is hence preferred.

#### 3.2.3 A graphical presentation

We use a simple example to illustrate graphically the important features of the CTGDR. Assume there are three genes and four SNPs per gene. For simplicity, assume there are no clinical risk factors. For each SNP, the measurement takes the value of 0 or 1 with probability 0.7 and 0.3, respectively. The regression coefficient \( \beta = (1.2, 1.0, 0.8, 0.6, -1.2, -1.0, 0.0, 0.0, 0.0, 0.0) \). Survival times are generated from the Cox model with baseline hazard function equal to 1, and censoring times are generated independently as \( \exp(0.5) \) distributed. We simulate 100 i.i.d. samples.

All four SNPs within gene 1 have predictive power, two out of four SNPs within gene 2 have predictive power, and none of SNPs within gene 3 have predictive power. Within genes 1 and 2, different SNPs have different regression coefficients and hence different predictive power. Although smaller, this example shares the same key characteristics with real large-scale studies.

The 4-fold cross-validation selects \((\tau_1, \tau_2, k) = (0.95, 0.95, 428)\). For \((\tau_1, \tau_2) = (0.95, 0.95)\), we show the CTGDR estimates as a function of \( k \), the number of iterations, in Figure 1.

As can be seen from Figure 1, the CTGDR has the following operating characteristics. In Step 2, the gradients, which measure the predictive power of SNPs, are computed. Predictive SNPs are expected to have gradients larger than noisy ones. In Step 3, within each gene or in general a covariate cluster, the thresholding...
Thus, we are able to discriminate not only predictive genes from subjects in the testing set. We compute the predictive scores in addition, we use the CTGDR estimate to make predictions for known. Hence, we are able to look at the identification accuracy. cross-validation. In the simulation, true predictive covariates are training set. Optimal tuning parameters are selected using 4-fold and a testing set with sizes 2:1. We use the CTGDR to analyze the steps: (i) randomly partition the data into a training and a testing set only. Note that, in order to conduct a fair evaluation, we simulate 200 replicates and present summary results in the one described above. The only difference is that, prior to each repeat Steps (i–iii) 200 times and obtain 200 prediction logrank statistics. Under the null when the covariates do not have predictive power, the logrank statistics are asymptotically $\chi^2$ distributed. However, when $n < c d$, it is not clear whether the asymptotic result still holds. Thus, we generate the reference distribution of the prediction logrank statistics have median 5.25 [SD = 4.1; interquartile range (2.3, 6.33)], whereas the permutation logrank statistics have median 0.47 [SD = 1.53; interquartile range (0.10, 1.34)]. In addition, we note that the distribution of permutation logrank statistics is close to its asymptotic distribution. We conclude from Figure 2 that the prognosis signature identified using the CTGDR has satisfactory prediction performance.

### 4 NHL PROGNOSIS SIGNATURES

#### 4.1 DLBCL

**4.1.1 Prognosis signature** We apply the CTGDR approach. The 4-fold cross-validation selections $(r_1, r_2, k) = (0.5, 1.0, 640).$ With the optimal tunings, 96 covariate clusters representing 114 covariates are identified as having predictive power. Among the clinical risk factors, only age is predictive. Among the 331 genes and 1730 SNPs, 95 genes and 113 SNPs are predictive. We provide details on the identified prognosis signature in the Supplementary Material. We find that, the CTGDR is capable of identifying a small number of genes and a small number of SNPs within those genes. The regression coefficients for SNPs within the same genes are not necessarily the same. For example SNP ADAM19_01 and ADAM19_10, both of which belong to gene ADAM19, have regression coefficients $-0.0425$ and $-0.0098$, respectively.

**4.1.2 Evaluation** We evaluate the following two aspects of the prognosis signature. **Prediction performance:** we use a resampling approach to evaluating the prediction performance. This approach consists of the following steps: (i) randomly partition the data into a training and a testing set with sizes 2:1; (ii) compute the CTGDR estimate using the training set only. Note that, in order to conduct a fair evaluation, a new set of tuning parameters need to be computed using cross-validation; (iii) make predictions for subjects in the testing set. Compute the predictive scores $\beta Z$, dichotomize the scores at the median and create two risk groups. Compute the logrank statistic, which measures difference in survival between the two groups. Repeat Steps (i–iii) 200 times and obtain 200 prediction logrank statistics.

We show the logrank statistics in Figure 2. For a better view, we only plot the estimated densities. For reference, we also plot the estimated density of 200 $\chi^2$ distributed random numbers. We can see that the prediction and permutation logrank statistics are well-separated ($P$-value from two-sample Wilcoxon test < 0.001). The prediction logrank statistics have median 5.25 [SD = 4.1; interquartile range (2.31, 6.33)], whereas the permutation logrank statistics have median 0.47 [SD = 1.53; interquartile range (0.10, 1.34)]. In addition, we note that the distribution of permutation logrank statistics is close to its asymptotic distribution. We conclude from Figure 2 that the prognosis signature identified using the CTGDR has satisfactory prediction performance.

#### 3.2.4 Simulation

We conduct simulation to evaluate the CTGDR method. We set the values of covariates equal to those observed on the DLBCL patients. There are a total of 133 subjects and 1743 covariates, among which 13 are clinical covariates representing 5 risk factors, and 1730 are SNP measurements representing 331 genes. Among the 336 covariate clusters and 1743 covariates, 60 clusters and 120 covariates are assumed to be predictive for prognosis. In the Cox model, the 120 non-zero regression coefficients are generated as $N(0, 0.15, 0.02) / N(0, 0.15, 0.02)$ distributed with equal probabilities. The censoring times are generated independently from $\text{exp}(0.8)$. We randomly partition each simulated dataset into a training set and a testing set with sizes 2:1. We use the CTGDR to analyze the training set. Optimal tuning parameters are selected using 4-fold cross-validation. In the simulation, true predictive covariates are known. Hence, we are able to look at the identification accuracy. In addition, we use the CTGDR estimate to make predictions for subjects in the testing set. We compute the predictive scores $\beta Z$, dichotomize the scores at the median and create two risk groups. We compute the logrank statistic which measures difference in survival between the two groups.

For comparison, we also consider the TGDR approach, which has been previously used in genomic studies (Ma and Huang, 2005). The TGDR has comparable performance as other one-level selection methods such as the Lasso.

We simulate 200 replicates and present summary results in Table 1. We can see that the CTGDR identifies a smaller number of covariates and clusters, although the differences are not significant. By accounting for the hierarchical structure, the CTGDR identifies significantly more true positives. In addition, the CTGDR prediction significantly outperforms that of the TGDR.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Median SD</th>
<th>Median SD</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall count</td>
<td>138 39</td>
<td>156 31</td>
<td>0.19</td>
</tr>
<tr>
<td>Cluster</td>
<td>80 15</td>
<td>68 14</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>True positive</td>
<td>80 15</td>
<td>68 14</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Logrank</td>
<td>6.59 4.05</td>
<td>4.85 4.07</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Overall count: number of covariates or clusters identified. True positive: number of identified covariates or clusters truly with non-zero coefficients. Logrank: logrank statistic. From the two-sample Wilcoxon test.
We note that, the reproducibility evaluation is a byproduct of the prediction evaluation and incurs no additional computational cost.

We can see from Figure 3 that, most CTGDR identified covariates (in the analysis using all observations) have larger probabilities of being identified using resampled data (probabilities range from 0.11 to 0.89 with mean 0.37). In contrast, those not identified tend to have smaller probabilities (range from 0 to 0.34 with mean 0.02). Thus, we conclude that the CTGDR identified covariates are reasonably reproducible.

4.1.3 Biological findings
We search PubMed and Comparative Toxigenomics Database for evidences of associations between identified genes and lymphoma prognosis.

Many identified genes have been previously identified as NHL susceptibility genes, including genes BCL6, BRCA1, BRCA2, CASP3 and others. There are also genes that have been identified as markers for cancers other than lymphoma, including genes IRF3, PPAR, ALOX12, ALOX15 and others. Our study suggests that they may also be associated with development of DLBCL. Of note, since a candidate gene approach was adopted, it is not surprising that the identified genes are biologically meaningful. However, there are also genes that have not been identified as DLBCL associated. Further biomedical studies are needed to fully understand the associations between those genes and DLBCL prognosis. Details on biological findings are presented in the Supplementary Material.

For the identified genes, we search KEGG for their pathway information. The results are provided in the Supplementary Material. The identified genes represent the following pathways which have been established as associated with prognosis of multiple cancers or only with prognosis of lymphoma: apoptosis, multiple signaling pathways (B-cell receptor signaling pathway, calcium signaling pathway, Jak-STAT signaling pathway, MAPK signaling pathway, p53 signaling pathway, PPAR signaling pathway, toll-like receptor signaling pathway, VEGF signaling pathway and Wnt signaling pathway), cytokine–cytokine receptor interaction, focal adhesion, leukocyte transendothelial migration and natural killer cell-mediated cytotoxicity.

4.1.4 Additional statistical analysis
To gain more insights into the CTGDR approach and NHL data, we conduct the following additional analysis.

In the first set of analysis, we seek to evaluate the additional predictive power provided by genomic measurements. We only consider clinical risk factors and evaluate their predictive power using the resampling approach described in Section 4.1.2. The resampling logrank statistics have median 2.46 [SD = 1.33; interquartile range (1.02, 3.07)], which are considerably smaller than those when both genomic and clinical measurements are used. The two-sample Wilcoxon test of the 200 logrank statistics obtained here versus those obtained in Section 4.1.2 yields a P-value < 0.01, which suggests that including the genomic measurements can significantly improve prediction performance.

In the second set of analysis, we consider a single-marker approach. For each SNP, we fit a Cox model with 'SNP+clinical measurements' as covariates, and compute the P-value for the SNP. We rank those P-values, with the smallest P-value having rank 1. For SNPs identified with the CTGDR, we show their P-values and ranks in the Supplementary Material. We can see that certain CTGDR identified SNPs have very high ranks. For example, SNPs...
We find that most CTGDR identified covariates (in the analysis using 4-fold cross-validation selects (\(20\,4\))

4.2.1 Prognosis signature
We apply the CTGDR approach. The CTGDR and TGDR identify 75 common covariates, representing 68 covariate clusters. We also evaluate prediction performance of the TGDR using the approach described in Section 4.1.2 and find that the prediction logrank statistics have median 3.13 [SD = 4.10; interquartile range (1.70, 5.06); estimated density shown in Fig. 2]. The Wilcoxon two-sample test suggests that they are significantly smaller than those obtained using the CTGDR (\(P\)-value 0.046).

The second and third set of analyses suggest that the CTGDR can identify predictive SNPs significantly different from alternative approaches. In addition, the CTGDR has better prediction performance than the TGDR by accounting for the hierarchical structure.

4.2 FL
4.2.1 Prognosis signature
We apply the CTGDR approach. The 4-fold cross-validation selects \((τ_1, τ_2, k) = (0.6, 1, 169)\). With the optimal tunings, 132 covariate clusters representing 188 covariates are identified as having predictive power. Among the clinical risk factors, only age is predictive. Among the 331 genes and 1729 SNPs, 131 genes and 187 SNPs are identified. More detailed results are provided in the Supplementary Material.

4.2.2 Evaluation
We evaluate prediction performance and reproducibility using the resampling approach.

Prediction performance: for the evaluation of prediction, a plot similar to Figure 2 is obtained and presented in the Appendix (Supplementary Material). The prediction logrank statistics have median 5.29 [SD = 4.09; interquartile range (3.73, 7.66)] and are well separated from the permutation logrank statistics [median = 0.52; SD = 1.33; interquartile range (0.12, 1.40)]. We conclude satisfactory prediction performance of the prognosis signature.

Reproducibility: for the evaluation of reproducibility, a plot similar to Figure 3 is obtained and presented in the Supplementary Material. We find that most CTGDR identified covariates (in the analysis using all observations) have larger probabilities of being identified using resampled data (probabilities range from 0.90 to 1.00 with mean 0.46). In contrast, those not identified have smaller probabilities, ranging from 0.0 to 0.48 with mean 0.05. We conclude that the CTGDR identified covariates are reasonably reproducible.

4.2.3 Biological findings
Interrogation of PubMed and Comparative Toxicogenomics Database suggests similar biological findings as those for the DLBCL. Details on identified genes are presented in the Supplementary Material.

The identified genes represent the following KEGG pathways, which have been previously established as associated with the prognosis of lymphoma and other cancers: apoptosis (including genes CASP9, IL1A, IL3, IRAK2, IRAK3), cytokine–cytokine receptor interaction (genes IFNGR2, IL10, IL15, IL1A, IL3, IL4R and several others), focal adhesion (RAC1, RAC2), leukocyte transendothelial migration and multiple signaling pathways (B-cell receptor signaling pathway, calcium signaling pathway, Fc epsilon RI signaling pathway, insulin signaling pathway, Jak-STAT signaling pathway, MAPK signaling pathway, p53 signaling pathway, toll-like receptor signaling pathway, VEGF signaling pathway, Wnt signaling pathway).

4.2.4 Additional statistical analysis
In the first set of analysis, when only the clinical risk factors are used, the prediction logrank statistics have median 2.60 [SD = 1.58; interquartile range (1.12, 3.56)] and are considerably smaller than those when both clinical and genomic measurements are used. The two-sample Wilcoxon test of the 200 logrank statistics obtained here versus those obtained in Section 4.2.2 yields a \(P\)-value < 0.01. We conclude that including the genomic measurements can significantly improve prediction performance.

In the second set of analysis, we consider the single-marker approach. We present the marginal \(P\)-values and ranks in the Supplementary Material. The observations are similar to those with the DLBCL.

In the third set of analysis, we apply the TGDR. A total of 242 covariates are identified, representing 129 covariate clusters. The CTGDR and TGDR identify 153 common covariates, representing 112 clusters. We evaluate prediction performance of the TGDR and find that the prediction logrank statistics [median = 2.87; SD = 4.42; interquartile range (1.73, 5.58); estimated density shown in Fig. 4] of the Supplementary Material are significantly smaller than those using the CTGDR (\(P\)-value from Wilcoxon test < 0.001).

With the additional analysis, we can conclude better prediction performance of the CTGDR and its ability to identify genes/SNPs missed by alternative approaches.

5 DISCUSSION
We adopt the CTGDR approach for the two-level gene and SNP-within-gene selection and show that it has satisfactory performance with association data. There are several other approaches that are also capable of two-level selection (Huang et al., 2009; Ma et al., 2009; Wang et al., 2009), however, they have only been studied under settings significantly different from association studies. Investigation of their performance under the present data/model setup is non-trivial and will not be pursued here.

In our numerical studies, we compare the CTGDR with TGDR. There are other one-level selection approaches that can be used to analyze association data, for example, the Lasso in Wu et al. (2009). However, their performance with association data and a survival outcome has not been investigated. The TGDR has a statistical framework closest to the CTGDR and has been shown to have performance comparable with other approaches including the Lasso. Thus, we focus on comparison with the TGDR only.

With the CTGDR, we group SNPs based on the genes they belong to. It is possible to further group genes into pathways. Thus, instead of a two-level hierarchical structure, there will be a three-level structure. In principal, it is possible to extend the CTGDR and conduct three-level selection. However, such an extension needs significant methodological development. In this article, our goal is to identify SNPs and genes predictive for NHL prognosis.
Thus, pathway analysis (e.g. gene set enrichment analysis) is not conducted.

We use resampling approaches to evaluating the overall prediction performance of identified signatures and reproducibility of each identified covariate. Some other aspects of the CTGDR may also be of interest. For example, it may be of interest to evaluate the robustness of the CTGDR approach and the relative increment of predictive power contributed by each individual covariate. To the best of our knowledge, such evaluation approaches have not been well developed and are worth separate investigations.

The association study we consider took a candidate gene approach and measured 1764 SNPs. Thus, it is possible to directly model the joint effects of all SNPs. In an array-based genome-wide association study, ∼500000 or more SNPs may be measured. Direct modeling the joint effects of all SNPs may demand prohibitively high computational power. As pointed out by Wu et al. (2009), in those studies, researchers often have a rough estimate of the number of ‘interesting SNPs’. Thus, it is possible to first conduct a prescreening to reduce the number of SNPs, and then carry out analysis of joint effects.

6 CONCLUSIONS

Lymphoma prognosis is affected by both clinical and genomic risk factors. In this article, we analyze an association study, and construct DLBCL and FL prognosis signatures. Since individual SNPs do not have sufficient predictive power, and since progression of lymphoma is affected by the joint effects of multiple genomic defects or mutations, we consider the joint modeling of multiple genes/SNPs and clinical risk factors. Since data obtained in association studies have the natural two-level gene and SNP-within-gene hierarchical structure, we adopt the CTGDR approach for selection at both levels.

We identify prognosis signatures for DLBCL and FL. Other subtypes can be analyzed in a similar manner. However, since sample sizes for other subtypes are very small, we do not pursue such analysis. Our analysis suggests that (i) the CTGDR can identify a small number of genes and a small number of SNPs within genes with satisfactory prediction performance; (ii) many identified genes and their corresponding pathways have been previously established as associated with lymphoma prognosis; (iii) including the genomic measurements can greatly improve prediction performance (beyond using clinical measurements alone); (iv) the CTGDR can identify genes/SNPs significantly different from those using single-marker or one-level selection approaches; and (v) the two subtypes have significantly different prognosis signatures.

ACKNOWLEDGEMENTS

We would like to thank the associate editor and three referees for very insightful comments, which have led to significant improvement of this article. This research was approved by the DPH HIC. Certain data used in this study were obtained from the Connecticut Department of Public Health. The authors assume full responsibility for analyses and interpretation of these data.

Funding: National Institutes of Health (LM009754, CA120988 and CA62006); National Science Foundation (DMS0805670 and DMS0904181); Yale Cancer Center (Hull Argall & Anna Grant 22067A); National Institutes of Health (Fogarty training grants ID43TW008323-01 and ID43TW007864-01).

Conflict of Interest: none declared.

REFERENCES


Ma,S. et al. (2007) Supervised group Lasso with applications to microarray data analysis. BMC Bioinformatics, 8, 60.


NHL prognosis signatures