Enriching the Analysis of Genomewide Association Studies with Hierarchical Modeling

Gary K. Chen and John S. Witte

Genomewide association studies (GWAs) initially investigate hundreds of thousands of single-nucleotide polymorphisms (SNPs), and the most promising SNPs are further evaluated with additional subjects, for replication or a joint analysis. Deciding which SNPs merit follow-up is one of the most crucial aspects of these studies. We present here an approach for selecting the most-promising SNPs that incorporates into a hierarchical model both conventional results and other existing information about the SNPs. The model is developed for general use, its potential value is shown by application, and tools are provided for undertaking hierarchical modeling. By quantitatively harnessing all available information in GWAs, hierarchical modeling may more clearly distinguish true causal variants from noise.

Genomewide association studies (GWAs) are quickly becoming a popular design for deciphering the genetic basis of complex phenotypes. GWAs first evaluate hundreds of thousands of SNPs across the genome and then follow up on the most-promising SNPs. Gauging which SNPs merit further investigation is extremely important, since SNPs not selected could be false-negative results, whereas those chosen could lead to false-positive associations. The conventional approach entails simply selecting SNPs with the smallest association *P* values from standard maximum-likelihood tests.[1] This approach, however, ignores the extensive information known about the SNPs, such as whether they are in regions previously linked or associated with the phenotype, conserved across species, or functional.

Instead of assuming that every SNP measured in a GWA is a priori equally likely causal, one can quantitatively incorporate existing information about the SNPs into the analysis. For example, one can employ a false-discovery rate on stratified data,[2] rank *P* values on the basis of a weighting function that incorporates prior information[3] (e.g., linkage or association evidence), or weight each SNP's association *P* value by how well it tags other unmeasured SNPs.[4] *P* values derived from these strategies appear to give better rankings than do conventional *P* values.[2–4] The ensuing ranking of results could then be used to determine which SNPs should be further evaluated. As the dimensionality of SNP information grows, however, it may become increasingly difficult to evaluate data with some of these approaches, because of sparse strata.

One can surmount this problem by moving to a hierarchical modeling framework that simultaneously combines various types of a priori information. Previous theoretical and applied work indicates the potential value of hierarchical modeling, especially for evaluation of large amounts of data on a limited number of subjects (i.e., precisely the situation faced by GWAs).[5–9] Related work has also shown how this approach can be used in association studies of candidate genes or regions.[10–15] Here, we extend this approach to GWAs; show, by example, the potential value of hierarchical modeling; and provide tools for undertaking these analyses.

To develop the hierarchical model, first assume that one has undertaken a GWA of the relationship between an enormous number of SNPs (*M* total) and a particular phenotype, which can be quantitative or qualitative. The SNPs are genotyped on the initial population of study subjects (*N* total individuals), and the ensuing data are analyzed to test genomewide for the association between each of the *M* SNPs and the phenotype.

If the phenotype is quantitative, one can test for an association with the *m*th SNP using the linear regression

$$\mathbf{y} = \mu_{\mathrm{m}} + \mathbf{x}_{\mathrm{m}}\beta_m \ , \qquad (1)$$

where $\mathbf{y} = (y_1, \ldots, y_N)$ is a vector of the N subjects' phenotype values, $\mathbf{x_m} = (x_{m1}, \ldots, x_{mN})$ is a vector of the subjects' genotype values for the *m*th SNP coded here in a log-additive manner, β_m is the regression coefficient corresponding to the *m*th SNP, and μ_m is the intercept term. (If the phenotype is qualitative, a logistic-regression model could be used instead of a linear one.) Fitting equation (1) to the data gives the maximum-likelihood coefficient estimate $\hat{\beta}_m$ for the association between SNP *m* and the phenotype (our "first-stage" estimates). The statistical significance of this association can be tested with a Wald statistic, given by $\hat{\beta}_m$ divided by its SE.[16] The *P* values obtained in this manner across all *M* SNPs can then be ranked in ascending order, to decide which SNPs to investigate further.

As noted above, however, this conventional approach ignores existing information about the *M* SNPs and assumes that they are all equally likely to impact the phe-

From the Department of Epidemiology and Biostatistics and Institute for Human Genetics, University of California–San Francisco, San Francisco
Received January 24, 2007; accepted for publication April 30, 2007; electronically published June 26, 2007.
Address for correspondence and reprints: Dr. John S. Witte, Department of Epidemiology and Biostatistics, University of California–San Francisco, 513 Parnassus, Room S965, San Francisco, CA 94143-0794. E-mail: JWitte@ucsf.edu
Am. J. Hum. Genet. 2007;81:397–404. 0002-9297/2007/8102-0020$15.00
DOI: 10.1086/519794

notype. Instead, one can incorporate information about the SNPs into a hierarchical model, in an attempt to improve the ranking of the *P* values for association. In particular, we can add to equation (1) a second-stage linear model for the coefficients β_m

$$\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\pi} + \mathbf{U} \text{ , where } \mathbf{U} \sim N(0,\tau^2\mathbf{T}) \text{ ,} \quad (2)$$

where $\boldsymbol{\beta}$ is a vector of *M* first-stage coefficients, **Z** is an M × K second-stage design matrix that incorporates known information on *K* factors about the SNPs, $\boldsymbol{\pi}$ is a *K*-element column vector of coefficients corresponding to the effects of these *K* factors on the phenotype, and **U** is the error term, assumed to be normally distributed with zero mean and variance $\tau^2\mathbf{T}$. The *ij*th element of **Z** indicates whether SNP *i* exhibits known factor *j,* such as being in a linkage region or functional. An example **Z** matrix is given in table 1 (discussed in detail below). Ultimately, model (2) evaluates the *K* second-stage covariates for their effect on the first-stage estimates through the *K*-element vector π, with error term **U** within a multivariate regression framework. In doing so, this higher-level model provides a "knowledge-based" estimate of the SNP effects, which can be combined with the conventional maximum-likelihood estimates in equation (1) to improve the ranking of results from a GWA.

The *M*-dimensional second-stage variance-covariance matrix $\tau^2\mathbf{T}$ in equation (2) reflects the residual variation in the first-stage regression coefficients after the second-stage covariates are taken into consideration; it can be either estimated iteratively (empirical Bayes) or prespecified by an investigator (semi-Bayes).[17] If the latter, $\tau^2\mathbf{T}$ should reflect the widest range of expected residual effects remaining for each SNP. One can formulate the structure of $\tau^2\mathbf{T}$ in several ways. In the simplest case, one might assume a common variance τ^2 across all SNPs, where **T** is the identity matrix. Alternatively, one can model correlation between nearby SNPs as a function of genetic distance by populating the off-diagonal entries of **T** with positive values.[13]

Our implementation of $\tau^2\mathbf{T}$ does not assume a correlation structure among the SNPs (i.e., the off-diagonal entries in **T** are set equal to 0). This allows for jointly analyzing a large number of SNPs with modest computational time by substituting most matrix operations with vector operations. Assignment of the diagonal values in $\tau^2\mathbf{T}$ is predicated on the idea that SNPs with stronger prior evidence (e.g., in linked regions) should be more heavily weighted. A general form for element t_{mm} of the diagonal of **T** for SNP *m* is

$$t_{mm} = \frac{1}{e^{\nu f(\mathbf{z_{m\cdot}})}} \text{ ,} \quad (3)$$

where $f(\mathbf{z_{m\cdot}})$ represents a weighting function of covariate values at row *m* of **Z**, and ν is a normalizing constant. One may simply choose a column in **Z** that provides a reasonable basis for weighting (e.g., prior linkage or association scores) and assign $f(\mathbf{z_{m\cdot}})$ to be the value at row *m* in that column of **Z**.

Alternatively, one might designate a prior weighting on the basis of a composite model that includes more than one covariate, defining $f(\mathbf{z_{m\cdot}})$ as a weighted sum of the covariates

$$f(\mathbf{Z}_{m\cdot}) = \sum_{i=1}^{K} \omega_i \mathbf{Z}_{mi} \text{ ,} \quad (4)$$

where *K* is the set of covariates with compatible units of measure (e.g., LOD scores) and ω weights the relative importance of the covariates (e.g., on the basis of a factor inversely proportional to the false-positive report probability[18]). A value of zero for the weighting function $f(\mathbf{z_{m\cdot}})$ implies that we do not believe that, beyond information contained in **Z**, SNP *m* is more likely to be associated with the phenotype than is any other SNP. When $f(\mathbf{z_{m\cdot}}) = 0$, equation 3 implies that the second-stage SD is equal to τ, whereas positive values reduce and negative values inflate the second-stage SD relative to τ. Thus, τ serves as a baseline residual SD for the SNP effects.

Because units of measure may vary across definitions of $f(\mathbf{z_{m\cdot}})$, we can normalize the weighting function through the following constant, ν,

$$\nu = \frac{\ln\tau^2 - \ln\rho^2}{\max\left[f(\mathbf{Z}_{m\cdot})\right]} \text{ ,} \quad (5)$$

where ρ denotes the residual precision of our second-stage estimate at the SNP with maximum prior evidence. This constrains the minimum SD across all *M* SNPs to a value specified by ρ. Like τ, ρ can be either prespecified or estimated empirically.

Once **Z** and $\tau^2\mathbf{T}$ have been specified, estimates for the second-stage regression coefficients in model (2) are solved through weighted least squares as

$$\tilde{\boldsymbol{\pi}} = (\mathbf{Z}^{\mathrm{T}}\mathbf{S}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{S}\hat{\boldsymbol{\beta}} \text{ and } \mathbf{S} = [\hat{\mathbf{V}} + \tau^2\mathbf{T}]^{-1} \text{ ,} \quad (6)$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}$ are the conventional maximum-likelihood estimates of the regression coefficients and variance-covariance matrix, respectively, for the *M* SNPs from fitting the linear model (1). We consider the absolute values of $\hat{\boldsymbol{\beta}}$, because a particular allele may either increase or decrease an individual's risk of the phenotype.

Finally, the hierarchical modeling estimate $\tilde{\boldsymbol{\beta}}$, which can be considered a posterior estimate of association for the *M* SNPs in a GWA, is determined as a variance-weighted average of the first- (eq. [1]) and second-stage (eq. [2]) estimates of the coefficients $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\pi}}$,

$$\tilde{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{W})\hat{\boldsymbol{\beta}} + \mathbf{W}\mathbf{Z}\tilde{\boldsymbol{\pi}} \text{ and } \mathbf{W} = \mathbf{S}\hat{\mathbf{V}} \text{ .} \quad (7)$$

Here, **W** is an M × M matrix that determines how much the maximum-likelihood (first-stage) estimates $\hat{\boldsymbol{\beta}}$ are reduced toward the second-stage estimates $\mathbf{Z}\tilde{\boldsymbol{\pi}}$. In particular,

Table 1. Example Second-Stage Design (Z) Matrix for Hierarchical-Modeling Approach

			Indicator Variable by Functional Category							LD Sum by Functional Category					
SNP	Intercept	Conservation	mRNA UTR	Nonsynonymous Coding	Intron	Locus	Synonymous Coding	Conservation	All SNPs	mRNA UTR	Nonsynonymous Coding	Intron	Locus	Synonymous Coding	Linkage Scores
1	1	21	0	1	0	0	0	42	2	1	0	1	0	0	4.4
2	1	32	1	0	0	0	0	31	2	0	1	1	0	0	5.5
3	1	10	0	0	1	0	0	53	2	1	1	0	0	0	4.3
4	1	15	0	0	0	1	0	0	0	0	0	0	0	0	3
5	1	14	0	0	1	0	1	0	0	0	0	0	0	0	2
6	1	9	0	0	0	1	0	0	0	0	0	0	0	0	2
7	1	31	1	0	0	0	0	84	4	0	0	2	1	1	2
8	1	31	0	0	1	0	0	84	4	1	0	0	1	1	2
9	1	31	0	0	1	0	0	84	4	1	0	0	1	1	1
10	1	31	0	0	0	1	0	84	4	1	0	2	0	0	.8
11	1	21	0	0	0	0	1	94	4	1	0	0	1	1	.2

NOTE.—Information about SNPs was obtained from existing resources. To demonstrate how one can use functional annotation among correlated SNPs, we assume that SNPs 1–3 and 7–10 are in LD ($r^2 \geqslant 0.8$) with neighboring SNPs.

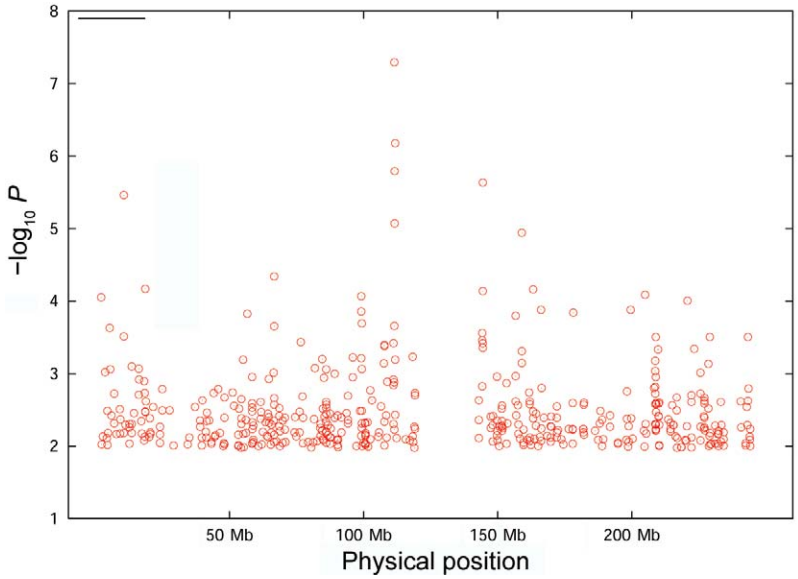

Figure 1. The smallest 500 $-\log_{10} P$ values estimated from ordinary linear regression of the *CHI3L2* gene–expression phenotype on the genotypes of 57 CEU individuals across chromosome 1. The causal SNP *rs755467* is shown at 111.48 Mb with a $\log_{10}$ (*P* value) of 7.29.

if $\hat{\mathbf{V}}$ is large relative to $\tau^2\mathbf{T}$, less weight will be given to $\hat{\boldsymbol{\beta}}$—and more weight will be given to $\mathbf{Z}\tilde{\pi}$—in estimating $\tilde{\boldsymbol{\beta}}$ (and vice-versa). Note that, whereas $\tilde{\boldsymbol{\beta}}$ are not asymptotically unbiased estimators, extensive previous theoretical and simulation work shows that $\tilde{\boldsymbol{\beta}}$ are consistent estimators, and that Wald procedures from $\tilde{\boldsymbol{\beta}}$ work well in typical finite samples.[5,7,9,19,20] Thus, Wald statistics testing $\tilde{\boldsymbol{\beta}}$ can be used to provide GWA rankings on the basis of information from both maximum-likelihood estimates and the additional information contained in the second-stage covariates.

To demonstrate the use and value of hierarchical modeling, we present two examples that are based on data from a GWA between SNPs and gene-expression levels.[21] These data include SNP genotypes from HapMap (International HapMap Project) for 57 unrelated individuals of European ancestry (CEU),[22] the same individuals used in the association study by Cheung et al.[21] We also obtained phenotype information about these individuals for 8,793 gene-expression levels from the Gene Expression Omnibus database at National Center for Biotechnology Information (NCBI) (accession number GSE2552); data were $\log_2$ transformed to alleviate any nonnormal characteristics of the trait distributions.[21]

The first example highlights construction of the second-stage design matrix **Z** with existing information and how to develop a weighting function for the second-stage covariates, as in equation 4. For focus, we studied a region on chromosome 1 where there was strong linkage evidence and an association between the regulatory SNP *rs755467* at the chitinase 3-like 2 (*CHI3L2* [MIM 601526]) promoter and the gene's expression; this finding was confirmed through luciferase reporter and haplotype-specific chromatin immunoprecipitation assays.[21] In light of this finding, we assumed that *rs755467* is causal for *CHI3L2* expression and then compared how well conventional maximum-likelihood and hierarchical-modeling approaches worked to rank SNPs within the surrounding region.

To determine the maximum-likelihood ranking of SNPs, we undertook ordinary linear-regression analyses of the associations between each of 39,186 SNPs on chromosome 1 and *CHI3L2* expression levels (under the assumption of a log-additive genotypic effect). To remove correlated and noninformative SNPs, these SNPs include those on the Illumina 550K SNP panel that were polymorphic in the 57 CEU individuals. Results from this initial ("first-stage") analysis are given in figure 1. In particular, the 500 SNPs with the smallest *P* values for association with *CHI3L2* are plotted in red by chromosomal location, with use of $-\log_{10}$ (*P* values), so high points indicate small *P* values. The smallest association *P* value ($P < 10^{-7}$) is for SNP *rs755467* (the "causal" SNP) at 111.48 Mb near the centromere (i.e., the large gap in the center of the graph).

For the hierarchical model, we incorporated four classes of existing information about the SNPs into a second-stage design matrix **Z**: conservation, functional category, tagging, and linkage. This information is incorporated into 16 columns of **Z**. Table 1 gives examples of this information for 11 hypothetical SNPs. The first column of **Z** corresponds to an intercept and is all ones. Column 2 of **Z** quantifies prior evidence of conservation, since SNPs within conserved regions may be more likely functional.[23] These data, obtained from the conserved elements database at the UCSC Genome Browser Web site, are LOD scores computed from the phastCons program,[24] which

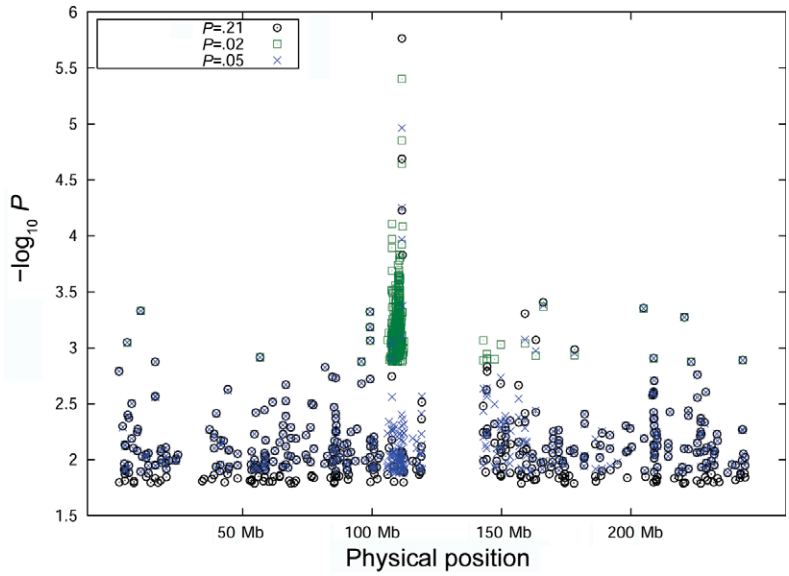

Figure 2. A comparison of the smallest 500 $-\log_{10} P$ values from the *CHI3L2* example with use of hierarchical models across three values of the SD parameter ρ. Larger values of ρ reduce the effect of reduction toward the second-stage mean at the region with strong prior evidence (i.e., linked region in center), whereas smaller values increase the reduction.

assesses the strength of evidence of conservation across 17 species. SNPs located within any region of conserved DNA were assigned the LOD score at that segment. Columns 3–7 of the **Z** matrix contain indicator variables for functional category (i.e., mRNA UTR, nonsynonymous coding, intron, locus, and synonymous coding). Annotation for all SNPs was obtained from the dbSNP, NCBI FTP, and Ensembl sites.

Columns 8–15 in **Z** incorporates information on tagging, since SNPs in linkage disequilibrium (LD) with many other markers may be more likely in LD with causal variants than would SNPs in LD with few markers.[4] Here, we defined SNPs in LD with a given SNP as those mapped within a 500-kb window centered at that SNP, with $r^2 \geqslant 0.8$. We assigned each element in column 8 of **Z** as the total number ("LD sum") of other SNPs in the entire HapMap Phase 2 panel (International HapMap Project) in LD with the SNP at that row.[25] Columns 9–14 of the design matrix combine the LD-sum information with the information described for columns 2–7, to reflect the notion that SNPs in LD with a conserved or functionally important SNP may be distributed differently from SNPs in LD with any SNP in general. Values in column 9 are assigned as the sum of conservation LOD scores for SNPs in LD with the SNP at that row. Values in columns 10–15 are assigned as the total number of functionally annotated SNPs in LD with the SNP at that row, where columns 10–14 are ordered as described for columns 3–7 and column 15 represents SNPs in LD with splice-site SNPs (column 15 of **Z** not shown in table 1). Because these columns are constructed from a dense HapMap SNP panel (International HapMap Project), these columns are particularly informative when a set of SNPs chosen for analysis may not be sufficiently annotated to warrant indicator columns. Finally, the last column of **Z** incorporates prior evidence of linkage. LOD scores were calculated as described elsewhere[26] from linkage analysis of 2,882 SNP genotypes to *CHI3L2* expression, with use of five CEPH families that were unrelated to the 57 individuals in our sample; here, we used the program SOLAR.[27] LOD scores were also incorporated into the diagonal entries of the second-stage covariance matrix **T** by assigning the weighting function $f(\mathbf{z}_{\mathbf{m}\cdot})$ simply as the LOD score for the region in which a particular SNP was located.

Before fitting the hierarchical model, we first estimated an overall second-stage SD τ and a minimum SD ρ. Using equation (2) as the basis of a posterior distribution, we estimated these parameters using the WinBUGS program,[28] which implements a Markov chain–Monte Carlo (MCMC) Gibbs sampler. WinBUGS converged to estimates of $\hat{\tau} = 0.22$ and $\hat{\rho} = 0.21$. To assess the sensitivity of our model to these values, we experimented with other values as well. As can be seen from equation 7, adjusting the value of τ or ρ alters the degree of reduction of the first-stage estimates toward their second-stage estimates. In light of the highly significant LOD scores (>7) for linkage in the same region as the SNP association, the empirical estimate of $\hat{\rho} = 0.21$ might yield a conservative weighting function. This likely reflects a poor fit between the large number of high LOD scores in the **Z** matrix and the small number of statistically significant SNPs at the first stage in this dense data set. Decreasing ρ from 0.21 to 0.05 or 0.02 strikingly increases the influence of the LOD scores on the top-ranked SNPs from the hierarchical model, particularly for those in the linkage region (fig. 2). A visual inspection shows that, in contrast to $\rho = 0.02$, the more

conservative value of $\rho = 0.05$ allows SNPs outside the linkage region that may be potentially interesting to be included in the set of top 500 candidates for follow-up studies.

Therefore, to compare the maximum-likelihood and hierarchical models we used parameter values of $\tau = 0.22$ and $\rho = 0.05$. As above, *P* values from Wald statistics were calculated, and the top 500 SNPs (i.e., those with the smallest *P* values) from each method were plotted (fig. 3). A cursory inspection of the figure shows that, in contrast to maximum-likelihood estimates, a larger proportion of the top-ranked SNPs from the hierarchical model are more consistently clustered around the true causal SNP, whereas SNPs outside the linkage region are included as well. To evaluate this phenomenon more thoroughly, we counted the total number of SNPs that were mapped within windows of various sizes centered at the causal SNP. Figure 4 shows that, in comparison with the maximum-likelihood approach, the hierarchical model increases the proportion of SNPs near the causal variant that are captured, regardless of window size.

Figure 3 also shows that the top-ranked *P* values from hierarchical modeling are slightly larger than those from the single-stage maximum-likelihood approach. This is due in part to reduction of first-stage estimates toward their prior means and is especially apparent in the linked region, because of the stronger effect of the weighting function derived from linkage scores (i.e., smaller values of t_{mm} of $\mathbf{T}$, as shown in eq. [3] for linked SNPs). Note that, despite the smaller *P* values for the maximum-likelihood estimates, many of these putative associations may be spurious, and following them all up may lead to inefficient use of genotyping resources. As illustrated by the horizontal bar in figure 3, if one were to consider a $P < .001$ cut-off when selecting SNPs for follow-up studies, 67 SNPs would be selected when the maximum-likelihood approach was used versus only 17 with the hierarchical model.

The second example explores how information contained in the hierarchical model's second-stage design matrix $\mathbf{Z}$ impacts the ranking of associated SNPs. Here, we focused on the ENCODE regions, which have been resequenced and thus have more-thorough SNP information than do other regions of the genome.[29] In particular, we examined ENCODE region ENm010 (on chromosome 7), because a conventional linear-regression analysis indicates a strong association between SNP *rs11564053* in this region and expression of the cell-cycle progression (*CCPG1*) gene ($P < 10^{-30}$). We evaluated the association between *CCPG1*'s expression and the 758 SNPs in this region on the Illumina 550K panel.

For the hierarchical model, we constructed a second-stage design matrix $\mathbf{Z}$ in the same manner as the first example, although we did not include column 16 (i.e., the linkage column) and other columns, because of lack of data. From WinBUGS, the second-stage SD was estimated as $\hat{\tau} = 0.55$. We set $\rho = \hat{\tau}$, which assumes that the residual second-stage SDs are equal across all SNPs. We then evaluated the sensitivity of the hierarchical model to the covariates included in $\mathbf{Z}$. In particular, we first undertook a hierarchal regression analysis of the association between the 758 SNPs and *CCPG1* expression, including all covariates in $\mathbf{Z}$. We then repeated this analysis, but now only including in $\mathbf{Z}$ subsets of the covariates representing three categories of prior information described above—conservation scores (column 3), functional categories (columns

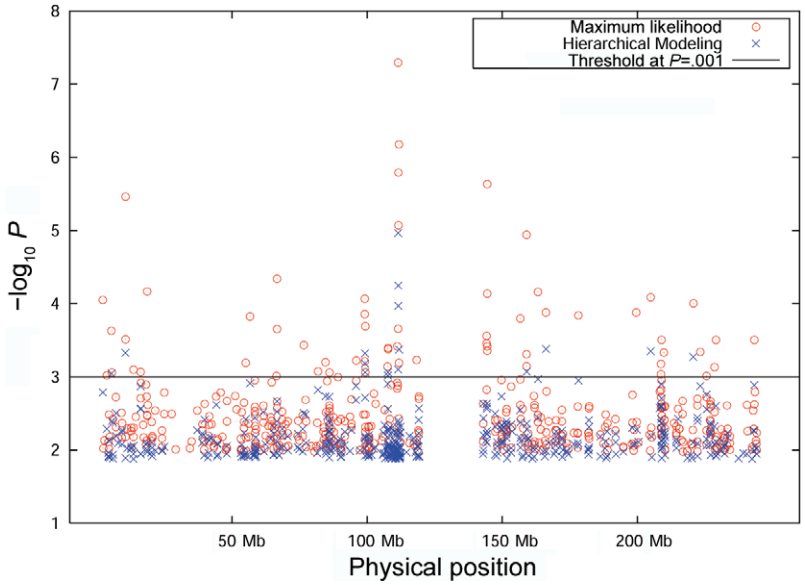

Figure 3. A comparison of the smallest 500 $-\log_{10} P$ values estimated from ordinary linear regression (*in red*, as shown in fig. 1) and the hierarchical model, with $\rho = 0.05$ estimates superimposed in blue.

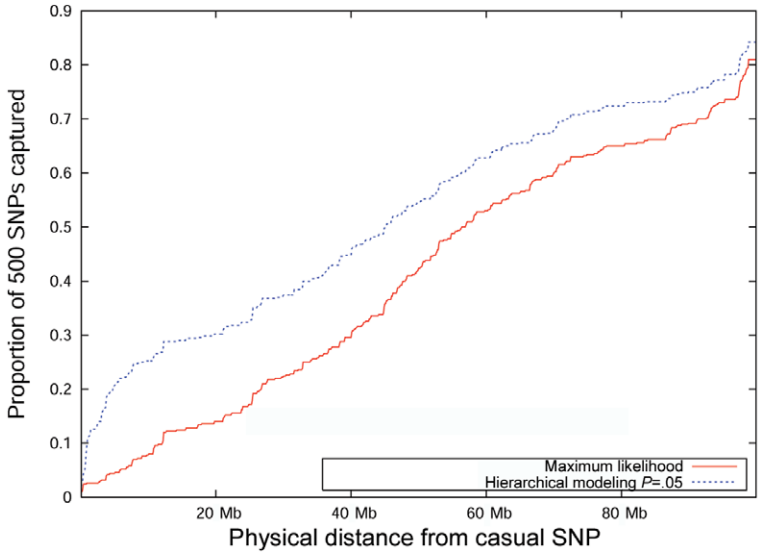

Figure 4. Proportion of the top 500 SNPs located across windows centered at the causal variant for *CHI3L2* gene expression for ordinary linear regression and for the hierarchical model. The *X*-axis denotes the distance from the causal SNP to either edge of a window.

4–6), and LD-sum columns (columns 7–12). The rankings of all 758 SNPs that were based on each of these four **Z** matrix formulations were compared against each other. Using the Kendall-Tau statistic, a nonparametric test for correlation, we found that rankings were significantly correlated ($P < 10^{-7}$) between all six possible pairings of models and hence did not appear to be overly sensitive to the exact formulation of **Z**.

Finally, to assess whether our implementation of hierarchical modeling would yield similar posterior estimates to those provided by an alternate implementation, we revisited the model we designed in WinBUGS. Specifically, we compared hierarchical regression coefficients $\tilde{\beta}$ as obtained from equations (1)–(7) versus those calculated from WinBUGS. The second-stage coefficient estimates $\tilde{\pi}$ were estimated using both methods and substituted into equation (7) to determine $\tilde{\beta}$. Whereas $\tilde{\pi}$ differed slightly between the two methods, they did not lead to materially different $\tilde{\beta}$ estimates; for each of the 758 SNPs, the latter were within 1 SE of each other. Moreover, whereas some of the $\tilde{\pi}$ estimates obtained from the two methods had opposite signs—suggesting opposite effects on the phenotype—these differences appeared limited, because most of these values of $\tilde{\pi}$ were very close to zero.

There are a number of issues to consider with hierarchical modeling of GWAs. Specifying a comprehensive second-stage design matrix **Z** for SNPs in genomic regions with limited annotation will be difficult and can lead to colinearity issues. Fortunately, this will become less of an issue as annotation data become more abundant across the genome. Moreover, our second example and previous work[9] indicate that hierarchical modeling is not overly sensitive to the second-stage design matrix **Z**. One must also be careful in specifying the second-stage residual SD parameters τ and ρ, which are essentially smoothing parameters. These parameters influence posterior estimates of the disease effects by reducing the variance inherent in maximum-likelihood estimates at the cost of introducing some bias.[19] However, for relatively small-scale epidemiologic studies, introducing a certain degree of bias from informative priors can be well justified.[30] Multiple potential values should be considered in the evaluation of the sensitivity of one's results to the second-stage parameter estimation or specification. One can estimate these with an empirical Bayes approach,[7] although we found that doing so resulted in setting them to zero values. Hence, we simply prespecified them with a semi-Bayes approach. We found that an MCMC approach provided us with good starting values for the unknown parameters. Here, visual inspection and subject-matter knowledge about potential residual associations for the SNPs can also help guide sensible values for τ and ρ.[30] For example, given a predetermined number of top-ranked SNPs that can be selected for further study, one might specify a value of ρ that leads to selection of a certain proportion of SNPs in regions with the strongest a priori evidence of association.

In summary, we have illustrated how a hierarchical method can be used to help determine an optimal ranking of SNPs for follow-up in GWAs. By including existing information and borrowing strength from similarities among SNPs in a hierarchical model, one can enrich the overall GWAs signal. We provide resources on the J.S.W. lab home page to help facilitate the development of these models. Future work can use these tools to further study the properties of hierarchical modeling and to apply this approach to GWAs.

Acknowledgments

We thank the reviewers for numerous helpful suggestions and Eric Jorgenson and Sander Greenland for comments on the hierarchical model. This research was funded by National Institutes of Health grants R01 CA88164 (to J.S.W) and R25T CA112355 (fellowship to G.K.C.).

Web Resources

The accession number and URLs for data presented herein are as follows:

Ensembl, http://www.ensembl.org/
Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi (for phenotype data about 57 CEU individuals [accession number GSE2552])
International HapMap Project, http://www.hapmap.org/downloads/index.html.en (for genotype and LD data about SNPs)
J.S.W. lab, http://www.epibiostat.ucsf.edu/witte_lab/
NCBI FTP, http://www.ncbi.nlm.nih.gov/Ftp/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *CHI3L2*)
UCSC Genome Browser, http://genome.ucsc.edu/cgi-bin/hgTables?command=start (for SNP-conservation)

References

1. Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002) Two-stage designs for gene-disease association studies. Biometrics 58:163–170
2. Sun L, Craiu RV, Paterson AD, Bull SB (2006) Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genet Epidemiol 30:519–530
3. Roeder K, Bacanu S-A, Wasserman L, Devlin B (2006) Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78:243–252
4. Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38:663–667
5. Morris C (1983) Parametric empirical Bayes inference: theory and applications. J Am Stat Assoc 78:47–65
6. Greenland S (1992) A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. Stat Med 11:219–230
7. Greenland S (1993) Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. Stat Med 12:717–736
8. Witte JS, Greenland S, Haile RW, Bird CL (1994) Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. Epidemiology 5:612–621
9. Witte JS, Greenland S (1996) Simulation study of hierarchical regression. Stat Med 15:1161–1170
10. Thomas D, Langholz B, Clayton D, Pitkaniemi J, Tuomilehto-Wolf E, Tuomilehto J (1992) Empirical Bayes methods for testing associations with large numbers of candidate genes in the presence of environmental risk factors, with applications to HLA associations in IDDM. Ann Med 24:387–392
11. Witte JS (1997) Genetic analysis with hierarchical models. Genet Epidemiol 14:1137–1142
12. Kim LL, Fijal BA, Witte JS (2001) Hierarchical modeling of the relation between sequence variants and a quantitative trait: addressing multiple comparison and population stratification issues. Genet Epidemiol Suppl 21:S668–S673
13. Conti DV, Witte JS (2003) Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. Am J Hum Genet 72:351–363
14. Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, Witte JS (2004) Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. Cancer Epidemiol Biomarkers Prev 13:1013–1021
15. Liu X, Jorgenson E, Witte JS (2005) Hierarchical modeling in association studies of multiple phenotypes. BMC Genet Suppl 6:S104
16. Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions Am Math Soc 54:426–482
17. Greenland S, Poole C (1994) Empirical-Bayes and semi-Bayes approaches to occupational and environmental hazard surveillance. Arch Environ Health 49:9–16
18. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 96:434–442
19. Efron B, Morris C (1975) Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 70:311–319
20. Greenland S (1997) Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analyses. Stat Med 16:515–526
21. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365–1369
22. The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796
23. Mooney SD, Klein TE (2002) The functional importance of disease-associated mutation. BMC Bioinformatics 3:24
24. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050
25. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265
26. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–747
27. Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211
28. Spiegelhalter DJ, Thomas A, Best NG (1999) WinBUGS version 1.2 user manual. Medical Research Council Biostatistics Unit, Cambridge, United Kingdom
29. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640
30. Thomas DC, Witte JS, Greenland S (2007) Dissecting effects of complex mixtures: who's afraid of informative priors? Epidemiology 18:186–190