Quantitative Trait Nucleotide Analysis Using Bayesian Model Selection

JOHN BLANGERO,[1] HARALD H. H. GÖRING,[1] JACK W. KENT JR.,[1] JEFF T. WILLIAMS,[1] CHARLES P. PETERSON,[1] LAURA ALMASY,[1] AND THOMAS D. DYER[1]

Abstract Although much attention has been given to statistical genetic methods for the initial localization and fine mapping of quantitative trait loci (QTLs), little methodological work has been done to date on the problem of statistically identifying the most likely functional polymorphisms using sequence data. In this paper we provide a general statistical genetic framework, called Bayesian quantitative trait nucleotide (BQTN) analysis, for assessing the likely functional status of genetic variants. The approach requires the initial enumeration of all genetic variants in a set of resequenced individuals. These polymorphisms are then typed in a large number of individuals (potentially in families), and marker variation is related to quantitative phenotypic variation using Bayesian model selection and averaging. For each sequence variant a posterior probability of effect is obtained and can be used to prioritize additional molecular functional experiments. An example of this quantitative nucleotide analysis is provided using the GAW12 simulated data. The results show that the BQTN method may be useful for choosing the most likely functional variants within a gene (or set of genes). We also include instructions on how to use our computer program, SOLAR, for association analysis and BQTN analysis.

In this era of genomic science our current approach to understanding the genetic architecture of a complex phenotype usually follows a specific trajectory. First, the underlying quantitative trait locus (QTL) is localized using a genomic scan of a potentially large chromosomal region. This localization frequently is accomplished either by linkage analysis using data on the cosegregation of phenotypes and genetic markers in families or, less often, by genome-wide studies of phenotype–genetic marker association in sets of unrelated individuals. Second, the chromosomal location is refined by saturating the positional candidate region with additional genetic markers and simultaneously examining both linkage disequilibrium (the effective signal of which spans a much smaller region than the

[1]Department of Genetics, Southwest Foundation for Biomedical Research, 620 NW Loop 410, San Antonio, TX 78245-0549.

Human Biology, October 2005, v. 77, no. 5, pp. 541–559.

KEY WORDS: STATISTICAL GENOMICS, MODEL AVERAGING, SEQUENCE DATA, SINGLE NUCLEOTIDE POLYMORPHISMS, BAYESIAN QUANTITATIVE TRAIT NUCLEOTIDE (BQTN) ANALYSIS.

linkage signal) and linkage to fine-map the QTL. Finally, if positional candidate loci are revealed by this process (perhaps in tandem with bioinformatic data mining) and appear to be in linkage disequilibrium with the putative alleles influencing the trait, we attempt to determine the actual functional variants that are responsible for the observed linkage signal. This final activity takes us from the QTL to the responsible nucleotide differences [the quantitative trait nucleotides (QTNs) (Long et al. 1998; Phillips 1999)] influencing the phenotype. The main corpus of extant statistical genetic methodology is largely focused on the first two tasks, whereas molecular sequencing and functional genetic analyses are traditionally relied on to pinpoint the actual genetic variants involved. In this paper we propose to develop a general approach to statistical functional genomic analysis that will bring rigorous statistical procedures to the final stage of identifying the specific variants involved in determining variation in disease risk. We anticipate that our statistical method will be used to prioritize variants for intensive molecular functional analysis to identify the actual mechanism underlying a variant's effect on a given phenotype. An empirical example of this method that examines functional variation in the Factor VII structural gene is provided in a companion paper [Soria et al. 2005 (this issue)].

Quantitative Trait Nucleotide Analysis

Given complete sequence data for a gene harboring a functional site, we can identify statistically which polymorphism, or polymorphisms, are most likely to be affecting our phenotype. Although determination of the mechanism by which a genetic variant leads to phenotypic variation will still require molecular investigation, it is possible to formulate a first-line statistical genetic approach to limit the number of genetic variants to be examined in the molecular laboratory and to prioritize them in terms of their likely importance in the population.

This approach requires exhaustive enumeration of all polymorphisms within the positional candidate loci and therefore requires initial resequencing of a substantial number of individuals to establish which sites are polymorphic in the population. The number of individuals to be sequenced can be established to reliably detect polymorphisms of a given frequency. However, it is still unknown whether common variants are of major importance as determinants of quantitative trait variation. Given our knowledge of the ubiquity of rare variants in genes and the large literature on multiple rare mutations in monogenic disease, it is likely that rare variants will play a role. In addition, there is growing empirical evidence to support the hypothesis that rare variants are important for human quantitative variation (Blangero 2004). Thus the size of a resequencing sample should be large enough to detect rare variants with a frequency of at least 0.05. Efficient selection strategies for choosing individuals who are more likely to have variant QTL alleles may help to reduce the amount of resequencing necessary

for discovery of relevant single nucleotide polymorphisms (SNPs). Once all polymorphisms are found, they must be typed in a large number of individuals for whom phenotypic information is available (e.g., the extended pedigree sample in which we conducted our linkage analyses). New microarray technologies may make this step much more efficient in the near future.

The QTN Model

The QTN model that we have used represents a simple extension of the classical variance component model. For example, assume that we have a candidate locus with m polymorphic nucleotide sites. Define a variate s_i for the ith SNP that takes the values of 1, 0, and -1 for the marker genotypes AA, Aa, and aa, respectively. In general, the additive genetic variance (σ^2_{ai}) associated with the ith marker is $H_i\alpha_i^2$, where H_i is the heterozygosity and α_i is one-half the displacement between the homozygous marker means. If the ith locus is nonfunctional but is associated with the phenotype because of linkage disequilibrium with the jth marker, which is a functional variant, then

$$\sigma^2_{ai} = \rho^2_{ij} H_j \alpha_j^2 = \rho^2_{ij} \sigma^2_{aj}, \tag{1}$$

where ρ is the correlation between the variables s_i and s_j. ρ_{ij} is also the correlation between the allelic values of the two loci and is thus one of the standard measures of linkage disequilibrium. Note that $\sigma^2_{ai} \leq \sigma^2_{aj}$; that is, the variance associated with a marker will generally be less than that resulting from the functional polymorphism unless the genotypes at the two loci are completely correlated. Using this framework, we model the phenotype as a linear combination of fixed effects and random variables:

$$p = \mu + \sum \alpha_i s_i + \sum \beta_l x_l + \sum q_k + g + e, \tag{2}$$

where the β_l are fixed-effect regression coefficients for any measured covariates (x_l) and the q_k, g, and e are random effects representing other QTLs, residual genetic effects, and random environmental effects, respectively. Estimation of the various fixed effects and variance components associated with the random effects can be performed using standard maximum-likelihood methods, such as those implemented in our computer package, SOLAR (Almasy and Blangero 1998).

Model Selection Using the Bayesian Information Criterion

Once the extent of polymorphism within the gene is assayed, Bayesian model averaging and model selection can be used to predict the most likely functional polymorphisms. We first applied this powerful methodological framework

to the study of multiple QTLs in linkage analyses (Blangero et al. 1999) to allow for a simple statistical method to establish the number of likely QTLs influencing a trait. Because the number of SNPs needed to evaluate a candidate gene may be large, there can be many possible models of QTN action. If we consider only additive QTN effects, there are 2^m possible models. Our approach is to evaluate all such models and to utilize Bayesian methods to estimate the probability that each SNP is functional.

In a Bayesian framework two competing hypotheses can be compared by evaluating the Bayes factor, which is the ratio of the integrated likelihoods of the competing models (Kass and Raftery 1995). Bayes factors provide a direct evaluation of the superiority of one model over another (Kass and Raftery 1995). When the prior probabilities of the models are equal, the Bayes factor is equal to the posterior odds. For the current exposition we assume equal prior probabilities for models, but it is straightforward to use informative prior probabilities, such as assuming a maximum number of functional sites and employing a truncated Poisson distribution to obtain prior probabilities for models of a given dimension.

A number of approximations relating to the Bayes factor have been proposed, of which the Bayesian information criterion (BIC) is both simple and accurate when used with *regular* models (Schwarz 1978; Raftery 1995). The BIC approximation is generally appropriate and (pseudo-) Bayesian inferences can be made with no additional computational burden. For QTN analyses using this approach the BIC is defined with reference to the null model. In the null model there are no fixed QTN effects, but random genetic effects (such as polygenic effects) are allowed to account for nonindependence within families. The BIC of the kth QTN model is given by

$$\mathrm{BIC}_k = -\Lambda_{k0} + \mathrm{df}_k \ln N_e, \tag{3}$$

where Λ_{k0} is the likelihood ratio test statistic comparing the QTN model with the null model, df_k is the degrees of freedom for the comparison, and N_e is the effective sample size. The effective sample size provides an estimate of the number of *independent* observations and can be estimated as

$$N_e = \frac{\hat{\sigma}_p^2}{2\,\mathrm{var}(\hat{\sigma}_p)}, \tag{4}$$

where $\hat{\sigma}_p$ is the maximum-likelihood estimate (MLE) for the phenotypic standard deviation in the null (i.e., polygenic) model. This formulation of the BIC is based on a first-order approximation and has the benefit of computational simplicity. However, more accurate approximations (entailing additional computational burden) exist and could be substituted (Raftery 1996; Neath and Cavanaugh 1997).

The BIC can be used to assess whether the QTN model explains sufficient variation in the phenotype to justify the number of parameters used. In general, BIC differences greater than 2 units are indicative of positive evidence of support

for one model over another with approximate posterior probabilities greater than 75% (Raftery 1995). Similarly, BIC differences of 6 units represent strong support favoring a model with 95% posterior probabilities. BIC differences greater than 10 units are associated with posterior probabilities greater than 99% and thus represent very strong support.

Bayesian Model Averaging in QTN Analysis

The BIC can also be used to formulate a simple model-averaging approach to estimation that explicitly allows for model uncertainty (Raftery 1995; Raftery et al. 1997). Let Y indicate all the data, including both phenotypic and genotypic information, and let M_k indicate the kth model. It can be shown that

$$p(Y \mid M_k) \propto \exp\left(-\tfrac{1}{2}\mathrm{BIC}_k\right). \tag{5}$$

Therefore the posterior probability of the model conditional on the data can be approximated by

$$p(M_k \mid Y) \approx \frac{\exp\left(-\tfrac{1}{2}\mathrm{BIC}_k\right)}{\sum_{l=1}^{K} \exp\left(-\tfrac{1}{2}\mathrm{BIC}_l\right)}. \tag{6}$$

Using this relationship and placing it in the context of QTN analysis, we find that the posterior probability that $\alpha_i \neq 0$ is given by $\sum_{K_i} p(M_k \mid Y)$, where K_i denotes the set of models for which $\alpha_i \neq 0$. This is the posterior probability that the ith SNP is functional (assuming that all genetic variation has been assayed within the candidate locus) or highly correlated with an untyped functional effect. We term this the posterior probability of effect (PPE). The posterior mean α_i is given by

$$E[\alpha_i \mid Y, \alpha_i \neq 0] \approx \sum_{K_i} \hat{\alpha}_i \, p(M_k \mid Y), \tag{7}$$

and the posterior standard deviation is given by

$$\mathrm{Var}[\alpha_i \mid Y, \alpha_i \neq 0] \approx \sum_{K_i} \left[\mathrm{Var}(\hat{\alpha}_i) + (\hat{\alpha}_i)^2\right] p(M_k \mid Y) - E[\alpha_i \mid Y, \alpha_i \neq 0]^2. \tag{8}$$

The main utility of this approach is that it directly takes into account model uncertainty and provides an estimate of our faith that a given SNP is itself functional or in high linkage disequilibrium with a variant not currently assayed.

Application

To evaluate the Bayesian model selection approach to QTN analysis, we analyzed simulated data that we generated as part of GAW12 (Almasy et al. 2001). We simulated 24 pedigrees containing 1,000 phenotyped individuals, STR markers, and approximately 12 kb of sequence data at a positional candidate gene. There was a single functional site at sequence position 5782, which accounted for 24% of the phenotypic variance in the simulated quantitative trait. Our challenge was to see whether we could accurately determine the true functional polymorphism and estimate its effect size. We focused on the quantitative trait, Q1, and one of the actual QTLs (*GENE6*) that influence it. Comprehensive sequence data were available for this gene. Because of the computational burden involved, we present the results from a first replicate of the GAW12 simulation. All analyses were performed using SOLAR, which now incorporates Bayesian model selection of QTN models. These analyses used all 1,000 phenotyped members of the pedigrees and the pedigree information (see appendix).

Results

We selected all SNPs whose less common allele had a frequency of 0.03 or greater. With this criterion we identified 23 polymorphic nucleotide sites. Figure 1 shows the general pattern of disequilibria among all pairs of markers. There are several block-type structures in this simulated data, generating regions of high linkage disequilibrium. The magnitude of disequilibrium was rather high, with an average of 0.36 and a standard deviation of 0.32. Further analysis of the correlation matrix among SNP genotypes allowed us to estimate the effective number of SNPs using the method of Cheverud (2001). From this analysis we estimated the effective number of independent SNPs to be 10.55, which is 45.9% of the total. This number can be used to provide a multiple test correction for standard marginal association tests. Using this approach, we would require a p value of 0.00487 to maintain an experiment-wide Type I error rate of 0.05.

Table 1 shows the positions of each polymorphism, the frequencies of the minor alleles, their heterozygosities, the disequilibrium correlation of each with the functional site, and their distance (in bp) from the functional site. The mean disequilibrium correlation between the SNPs and the functional SNP was a very high 0.51, with a standard deviation of 0.39. Because of this high disequilibrium with the functional variant, 13 of the SNPs (including the functional variant) showed significant evidence of a marginal association with the Q1 phenotypes after controlling for the multiple tests (shown in Figure 2 as signals above the dashed line). Figure 2 also shows the results obtained for the widely used quantitative trait transmission disequilibrium test (QTDT) (Abecasis et al. 2000). Similarly, Table 1 shows the estimated relative variance (h^2_m) associated with each marker. Clearly, it would be easy to detect an association of this candidate gene

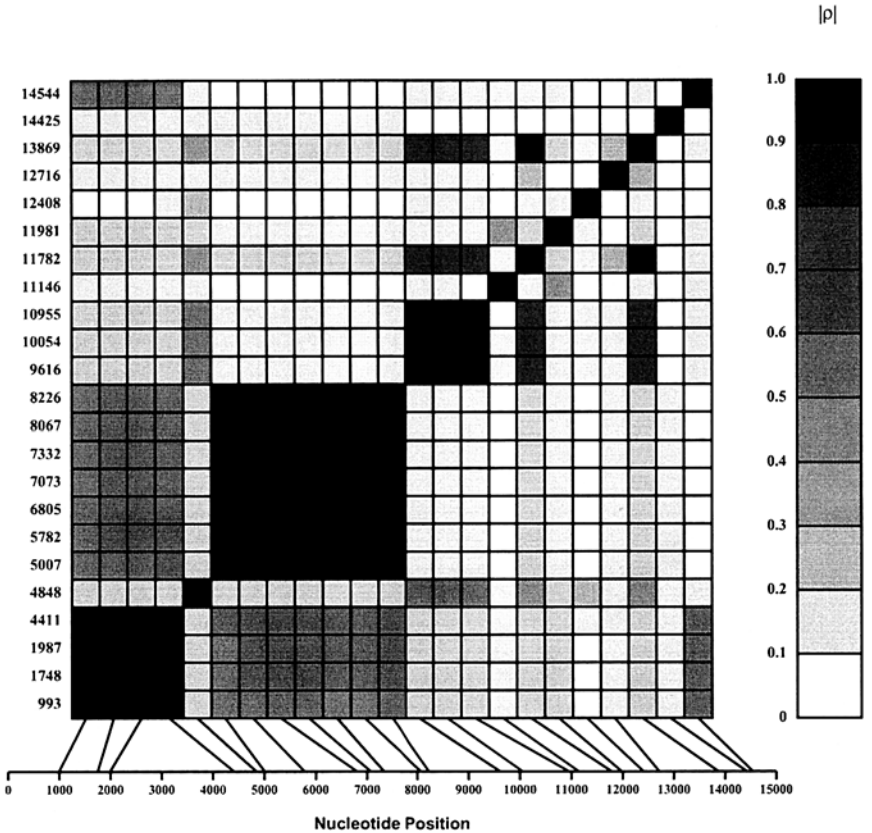

Figure 1. Linkage disequilibrium within *GENE6*.

with the putative QTL. The power to detect even an association that leads to less than 1% of the phenotypic variation with the marker is high in this data set.

As is evidenced by some of the repeated entries in this table, there were several highly correlated blocks of polymorphisms. The examination of the pairwise disequilibrium correlation matrix revealed the following sets of markers that showed near unit correlation. For such a marker sets it is statistically impossible to discriminate among members of the set with regard to the identification of functional effects. We term such sets of markers isocorrelated redundant variant (IRV) sets. In the current large sample we have set a linkage disequilibrium correlation cutoff of 0.975 to define the IRV sets within this gene. For smaller sample sizes it may be necessary to use a lower correlation (say, 0.90) to define such sets because the statistical resolution needed to separate such non-unit-correlated effects is a function of sample size. Using this definition, we enumerated

Table 1. Summary Statistics and Initial Marginal Tests of Association

SNP	*Frequency*	*H*	ρ	*Distance (bp)*	h^2_m
993	0.1958	0.3149	0.6955	4789	0.145
1748	0.1958	0.3149	0.7009	4034	0.145
1987	0.1958	0.3149	0.7009	3795	0.145
4411	0.1937	0.3124	0.7047	1371	0.151
4848	0.3676	0.4649	0.2202	934	0.013
5007	0.1027	0.1843	0.9941	775	0.281
5782	0.1041	0.1865	1.0000	0	0.281
6805	0.1027	0.1843	0.9980	1023	0.281
7073	0.1027	0.1843	0.9980	1291	0.281
7332	0.1027	0.1843	0.9980	1550	0.281
8067	0.1027	0.1843	0.9980	2285	0.281
8226	0.1041	0.1865	0.9961	2444	0.279
9616	0.2317	0.3560	0.1789	3834	0.005
10054	0.2332	0.3576	0.1795	4272	0.005
10955	0.2254	0.3492	0.1847	5173	0.006
11146	0.0384	0.0738	0.0707	5364	0.000
11782	0.2965	0.4172	0.2250	6000	0.015
11981	0.1311	0.2277	0.1489	6199	0.006
12408	0.0695	0.1292	0.0701	6626	0.000
12716	0.0691	0.1286	0.1128	6934	0.008
13869	0.2965	0.4172	0.2237	8087	0.015
14425	0.0622	0.1167	0.1302	8643	0.003
14544	0.0839	0.1537	0.0840	8762	0.002

the following four IRV sets: {993, 1748, 1987, 4411}, {5007, 5782, 6805, 7073, 7332, 8067, 8226}, {9616, 10054, 10955}, and {11782, 13869}. By using only a single representative marker from each of these IRV sets, we effectively reduced the number of SNPs to be evaluated to 11.

The 11 SNPs were then used in the Bayesian model selection procedure. A total of $2^{11} = 2{,}048$ models of QTN action were evaluated. This is a substantial reduction from the $2^{23} = 8{,}388{,}608$ possible models before the establishment of the IRV sets. Table 2 shows the results of the BQTN analysis. Only the IRV sets containing the true causal polymorphic variant at site 5782 show strong evidence of being functional, as reflected by the posterior probability. Because the variant at site 5782 is the true functional variant, the Bayesian QTN analysis has been successful for this replicate. Table 2 also shows the estimates of the α_i for a number of models, including a saturated model, and the results from Bayesian model averaging. Given the true generating value of 3.67, the Bayesian model-averaging procedure provides the most accurate estimate of effect size.

Figure 3 shows the results of the quantitative trait linkage analysis for this trait with a LOD score over 6 positioned at 42 cM. Obviously, there is very strong evidence for a QTL in this simulated example. More important, after

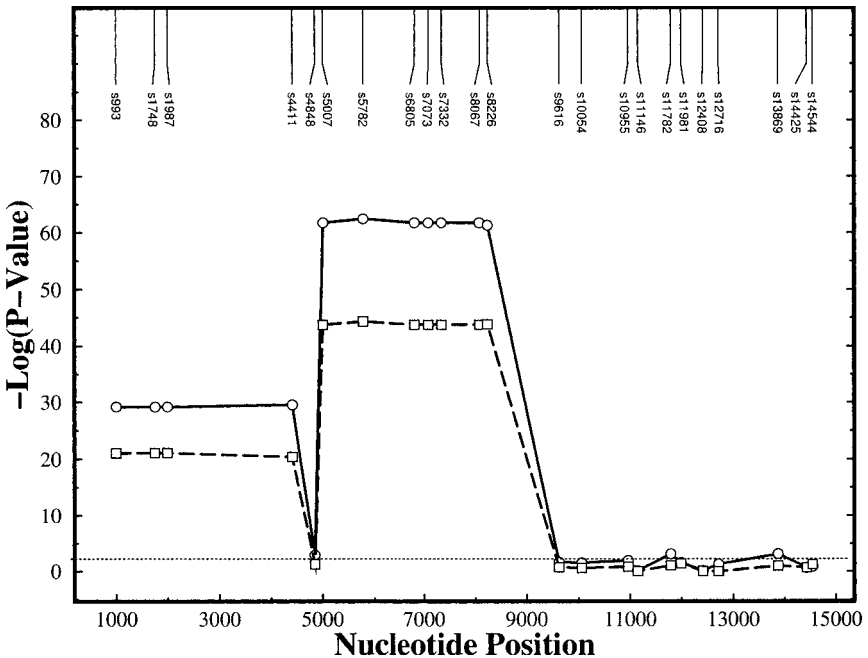

Figure 2. SNP association plot (QTNM plot). Marginal association analysis of simulated SNPs in simulation replicate 1. Circles, measured-genotype analysis; squares, QTDT analysis. Dotted line: $p = 0.00487$ (significance threshold with correction for multiple tests).

Table 2. Results of Bayesian QTN Analysis

	$\hat{\alpha}$		
SNP	*Saturated*	*Bayesian Model Averaging*	*Posterior Probability*
993, 1748, 1987, 4411	1.882	0	0
4848	0.023	0	0
5007, 5782, 6805, 7073, 7332, 8467, 8226	2.254	4.073	>0.9999
9616, 10054, 10955	0.483	0	0
11146	0.189	0	0
11782, 13869	−0.445	0	0
11981	−0.167	0	0
12408	0.229	0	0
12716	0.352	0	0
14425	0.174	0	0
14544	−1.705	0	0

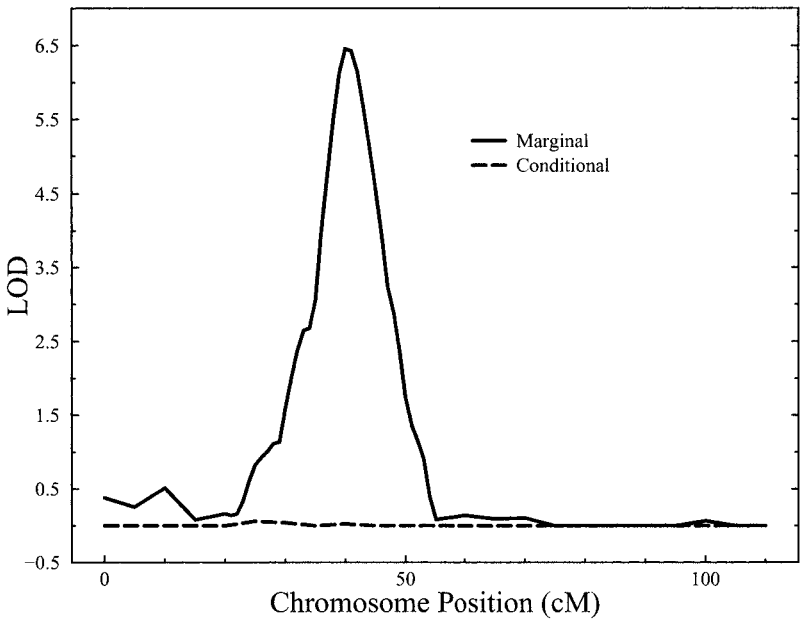

Figure 3. Quantitative trait linkage analyses of the original Q1 phenotype and Q1 conditional on the putative functional polymorphism as determined by BQTN analysis.

choosing the putative functional site using BQTN analysis, we can perform conditional linkage analyses to see whether the linkage signal is eliminated (Sun et al. 2002; Almasy and Blangero 2004). Figure 3 also shows the conditional LOD function when the variant at site 5782 is controlled for. In this case the LOD is completely abolished, suggesting that we have successfully captured all the allelic variation determining the QTL.

Discussion

In this paper we have presented a general method for identifying the most likely functional polymorphisms in positional candidate genes. Although the Bayesian QTN method can be computationally intensive, it appears to have the ability to objectively prioritize variants for more costly molecular functional characterization. For the example, we used complete enumeration of all additive models of gene action. However, for large numbers of variants this approach is intractable. Luckily, there are computationally efficient algorithms to reduce the model search space to those most likely to be important. For most such cases we advocate and use the Up algorithm proposed by Madigan and Raftery (1994).

However, Monte Carlo Markov chain methods also exist to sample the appropriate model space (Raftery et al. 1997).

Although we have shown only simulation-based results for a gene with a single functional site, the BQTN method can be used to accurately identify (or, more accurately, prioritize) multiple likely functional polymorphisms within genes. For example, in a companion paper [Soria et al. 2005 (this issue)], we dissect the effects of multiple variants in the Factor VII structural gene on Factor VII clotting levels. The BQTN method also has been recently used to identify the novel gene SELS (selenoprotein S) as a major player in the mediation of plasma cytokine variation (Curran et al. 2005). In that study, after exhaustive resequencing, we identified a promoter variant that has a high likelihood of being functional. Subsequent gold-standard molecular characterization strongly supported our purely statistical prediction (Curran et al. 2005).

With the dramatic improvements in resequencing technologies, it is likely that in the future most studies will routinely resequence a large number of individuals from the linkage sample to identify all polymorphisms within a positional candidate region. If we have prior evidence for particular candidate genes in a linkage region, we may pursue these candidates first in the sequencing and polymorphism discovery effort. Similarly, standard candidate gene studies in sets of unrelated individuals will move toward comprehensive resequencing. The BQTN method will be of great use in both of these situations. By using this approach, statistical prioritization of putative functional variants can lead to substantial cost savings by minimizing the classical wet laboratory analyses required to establish the molecular mechanism of associated DNA variants.

Finally, we have used this paper to provide information on how our computer program, SOLAR, can be used for quantitative trait association analysis, including BQTN and QTDT analyses. All the procedures (including the various plots) are available in SOLAR (see appendix). These techniques can be used in a wide range of study designs, from studies of unrelated individuals to studies of extended pedigrees of arbitrary size and complexity.

Acknowledgments This research was supported in part by the National Institutes of Health through grants AA08403, HL45522, HL28972, HL70751, GM31575, and MH59490. This analysis was SOLAR powered. SOLAR is available at http://www.sfbr.org/sfbr/public/software/solar/index.html.

Received 12 August 2005.

Literature Cited

Abecasis, G. R., S. S. Cherny, W. O. Cookson et al. 2002. Merlin: Rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 30:97–101.

Abecasis, G. R., W. O. Cookson, and L. R. Cardon. 2000. Pedigree tests of transmission disequilibrium. *Eur. J. Hum. Genet.* 8:545–551.

Almasy, L., and J. Blangero. 1998. Multipoint quantitative trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* 62:1198–1211.

Almasy, L., and J. Blangero. 2004. Exploring positional candidate genes: Linkage conditional on measured genotype. *Behav. Genet.* 34:173–177.

Almasy, L., J. D. Terwilliger, D. Nielsen et al. 2001. GAW12: Simulated genome scan, sequence, and family data for a common disease. *Genet. Epidemiol.* 21(suppl. 1):S332–S338.

Blangero, J. 2004. Localization and identification of human quantitative trait loci: King Harvest has surely come. *Curr. Opin. Genet. Dev.* 14:233–240.

Blangero, J., J. T. Williams, S. J. Iturria et al. 1999. Oligogenic model selection using the Bayesian information criterion: Linkage analysis of the P300 Cz event-related brain potential. *Genet. Epidemiol.* 17(suppl. 1):S67–S72.

Cheverud, J. M. 2001. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 87:52–58.

Clayton, D. 2000. *SNPHAP: A Program for Estimating Frequencies of Large Haplotypes of SNPs.* Available at http://www-gene.cimr.cam.ac.uk/clayton/software/

Curran, J. E., J. B. M. Jowett, K. S. Elliott et al. 2005. Genetic variation in selenoprotein S influences inflammatory response. *Nat. Genet.* 37:1234–1241.

Havill, L. M., T. D. Dyer, D. K. Richardson et al. 2005. Quantitative trait linkage disequilibrium (QTLD) test: A more powerful alternative to QTDT for use in the absence of population stratification. *BMC Genet*. (in press).

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90:773–795.

Long, A. D., R. F. Lyman, C. H. Langley et al. 1998. Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics* 149:999–1017.

Madigan, D., and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Stat. Assoc.* 89:1535–1546.

Neath, A. A., and J. E. Cavanaugh. 1997. Regression and time series model selection using variants of the Schwarz Information Criterion. *Commun. Stat. Theory Meth.* 26:559–580.

Phillips, P. C. 1999. From complex traits to complex alleles. *Tr. Genet.* 15:6–8.

Raftery, A. E. 1995. Bayesian model selection in social research. In *Sociological Methodology 1995*, Peter V. Marsden, ed. Oxford: Blackwell, 111–195.

Raftery, A. E. 1996. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* 83:251–266.

Raftery, A. E., D. Madigan, and J. A. Hoeting. 1997. Bayesian model averaging in linear regression models. *J. Am. Stat. Assoc.* 92:179–191.

Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6:461–464.

Sobel, E., and K. Lange. 1996. Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker sharing statistics. *Am. J. Hum. Genet.* 58:1323–1337.

Soria, J. M., L. Almasy, J. C. Souto et al. 2005. The F7 gene and clotting factor VII levels: Dissection of a human quantitative trait locus. *Hum. Biol.* 77(5):561–575 (this issue).

Sun, L., N. J. Cox, and M. S. McPeek. 2002. A statistical method for identification of polymorphisms that explain a linkage result. *Am. J. Hum. Genet.* 70:399–411.

Appendix: SNP Processing and Analysis in SOLAR

The genetic analysis package SOLAR has been extended to include a set of commands for processing and analyzing SNP genotype data. In this appendix we show the commands that were used in the preparation of this paper.

Genotype Loading. The first step in SOLAR SNP processing is to read in the genotype data and the SNP locations. The simulated GAW12 data analyzed in this paper are free of genotyping errors, and the pedigree data are accurate. In actual studies pedigree and genotype cleaning procedures should be carried out before SOLAR SNP processing. The following command was used to load the GAW12 SNP data:

```
solar> load snp gaw12-snps.1 gaw12-snps.map
```

where "gaw12-snps.1" is a SOLAR marker file containing the GAW12 simulated SNP genotypes, and "gaw12-snps.map" is a SOLAR map file in which the SNP locations are given in base pairs.

Allele-Frequency Estimation. SOLAR computes estimates of the SNP allele frequencies as part of the genotype loading process. If the data set includes related individuals, these estimates should be refined using the "freq mle -hwe" command, which computes maximum-likelihood estimates of the frequencies, taking relatedness into account. This command also tests whether the alleles at each SNP are in Hardy-Weinberg equilibrium.

```
solar> freq mle -hwe
Running allfreq for marker 993 ... iter 6 delta loglike = 0.0584
Running genfreq for marker 993 ... iter 8 delta loglike = 0.7915
Running allfreq for marker 1748 ... iter 6 delta loglike = 0.0584
Running genfreq for marker 1748 ... iter 8 delta loglike = 0.7915
...
```

The "snp show" command displays a summary of the SNP data.

```
solar> snp show
genotype file: gaw12-snps.1
location file: gaw12-snps.map

snp    locn(bp)  #typed  %typed  alleles           SE(freq)  HWE p-val
-----------------------------------------------------------------------
993    993       1000    66.8    1 0.8042  2 0.1958  0.015142  0.2259552
1748   1748      1000    66.8    1 0.8042  2 0.1958  0.015142  0.2259552
1987   1987      1000    66.8    2 0.8042  1 0.1958  0.015213  0.2259552
4411   4411      1000    66.8    2 0.8063  1 0.1937  0.015106  0.2115923
4848   4848      1000    66.8    1 0.6324  2 0.3676  0.018655  0.3707339
5007   5007      1000    66.8    1 0.8973  2 0.1027  0.011433  0.0588408
5782   5782      1000    66.8    1 0.8959  2 0.1041  0.011472  0.0518458
6805   6805      1000    66.8    1 0.8973  2 0.1027  0.011433  0.0588408
7073   7073      1000    66.8    1 0.8973  2 0.1027  0.011433  0.0588408
7332   7332      1000    66.8    1 0.8973  2 0.1027  0.011433  0.0588408
8067   8067      1000    66.8    1 0.8973  2 0.1027  0.011433  0.0588408
8226   8226      1000    66.8    2 0.8959  1 0.1041  0.011666  0.0518458
9616   9616      1000    66.8    2 0.7683  1 0.2317  0.015964  0.4589104
```

```
10054  10054     1000    66.8    2 0.7668  1 0.2332  0.015902  0.5165546
10955  10955     1000    66.8    2 0.7746  1 0.2254  0.015884  0.4838533
11146  11146     1000    66.8    1 0.9616  2 0.0384  0.007233  0.3697457
11782  11782     1000    66.8    2 0.7035  1 0.2965  0.017275  0.3287377
11981  11981     1000    66.8    1 0.8689  2 0.1311  0.012873  0.0925272
12408  12408     1000    66.8    1 0.9305  2 0.0695  0.009592  0.9919025
12716  12716     1000    66.8    1 0.9309  2 0.0691  0.009544  0.1791409
13869  13869     1000    66.8    2 0.7035  1 0.2965  0.017275  0.3287377
14425  14425     1000    66.8    1 0.9378  2 0.0622  0.008977  0.1463692
14544  14544     1000    66.8    1 0.9161  2 0.0839  0.010438  0.8477422
```

Conversion of Genotypes to Phenotypes. Before performing the BQTN analysis, the SNP genotypes must be recoded as covariates with the "snp covar" command. These covariates are written to the file "snp.genocov." By default, missing genotypes are inferred from the results of a previously conducted haplotype analysis (this action can be turned off with the "-nohaplos" option). Currently, SOLAR supports haplotyping using either SimWalk2 (Sobel and Lange 1996) or Merlin (Abecasis et al. 2002).

For this paper the input files needed for a SimWalk2 haplotype analysis were created with the "snphap prep sw2" command. The output of the haplotype analysis was collected into a single file, "swhaplos.out," which was then post-processed with the "snphap import sw2 -f swhaplos.out" command to create the file "snp.haplotypes."

Haplotype frequencies can be estimated either by simple counting using the "snphap count" command or by using the E-M algorithm in the program SNPHAP (Clayton 2000). In the latter case the SNPHAP input file is created by the "snphap freq prep" command, and the SNPHAP output is post-processed with the "snphap freq" import command. The haplotype frequencies are stored in the file "snp.haplofreqs." For this paper we used the simple counting method.

The "snphap show" command displays a summary of the SNP haplotypes.

```
solar> snphap show

Total #Haplotypes:       82
Haplotype Diversity:     0.891504

Per Cent Coverage:       80%     90%     95%     99%
#Haplotypes Needed:      8       11      25      63

     91144556778891111111111
     97948078030260011122344
     34814080736210917947845
     8718725327665548801624

Frequency      Cum.Freq.      4562186954
0.199353       0.199353       11111111111111111111111
0.177443       0.376796       11112111111122212111211
0.098060       0.474856       22221222222211111111111
```

```
0.085129        0.559986        11111111111111111211111
0.076149        0.636135        22221111111111111111112
0.072198        0.708333        11112111111111111111111
0.049569        0.757902        11111111111111112112211
0.048851        0.806753        11112111111111111121111
0.046336        0.853089        11111111111111121211111
0.045977        0.899066        11111111111111111111121
0.008980        0.908046        11111111111122212111211
. . .
```

A different covariates file, "snp.qtldcov," must be generated before performing the quantitative trait linkage disequilibrium (QTLD) association test procedure, described later. This file is created with the "snp qtld" command and contains the results of allelic transmission scoring algorithms, described by Havill et al. (2005).

Preliminary Analysis. The "snp ld" command calculates an estimate of pairwise correlations (linkage disequilibrium) among the SNPs. Two files are created, "snp.ld.pos" and "snp.ld.dat," which contain the SNP base-pair locations and pairwise correlations, respectively. The first few lines of the "snp.ld.dat" file for the GAW12 data are shown here:

```
M1      M2      DISEQ
1       1       1.000000
1       2       0.989622
1       3       0.991929
1       4       0.974673
1       5       0.286414
1       6       0.691418
1       7       0.695463
1       8       0.697454
1       9       0.697454
```

The "-plot" option produces a PostScript plot of these correlations (see Figure 1). Isocorrelated redundant variant sets can be identified by inspection from the output of this procedure.

Initial Association Tests. The QTLD test procedure (Havill et al. 2005) estimates SNP association by means of established techniques [measured genotype analysis, quantitative trait transmission disequilibrium test (QTDT) (Abecasis et al. 2000)] plus a modification of QTDT that draws information from founders when population stratification is absent. Before the QTLD test procedure is run, a previously maximized model that includes the focal trait must be loaded as the base model. It is recommended that this model include a linkage component [see Havill et al. (2005) for a discussion]. In addition to the trait of interest, the phenotypes that are loaded must include the QTLD covariates generated by the "snp qtld" command. The QTLD test procedure is then invoked with the command

"qtld." For this paper we loaded the trait data and QTLD covariates with the following command:

```
solar> load phenotypes gpheno.1 snp.qtldcov
gpheno.1: ID ALIVE AGE HHID EF1 EF2 AFFECT AGEON Q1 Q2 Q3 Q4 Q5

snp.qtldcov: FAMID ID b_993 b_1748 b_1987 b_4411 b_4848 b_5007...
```

where "gpheno.1" is a SOLAR phenotypes file containing the GAW12 simulated trait data and "snp.qtldcov" is the name of the file created by "snp qtld."

In the simulated data, trait Q1 is influenced by a QTL located approximately 42 cM from pter on chromosome 19. Therefore we included in our base model the estimated multipoint identical-by-descent allele-sharing matrix for that location. The base model and the output of the QTLD procedure are as follows:

```
solar> load model q1/null1
solar> model
solarmodel 4.0.0
matrix load/data/GAW12/sim/MIBD/GEN1/mibd.19.42.gzmibd1
trait q1
parameter   mean = 16.90497367      Lower 6.69             Upper 30.41
parameter   sd = 3.268729144        Lower 0                Upper 18.98184286
parameter   e2 = 0.3499080734       Lower 0.2563065695     Upper 0.4563065695
parameter   h2r = 0.2846801239      Lower 0                Upper 0.7080627736
parameter   bsex = 1.855859361      Lower -29.65           Upper 29.65
parameter   bage = 0.09887866095    Lower -0.4706349206    Upper 0.4706349206
parameter   bef1 = -0.000740045234  Lower -0.02210064178   Upper 0.02210064178
parameter   h2q1 = 0.3654118028     Lower 0.2991015983     Upper 0.3985030963
covariate sex
covariate age
covariate ef1
constraint e2 + h2r + h2q1 = 1
omega = pvar*(phi2*h2r + I*e2 + mibd1*h2q1)
# mu = \{Mean+bsex*Female+bage*(age-x_age)+bef1*(ef1-x eef1)\}
option StandErr 0
loglike set -1594.645259
solar> qtld
```

P-values

Trait	SNP	Stratification	Measured Genotype	QTDT	QTLD
q1	993	0.115896	6.3567e-30	8.7837e-22	6.4389e-29
q1	1748	0.106483	6.3567e-30	7.9620e-22	6.1535e-29
q1	1987	0.106483	6.3567e-30	7.9620e-22	6.1535e-29
q1	4411	0.197117	2.7526e-30	3.9751e-21	4.6748e-29
q1	4848	0.485439	0.001113	0.044249	0.007125
q1	5007	0.254774	1.6141e-62	1.7729e-44	2.2583e-62
q1	5782	0.251496	3.6050e-63	4.5082e-45	3.2526e-63

```
q1      6805    0.254774    1.6141e-62    1.7729e-44    2.2583e-62
q1      7073    0.254774    1.6141e-62    1.7729e-44    2.2583e-62
q1      7332    0.254774    1.6141e-62    1.7729e-44    2.2583e-62
q1      8067    0.254774    1.6141e-62    1.7729e-44    2.2583e-62
q1      8226    0.216691    4.4402e-62    1.5589e-44    5.1816e-62
q1      9616    0.567318    0.018749      0.163450      0.040319
q1      10054   0.454980    0.024002      0.222060      0.049912
q1      10955   0.483963    0.011313      0.142698      0.027056
q1      11146   0.972205    0.750437      0.780974      0.900958
q1      11782   0.260246    0.000731      0.075211      0.005884
q1      11981   0.609274    0.026093      0.041067      0.059178
q1      12408   0.959506    0.682042      0.779453      0.880560
q1      12716   0.076701    0.047557      0.884763      0.160196
q1      13869   0.217205    0.000731      0.087687      0.006407
q1      14425   0.601825    0.184919      0.184518      0.293773
q1      14544   0.226891    0.129311      0.052552      0.172736
```

Bayesian QTN Analysis. Before the BQTN analysis is performed, both the focal trait and the covariates generated by the "snp covar" command must be loaded as phenotypes. For this paper, we used the command

```
solar> load phenotypes gpheno.1 snp.genocov
gpheno.1: ID ALIVE AGE HHID EF1 EF2 AFFECT AGEON Q1 Q2 Q3 Q4 Q5

snp.genocov: id famid nGTypes snp_993 snp_1748 snp_1987 snp_4411 ...
```

where "gpheno.1" is the GAW12 phenotypes file and "snp.genocov" is the name of the file created by "snp covar." After the phenotype files are loaded, the command "allsnp" adds as covariates any fields bearing the "snp_" prefix, ignoring other fields in the phenotype files (covariates already in the base model will be retained). If not all SNPs are to be included in the analysis, an alternative approach is to add the "-list *snp_list_filename*" option to the "bayesavg" command. The list file should have each desired SNP name on a separate line.

The BQTN analysis itself is started with the "bayesavg -qtn" command (plus the "-list *snp_list_filename*" option, if desired, to restrict the set of SNPs included in the analysis). The partial output from an analysis of the GAW12 data is as follows:

```
solar> load model q1/null1
solar> bayesavg -qtn -nostop -list snp-list
*** Testing covariates: snp_993 snp_4848 snp_5007 snp_9616 ...

*** N is 11
*** Number of models is 2048

*** Maximizing base model cov0 (unsaturated)
*** Loglikelihood of cov0 is -1594.645259

*** Samplesize is 1000
*** Estimated log(n) is 6.9077553
```

```
Model   BIC         Loglike     H2r        H2r SE     bsnp_993   ...
-------------------------------------------------------------- ...
cov0    0.0000      -1594.645   0.2846802  0          0          ...
*** Best BIC in degree 0 is 0.0 for model cov.base
cov1    -122.2208   -1530.081   0.4367194  0          2.3394365  ...
cov2    -3.7212     -1589.331   0.2769106  0          0          ...
. . .
cov2_3_4_5_6_7_8_9_10_11      -212.4575  -1453.878  0.60881    0    ...
*** Best BIC in degree 10 is -216.0017232 for model
cov1_3_4_5_6_7_8_9_10_11
cov1_2_3_4_5_6_7_8_9_10_11    -209.1007  -1452.102  0.6180175  0    ...

*** Sorting output file
*** Maximizing cov3 for standard errors
*** log(n) calculated from cov3 is 6.6341545
*** Re-sorting output file with changed BIC's
*** Number of Models in Window: 1
*** Window: cov3

Component    Average      Std Error      Probability
-----------------------------------------------
H2r          0.60521      0.062889046    1
snp_993      0            0              0
snp_4848     0            0              0
snp_5007     4.0728172    0.2162492      1
snp_9616     0            0              0
snp_11146    0            0              0
snp_11782    0            0              0
snp_11981    0            0              0
snp_12408    0            0              0
snp_12716    0            0              0
snp_14425    0            0              0
snp_14544    0            0              0

*** Averages written to q1/bayesavg_cov.avg
*** Model results written to q1/bayesavg_cov.out
*** Messages written to q1/bayesavg_cov.history
*** Model with best BIC loaded: cov3
```

As indicated at the end of this output, a number of output files were written to the trait directory using the "bayesavg" command. These include the file "bayesavg_cov.history," which is shown here:

```
*** Testing covariates: snp_993 snp_4848 snp_5007 snp_9616 ...

*** N is 11
*** Number of models is 2048

*** Maximizing base model cov0 (unsaturated)
*** Loglikelihood of cov0 is -1608.800021
```

```
*** Samplesize is 1000
*** Estimated log(n) is 6.9077553

*** Best BIC in degree 0 is 0.0 for model cov.base
*** Best BIC in degree 1 is -299.87672072 for model cov3
*** Best BIC in degree 2 is -294.29644544 for model cov3_8
*** Best BIC in degree 3 is -290.23809016 for model cov1_3_11
     *** No models with degree 3 were in window
*** Best BIC in degree 4 is -284.27019488 for model cov1_3_8_11
     *** No models with degree 4 were in window
*** Best BIC in degree 5 is -277.77713160 for model cov1_3_7_8_11
     *** No models with degree 5 were in window
*** Best BIC in degree 6 is -271.21666432 for model cov1_3_4_8_10_11
     *** No models with degree 6 were in window
*** Best BIC in degree 7 is -264.45432904 for model cov1_3_4_7_8_10_11
     *** No models with degree 7 were in window
*** Best BIC in degree 8 is -257.78160376 for model cov1_3_4_5_7_8_10_11
     *** No models with degree 8 were in window
*** Best BIC in degree 9 is -251.07810248 for model
cov1_3_4_5_6_7_8_10_11
     *** No models with degree 9 were in window
*** Best BIC in degree 10 is -244.31124721 for model
cov1_3_4_5_6_7_8_9_10_11
     *** No models with degree 10 were in window

*** Sorting output file
*** Maximizing cov3 for standard errors
*** log(n) calculated from cov3 is 6.6341544
*** Re-sorting output file with changed BIC's

*** Averages written to q1/bayesavg_cov.avg
*** Model results written to q1/bayesavg_cov.out
*** Model with best BIC loaded: cov3
```

Marginal tests for the Bayesian QTN analysis are performed using the "qtnm" command. Both tabular and graphical output are generated with this command (see Figure 2).

Bayesian QTN analysis can also be conducted for the SNP haplotypes. As with the SNP genotypes, the haplotypes are recoded as covariates and read in by the "load phenotypes" command. The haplotype recoding is performed using the "snphap covar" command.