DOI: 10.1534/genetics.105.048090

A Markov Chain Monte Carlo Technique for Identification of Combinations of Allelic Variants Underlying Complex Diseases in Humans

Alexander V. Favorov,*,[1] Timophey V. Andreewski,[†] Marina A. Sudomoina,[†,‡] Olga O. Favorova,[†,‡] Giovanni Parmigiani[§] and Michael F. Ochs**

**Bioinformatics Laboratory, GosNIIGenetika, Moscow 117545, Russia, [‡]Department of Molecular Biology and Medical Biotechnology, Russian State Medical University, Moscow 121552, Russia, [†]Cardiology Research Center, Moscow 121552, Russia, [§]Departments of Oncology, Biostatistics and Pathology, Johns Hopkins University, Baltimore, Maryland 21205-2011 and **Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111*

Manuscript received July 11, 2005
Accepted for publication August 3, 2005

ABSTRACT

In recent years, the number of studies focusing on the genetic basis of common disorders with a complex mode of inheritance, in which multiple genes of small effect are involved, has been steadily increasing. An improved methodology to identify the cumulative contribution of several polymorphous genes would accelerate our understanding of their importance in disease susceptibility and our ability to develop new treatments. A critical bottleneck is the inability of standard statistical approaches, developed for relatively modest predictor sets, to achieve power in the face of the enormous growth in our knowledge of genomics. The inability is due to the combinatorial complexity arising in searches for multiple interacting genes. Similar "curse of dimensionality" problems have arisen in other fields, and Bayesian statistical approaches coupled to Markov chain Monte Carlo (MCMC) techniques have led to significant improvements in understanding. We present here an algorithm, APSampler, for the exploration of potential combinations of allelic variations positively or negatively associated with a disease or with a phenotype. The algorithm relies on the rank comparison of phenotype for individuals with and without specific patterns (*i.e.*, combinations of allelic variants) isolated in genetic backgrounds matched for the remaining significant patterns. It constructs a Markov chain to sample only potentially significant variants, minimizing the potential of large data sets to overwhelm the search. We tested APSampler on a simulated data set and on a case-control MS (multiple sclerosis) study for ethnic Russians. For the simulated data, the algorithm identified all the phenotype-associated allele combinations coded into the data and, for the MS data, it replicated the previously known findings.

IT is generally accepted now that genetic susceptibility to diseases with a complex mode of inheritance is explained by the presence of multiple genes, each conferring a small to moderate contribution to the overall risk (Tabor *et al.* 2002). The complexity increases because similar disease-prone phenotypes may be produced by different genes in the same pathways as well as by alternative sets of genes providing disease heterogeneity. Due to the success of the human genome project (McPherson *et al.* 2001; Venter *et al.* 2001) and the development of high-throughput sequencing and genotyping technologies (Sherry *et al.* 2001; International HapMap Consortium 2003), there has been a rapid increase in the availability of genetic data for numerous polymorphous loci, including SNPs, repeat polymorphisms, and insertions/deletions. This allows the collection of large sets of genetic data, which could be key in the dissection of the genetic basis of complex diseases.

Standard analytical approaches developed for simple etiologies present problems when dealing with complex etiologies involving multiple genes (Thornton-Wells *et al.* 2004). An approach that has shown great promise in areas with similar dimensionality problems is Markov chain Monte Carlo (MCMC) exploration using a Bayesian statistical basis (Gilks *et al.* 1996). Bayesian methods use the MCMC technique to make inferences that take into account a study's data, as well as additional independent information. For instance, if genes were known to be in linkage disequilibrium, a measurement on the variant of one would provide information on the second, whether it was measured or not. Such information could be included through a prior probability distribution. In general, the final inference is represented by a posterior probability distribution, which includes information from the likelihood, derived from the fit of a model to the data, and prior knowledge of the subject encoded in the prior distribution.

In statistical genetics, Bayesian approaches have become popular in recent years as computational power has increased to a point where these methods can be fully utilized. In addition, the completion of the human

[1]*Corresponding author:* Bioinformatics Laboratory, GosNIIGenetika, Fersmana St., 3-1-31, Moscow 117312, Russia. E-mail: favorov@sensi.org

genome project has provided a substantial body of information on gene locations, potential linkages, and SNPs, which are often best incorporated in an analysis by Bayesian approaches (Rannala 2001). Numerous recent examples of the application of Bayesian methods in genetics include population studies, quantitative trait loci mapping, and family-based studies (reviewed in Beaumont and Rannala 2004).

While the analysis of models with potentially complex interaction is not new to statistics and artificial intelligence, the complexity and size of the data analyses we currently face cannot be efficiently tackled with existing methods. In special settings, such as case-control and discordant sib-pair studies with a moderate number of alleles, exhaustive pattern searches can be conducted using multifactor dimensionality reduction (Hahn *et al.* 2003). This method has been effective in identifying a four-way interaction among alleles, but the method is not highly scalable, and one can consider only one pattern, albeit complex, at any given time. Larger model spaces can be explored using statistical model search procedures such as stochastic search variable selection (George and McCulloch 1993). These require a substantial computational effort and often rely on model assumptions that are difficult to test. Recursive partitioning methods are also commonly used to investigate complex interactions. One example is logic regression (Kooperberg *et al.* 2001; Kooperberg and Ruczinski 2005), which can search for multiple patterns, each including interactions. However, most recursive partitioning approaches have a difficult time identifying complex interactions between predictors, when those are not showing significant main effects, a critical feature of epistasis.

Our approach to surmount these obstacles can be outlined as follows. We are interested in searching over a space of candidate pattern sets, in which each pattern can be a complex genotypic pattern with multiple alleles involved. Evaluation of each of the possible candidates is not feasible for realistic problems because of the number of alleles typed. This suggests a stochastic search approach using MCMC technologies (Gilks *et al.* 1996; Robert and Casella 1999; Liu 2001). Implementation requires an *a posteriori* distribution, reflecting the strength of the evidence provided by the data in favor of an association between each pattern included in the pattern set and the phenotype. Our approach is based on a practical approximation to such a posterior, built upon the distribution of a statistic for the nonparametric evaluation of the null hypothesis of no association between the patterns and phenotype. We deal with the confounding of the patterns by a procedure that is the equivalent of a statistical adjustment and that we term "pattern isolation." We say that a pattern is considered isolated from some other patterns if we remove the influence of all these other patterns on the trait level before we consider its association with the level. The algorithm is intended to identify sets of patterns that are associated with the trait when considered in mutual isolation.

TABLE 1

Raw data structure

Individual	Phenotype	Locus 1	Locus 2	Locus *G*
1	0.1	*a*, *c*	*d*, *d*	*f*, *s*
2	0.4	*c*, *f*	*a*, *b*	0, *a*
—	—	—, —	—, —	—, —
i	0.7	*a*, *a*	*c*, *b*	*a*, *c*

Rows correspond to individuals. Columns include a phenotype and a pair of alleles typed on two chromosomes at a given locus. A value of 0 (*e.g.*, locus G, individual 2) indicates that information about the corresponding allele is missing.

METHODS

Overview: The type of allelic patterns we seek are of interest in complex genetic diseases and include multiple alleles that are associated with a trait in combination rather than individually. We consider the general situation in which we have, for each individual, both a list of typed alleles at a fixed set of candidate loci and the phenotype of interest. Our method is based on ranks, so the phenotype can be measured as a continuous variable or as an ordinal categorical variable. While quantitative phenotypic measurements are powerful when available, it is useful in many applications to have a more general methodology that requires comparing individuals only to each other, as is the case with ranks.

Our approach is designed to search for correlations between complex genetic patterns and phenotype. These correlations are captured via differences in the distributions of phenotype across two subsets of the population, defined by whether a certain allelic pattern is present or not. We consider a broad range of possible genetic models by allowing every allele to potentially affect the phenotype irrespective of its counterpart on the other chromosome. For example, our approach covers dominant and recessive models, as well as their combinations. When looking for polygenic disease patterns, an important challenge arises from the fact that it is not sufficient to consider candidate patterns one by one, because one pattern may confound the measurement of association for another. Thus, we seek a set of patterns. While we do not consider explicitly the issue of removing the possible influence of environmental factors on the phenotype, such a generalization is possible by modifying the test statistic used to construct the likelihood.

Data structure: The typical raw data structure to which our algorithm applies is represented in Table 1, where each row corresponds to an individual. Measurements include a phenotypic variable and the results of

TABLE 2

Examples of patterns

Pattern	Locus 1	Locus 2	Locus *G*
1	0, 0	d, 0	0, 0
2	0, 0	*a*, 0	0, 0
3	0, *f*	0, 0	*b*, 0
4	0, *b*	0, 0	0, 0
—	—, —	—, —	—, —

A pattern is a set of allelic variants at multiple loci. Each row illustrates a possible pattern. The chromosomal order is not used in the present implementation of our algorithm.

genotyping a set of loci on the genome. While these would generally be SNPs, genotypes arising from the sequencing of genes or chromosomal regions would produce appropriate data as well. We set no limit to the number of different alleles that can be observed at a locus in the data set and assume that data are available for the two chromosomes at each locus, although we do not distinguish the two chromosomes presently. If we do not have information about an allele, we denote this with a zero in one of the two locations defining the locus.

Allelic patterns: An allelic pattern is defined here as follows. If there are L loci, a pattern is a $2L$-dimensional vector. Each entry corresponds to a locus-chromosome combination. Each value is either a label for a specific allele or a 0, if the variant is irrelevant for the phenotype. Patterns are illustrated in Table 2. We set no limit to the number of loci that can be involved in a pattern. Patterns in a set can be independently contributing to the phenotype or may act in concert. To account for this possibility we consider pattern sets, which are collections of patterns. Patterns are indexed by n and pattern sets by s. The total numbers of patterns is $N$and the total number of pattern sets is therefore $S = 2^N$. To keep the computation manageable, we restrict the search to pattern sets with a fixed number of patterns.

The number of loci involved in a single pattern controls the order of interaction among loci. The number of patterns in a set controls the number of genetic effects that need to be simultaneously considered to avoid masking and confounding effects. To search for pattern sets it is useful to define a data structure, called the pattern presence matrix, indicating whether a certain pattern is present or absent in each individual. This is illustrated in Table 3 and is the basic data structure used in the algorithm. We use the notation y_i for the phenotype of individual i and x_{in} for entry i, n of the pattern presence matrix, indicating whether pattern n is present in individual i. The symbols y and x_n without further subscripts represent the corresponding random variables. If we do not know the value of x_{in}, because we do not obtain the necessary genotypic information concerning the individual i, we omit this individual and the corresponding row in the presence matrix when considering that pattern. Such an individual is included in calculations for other patterns if the genotyping information allows the determination of whether the individual carries that pattern.

Pattern level comparisons: In our approach, the fundamental comparison (henceforth the "atomic" comparison) is between two groups of individuals whose presence matrix rows differ only in one column, *i.e.*, differ only by the presence or absence of a single pattern. This comparison brings about the concept of mutual isolation of the patterns. Geometrically, we could represent all the 2^N pattern configurations in which an individual may fall by vertices of a unitary hypercube of dimension N (see Figure 1). Any pair of vertices that differ only by the presence of a single pattern is connected by an edge. All parallel edges of the hypercube correspond to the same pattern difference between sets. An atomic comparison is a comparison of two adjacent configurations on the same edge of the hypercube. The generic edge is denoted by e and the pattern that is different between the nodes connected by the edge by $n(e)$. The set of all edges associated with pattern n is denoted by E_n, and it includes 2^{N-1} elements, each corresponding to a configuration of all patterns other than n. Statistically, an atomic comparison is a conditional comparison, while a comparison of all parallel edges at once would be a marginal comparison.

TABLE 3

Phenotype and pattern presence matrix

		Pattern presence matrix					
Individual	Phenotype	Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5	
1	0.1	1	0	0	0	0	—
2	0.4	0	1	1	0	0	—
—	—	—	—	—	—	—	—
i	0.7	0	0	0	0	0	—

Each column in the pattern presence matrix corresponds to a pattern, with each entry i, n indicating whether pattern n is present in individual i. For example, pattern 3 (see Table 2) consists of an f allele on chromosome 2 at locus 1 and a b allele on chromosome 1 at locus G.

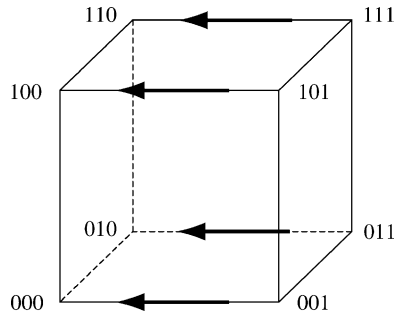

FIGURE 1.—A representation of a hypercube and comparisons between individuals who carry a pattern and those who do not. Each binary number indicates the presence (1) or absence (0) of a specific pattern in the set of patterns. This three-dimensional version represents the case of three patterns in the set.

Pattern $n(e)$ is associated to the phenotype, conditional on the particular configuration implied by e, if the two probability distributions, $p(y|x_1, \ldots, x_n = 0, \ldots, x_N)$ and $p(y|x_1, \ldots, x_n = 1, \ldots, x_N)$, differ. In particular, we say that the pattern is conditionally positively (negatively) associated with the phenotype if the distribution $p(y|x_1, \ldots, x_n = 1, \ldots, x_N)$ is stochastically larger (smaller) (PRATT and GIBBONS 1981) than $p(y|x_1, \ldots, x_n = 0, \ldots, x_N)$, and we say that the pattern is not associated if the two distributions are the same. To represent this association, we define α_e as follows:

$$\alpha_e = \begin{cases} +1 & \text{if } n_e \text{ is conditionally positively associated to the phenotype} \\ -1 & \text{if } n_e \text{ is conditionally negatively associated to the phenotype} \\ 0 & \text{if } n_e \text{ is not conditionally associated to the phenotype.} \end{cases} \quad (1)$$

While this representation does not cover every possible departure from no association, it covers the interesting cases in a parsimonious and nonparametric way. The α_e's are unknown random variables in our analysis.

Pattern set comparison: For pattern n we then define the random variable

$$\beta_n = \begin{cases} +1 & \text{if } \alpha_e = +1 \text{ for } \forall e \in E_n \\ -1 & \text{if } \alpha_e = -1 \text{ for } \forall e \in E_n \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

so that $|\beta_n|$ is 1 if the pattern is an independent predictor of phenotype independent of the configuration of other patterns and 0 otherwise. Finally, for a given pattern set s, we define

$$\gamma_s = \prod_{n \in s} |\beta_n|, \quad (3)$$

so that γ_s is 1 if all the patterns in the set are associated with the phenotype, irrespective of the type of association, and 0 otherwise.

If the β's were known, we could simply form a pattern set by including all patterns such that $\beta_n \neq 0$. In practice, β's can be inferred only statistically. Therefore, discovering associations is tantamount to making statistical inferences about which pattern sets yield $\gamma_s = 1$. A Bayesian analysis of this inference problem would postulate a complete probabilistic model for the data and derive a posterior probability $\pi(\gamma_s = 1|\text{data})$ for every s. Bayesian approaches have been successful in similar subset selection problems (GEORGE and MCCULLOCH 1993; CLYDE *et al.* 1998) and have good properties in terms of protection from overfitting the data and controlling false discovery rates (BENJAMINI and HOCHBERG 1995; EFRON and TIBSHIRANI 2002; MÜLLER *et al.* 2004). There are two practical obstacles to the implementation of a full Bayesian model in this setting. First, the pattern sets cannot be fully enumerated in realistic settings. Second, it would be challenging to find distributional assumptions that are sufficiently flexible to model $p(y|x_n)$ accurately for every n.

To address the first problem we construct our likelihood function by building upon the atomic comparisons and by using the distribution of the rank sum statistic, which is a conservative nonparametric approach. We then develop an approximate posterior distribution and use it to drive a Markov chain Monte Carlo algorithm to sample pattern sets that have high $\pi(\gamma_s = 1|\text{data})$.

Edge-level likelihood: To construct the likelihood, consider edge e. Along this edge we compare the two groups of individuals that differ only by whether $x_{n(e)} = 0$ or $x_{n(e)} = 1$, while all other x's are the same. These two groups include g_e and h_e individuals, respectively. Looking at all pairwise comparisons of ranks across groups, we use the term inversion to refer to a case in which the rank of an individual from the $x_{n(e)} = 0$ group is larger than that of an individual from the $x_{n(e)} = 1$ group. The number of inversions associated with edge e is z_e. The distribution $p(z_e|\alpha_e = 0)$ of this number, conditional on $\alpha_e = 0$, is the Wilcoxon distribution (WILCOXON 1945; MANN and WHITNEY 1947), a well-studied distribution that can be evaluated exactly using recursive formulas (DINNEEN and BLAKESLEY 1973; PRATT and GIBBONS 1981; DI BUCCHIANICO 1996; PRIEBE and COWEN 1999). The Wilcoxon distribution depends only on g_e and h_e and is thus robust to changes in the shape of the distribution of the phenotype y within the two groups. It is symmetrical and is bounded between 0 and $g_e h_e$, with mean $g_e h_e/2$. When $g_e = 1$, it is flat between 0 and h_e, and vice versa. As g_e and h_e get larger, the distribution becomes closer to a Gaussian, although it is always flatter than the binomial distribution with success probability $\frac{1}{2}$ and $g_e h_e$ independent outcomes.

Next we turn to the distribution of z_e when $\alpha \neq 0$, *i.e.*, when the pattern is conditionally associated with the phenotype. To simplify the specification of the

likelihood, we chose to specify this distribution directly, on the basis of a set of conditions that we wish the likelihood to meet:

The two distributions conditional on a positive and a negative effect should satisfy the symmetry property

$$p(z_e = z|\alpha = 1) = p(z_e = g_e h_e - z|\alpha = -1),$$

reflecting the requirement that a given distance from the center $g_e h_e/2$ should provide the same evidence against $\alpha = 0$ irrespective of the side.

$p(z_e|\alpha = 1)$ [or $p(z_e|\alpha = -1)$] should be nonincreasing (or nondecreasing) as the number of inversions z_e increases (decreases).

The sum of the two distributions should be a constant, that is,

$$p(z_e = z|\alpha = 1) + p(z_e = z|\alpha = -1) = c.$$

This is an effective approach that can flexibly capture complex departures from null distributions in other genomics applications even when it does not model the data well (Parmigiani *et al.* 2002).

Both distributions should have flat tails.

The less populated the smaller of the two groups is, the flatter the likelihoods should be, similar to the Wilcoxon distribution.

Both distributions should be flat if $g_e = 1$ or $h_e = 1$.

Among the many possible distributions that meet these requirements, for simplicity, we chose to use the following three-line combination for $p(z_e = z|\alpha = 1)$. The likelihood has a uniform tail of height $2 \cdot (1 - 2^{-\min(h_e, g_e)})/(h_e \cdot g_e + 1)$ on the left outside the interval where $p(z_e|\alpha_e = 0) > 1/(h_e g_e + 1)$ and a uniform tail of height $2 \cdot 2^{-\min(h_e, g_e)}/(h_e \cdot g_e + 1)$ on the right outside the interval. A straight line connects these two tails. The likelihood term $p(z_e = z|\alpha = -1)$ is determined by symmetry.

Pattern-level likelihood: At this level we need to specify a joint probability distribution for the collection of inversion statistics for pattern n,

$$z_n = \{z_e\}_{e \in E_n},$$

conditional on the pattern indicators β_n. Theoretically, these could be computed from the specification in the previous section. However, this computation is not feasible in practice. To obtain a practical approximation that still preserves the important feature of being based entirely on the conditional comparisons rather than on the marginal comparisons, we set

$$p(z_n|\beta_n = 0) \cong \prod_{e \in E_n} p(z_e|\alpha_e = 0) \quad (4)$$

$$p(z_n|\beta_n = 1) \cong \prod_{e \in E_n} p(z_e|\alpha_e = 1) \quad (5)$$

$$p(z_n|\beta_n = -1) \cong \prod_{e \in E_n} p(z_e|\alpha_e = -1). \quad (6)$$

Exact evaluation of the left-hand side of Equation 4 would require a sum over the set of all 2^N possible combinations of α's for which β is zero. However, the formulation above is sufficiently sensitive to permit identifying patterns that are likely to have nonzero β's.

A large fraction of possible patterns do not occur in any of the observed genomes, while others occur in all. The data provide no information about these patterns. For brevity, we refer to these as "uninformative patterns." To improve sampling efficiency, our proposal distribution can be set to ignore these patterns during the sampling.

Posterior probabilities: We specify a prior distribution directly on the β's by assuming that $\pi(\beta_n = 0) = \pi_0$ for every n and that

$$\pi(\beta_n = 1) = \pi(\beta_n = -1) = (1 - \pi_0)/2.$$

We also assume that the β_n's are *a priori* independent.

Our approach here is to derive posterior probabilities and use them to develop a Metropolis algorithm, described in the next section. For each pattern, the posterior probabilities are calculated as

$$\pi(\beta_n = j|z_n) = \frac{p(z_n|\beta_n = j)\pi(\beta_n = j)}{\sum_{j \in \{-1,0,1\}} p(z_n|\beta_n = j)\pi(\beta_n = j)} \quad \text{for } j = -1,\ 0,\ 1. \quad (7)$$

From these probabilities we derive, with a further independence approximation, the posterior probabilities at the level of the set, that is,

$$\pi(\gamma_s = 1|\text{data}) = \prod_{n \in s} [1 - \pi(\beta_n = 0|z_n)]. \quad (8)$$

Because the likelihood function is based on conditional comparisons that evaluate a pattern after removing the effect of other patterns, this independence approximation is plausible in our setting.

Markov chain Monte Carlo sampling and validation: *Sampling from the posterior distribution:* Our algorithm to search for pattern sets that have high $\pi(\gamma_s = 1|\text{data})$ is an adaptation of the Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Robert and Casella 1999). We search pattern sets of a preset size, which is not a serious restriction. In particular, if the number of patterns in the set is overestimated, the posterior probability of all pattern sets will be diluted by the need to include an irrelevant pattern, but the ranking of pattern sets is unlikely to be profoundly altered.

At each step the sampler proposes a new pattern set by one of two alterations of the current pattern set: (1) a change in one allele in one pattern or (2) a recombination of patterns. The choice between these two is random. For case 1, the position of the change is generated randomly from a uniform distribution; then the new allele is generated randomly. The proposal probability of an allele having no influence, the "zero" allele

(see above), is a user-specified input to the sampler. Nonzero allele values are drawn from either a uniform distribution or a Dirichlet distribution on the basis of results of monoallelic pattern tests. For case 2, two patterns are chosen at random, cut at random at either one or two randomly chosen points, and recombined as in genetic recombination.

The proposed pattern sets are accepted or rejected in accordance with the Metropolis-Hastings scheme: the current and proposed states are compared by their sampled posterior values, weighted with the proposal distribution. If the proposed value is higher, the proposed change is accepted. Otherwise, it is accepted with a probability equal to the ratio of the proposed and current pattern set. If the proposed value is not accepted the chain remains at the old pattern set. The resulting sequence of sets forms a Markov chain, whose ergodic distribution is the posterior distribution we are exploring.

Informative patterns form only a small subset of all possible patterns. Hence sets made only of informative patterns are very rare. All other pattern sets, which contain at least one uninformative pattern, are ignored by the proposal distribution. Therefore starting at a random pattern set is impractical, and we choose an informative initial pattern set. We achieve this by selecting a set of patterns, each of which carries one allele that is correlated with trait level.

Initial iterations of the sample are discarded during the burn-in phase, which is terminated after the sampling of a set with posterior greater than a prefixed value (say 0.01). After that we store the posterior probabilities of the best B pattern sets sampled. At each step, we compare the sampled set to the sets in the list of the B best and update the list if appropriate. We also count how often the sampler visits each pattern set in the list.

Simulated data: The sampler was tested on simulated data designed to incorporate the main complexity of allelic interaction while also permitting specific variants at a locus to affect the propensity to disease in different ways. This maximized the genetic overlap to be separated by the algorithm. Three patterns were created, one with three alleles and two with two alleles involving three genetic loci. Each pattern's influence on the target phenotypic feature was characterized by a real number ("role"). The first and third patterns predisposed individuals to the phenotype, while the second was protective. For every pattern, patterns differing only in one allele (shadows) were created. Then, two copies of every pattern and three shadows were distributed randomly among the genotypes for 50 individuals. The patterns and examples of shadows are shown in Table 4.

All empty positions in the genotypes were filled with randomly generated alleles from a collection of alleles for each locus. The only restriction placed on the distribution was that each locus could contain at most two alleles. Then, a trait level was generated for each set, with a level equal to the sum of roles of all patterns contained in the gene set.

TABLE 4

The simulated patterns and examples of shadows used to generate the simulated data for the method test

Patterns	Role	Notation
0 0, *d* 0, *c* 0	1.2	*D*
0 *e*, 0 *c*, 0 0	−4	*E*
f 0, 0 *a*, *b* 0	5	*F*
Shadows	Role	
0 0, *c* 0, *c* 0	0	
0 *e*, 0 *d*, 0 0	0	
f 0, 0 *a*, *a* 0	0	

The strength of their impact for the phenotypic trait level in the data and their notations in Table 5 representing the test results are shown as well.

Tests were made for five different data sets constructed as described above. Five runs of the sampler were made for each set, with different random seeds and starting points. The results were compared to each other and to the input patterns. In addition, noise was added to the phenotypic measure to reflect errors in the estimation of disease level or complexities due to environmental factors. This noise was added to all individuals' level of disease.

Experimental data: We also illustrate our approach on a case-control study. The data consist of genotypes and personal data for 286 unrelated patients with clinically definite multiple sclerosis (MS) and 362 healthy unrelated controls, all of Russian descent. The data included results of genomic typing at polymorphic loci at or near genes of the autoimmune inflammatory response. At chromosome 6p21 there were the *DRB1* gene, repeat polymorphisms of $(AC)_n$ and $(TC)_n$ microsatellites, designated as TNFa and TNFb; SNPs −376A → G, −308G → A, and −238A → G in the *TNF* gene; and SNPs +252G → A and +319C → G in the *LT* gene. At the *TGFβ1* gene (19q13) there were SNPs −509C → T, +72 wild-type → C insertion, +869T → C (10Leu → Pro), +915G → C (25Arg → Pro), and +1632C → T (263Thr → Ile); at the *CTLA4* gene (2q33), −SNP +49A → G (17Thr → Ala); and at the *CCR5* gene (3p21), −wild-type/32-bp deletion. The genotypic data were partially missing; *i.e.*, data on distinct gene polymorphisms for some cases and controls were unknown.

RESULTS

The results of the simulations are summarized in Table 5. The sampler's output reflected the input in terms of matching the pattern, the relative rank of each pattern in terms of its association with the phenotype, and whether the pattern is protective or predisposing, in

TABLE 5

Summary of the results

Position in statistics			
Of the pattern set	Of the pattern (with its role sign)		
	D	*E*	*F*
1	2+	3−	1+
1	2+	3−	1+
1	2+	3−	1+
1	2+	3−	1+
1	2+	4−	1+
1	2+	3−	1+
1	2+	3−	1+
1	2+	4−	1+
1	2+	3−	1+
1	2+	3−	1+
1	2+	3−	1+
1	2+	3−	1+
1	2+	3−	1+
1	2+	3−	1+
1	2+	3−	1+
1	1+	3−	2+
1	1+	3−	2+
1	1+	3−	2+
1	1+	3−	2+
1	1+	3−	2+
1	2+	4−	1+
1	2+	5−	1+
1	2+	5−	1+
1	2+	5−	1+
1	2+	5−	1+

The experiment number is combined from the data set number and starting point number. The numbers represent the position of the original pattern set in the pattern set statistics and the positions of original patterns in pattern statistics with the sign of their influence. In all cases, the pattern set reflecting the true underlying patterns (*i.e.*, Equation 3) is the dominant result. Each pattern has the correct influence as well, with *D* and *F* from Table 4 predisposing and *E* protective.

24 of 25 runs. The first four test data sets gave 18 cases where the three original patterns were ranked as the top three and 2 cases where three patterns appeared as the first, second, and fourth. In both of the latter cases, the third most common pattern was the original predisposing pattern with an additional allele, which appeared slightly more often than the original protective pattern. The fifth data set gave poorer results, with two fake predisposing patterns being more common than the original protective one. The two patterns may have appeared in the simulated data because of a random duplication of one of original patterns. Of significance, however, is the reliability of the full pattern set, which is always identified as the most significant. This reflects the importance of the conditional comparisons shown in Figure 1.

When the samplers were started from different initial points and with different random number seeds for the same data, the same results were obtained, indicating that the sampler appears to be relatively insensitive to starting values. This is especially important in this case given the complexity of the posterior distribution and the steps taken to improve the efficiency of the Markov chain exploration.

TABLE 6

The results for the case with addition of phenotypic noise

s of Gaussian noise	Position in statistics			
	Of the pattern set	Of the pattern (with its role sign)		
		D	*E*	*F*
0	1	1+	3−	2+
0.1	1	1+	3−	2+
0.2	1	1+	3−	2+
0.3	3	1+	2−	3+
0.4	4	1+	3−	2+
0.5	7	2+	4−	1+
0.75	1	4+	2−	1+
1	>25	>25	3−	1+
1.5			14−	1+
2			3−	2+
2.5			13−	1+
3			15−	1+
3.5			>25	2+
4				>25

As can be seen, as the noise level rises, the identification of the pattern set fails, reflecting the necessity of identifying all patterns simultaneously in Equation 3. The statistics gathered on patterns alone still show some recovery of information.

Additional experiments were performed to test the stability of the algorithm for noisy trait level information. A line of data sets with the levels mixed with normally distributed random noise was generated and investigated with the sampler. The results are shown in Table 6. We see patterns shifting away one by one from the list of the best patterns as the noise is increased. It is interesting to note that the identification of the best pattern set failed after the first pattern fails to be found, in keeping with Equation 8, while the description generated solely from patterns still identifies some real features. However, this suggests that for reliable application of the algorithm in real data sets, the pattern set level statistics will provide the most reliable estimate of genetic factors underlying disease.

The algorithm was also applied to the MS case-control study performed on Russian patients with definite MS and healthy unrelated controls of the same ethnicity. The prior probability that a locus has no effect on the phenotype was set to 0.99. We were looking for sets of two or three patterns. The algorithm identified patterns that have a high probability of being associated with MS. All the patterns identified deal with carriership of alleles, without distinguishing homozygotes from heterozygotes. Two of the patterns are the single predisposing alleles HLA class II *DRB1**15(2) and TNFa9 microsatellite. These data represent a validation of

MS associations with *DRB1**15(2) (Boiko *et al.* 2002) and TNFa9 (Gusev *et al.* 1997), which we previously identified in an independent cohort of ethnic Russians. The third pattern identified by the algorithm is a predisposing biallelic combination of *CCR5*Δ32 with *DRB1**04, which was recently described for this data set (Favorova *et al.* 2002).

DISCUSSION

Over the last century, statistical genetics has created powerful tools for the identification of genetic variants leading to disease. The emergence in the last decade of technologies capable of sequencing entire genomes in moderate time frames and of technologies that permit rapid high-throughput measurement of genetic polymorphisms such as SNPs is leading to large data sets that have the potential to unlock the bases of many complex, polygenic diseases. However, the methods developed for small data sets are not easily adapted to the problems that arise with high-dimensionality data. Here we demonstrate a new approach based on two key features: (1) the efficient sampling of the space of all possible genetic pattern sets tied to a phenotypic trait and (2) the use of the Wilcoxon-Mann-Whitney nonparametric statistics to provide a measure of association with phenotype. At this stage, we do not classify all identified patterns by haplotypes, because every pattern regardless of haplotype is a valid pattern.

The consideration above focuses solely on genetic variations and does not consider environmental influences and other possible factors. However, the method can include such variables where appropriate, since the search algorithm is generic and does not rely on any fundamental genetic structures. For instance, environmental factors could be encoded in the same manner as genetic loci without effect on the mathematical structure of the search. Such an approach would need to be carefully constructed to ensure the logical structure of patterns; however, this should not be an insurmountable problem. In addition, phenotypic levels could be modified to adjust for expected environmental or other effects.

While we present our method in the context of a case-control design, adaptation to family data is possible upon selection of the appropriate nonparametric statistics. Extension to any rank-based statistic with finite range is straightforward. The MS test showed that the method is good for partially missing data; *e.g.*, alleles *CCR5* and *DRB1* that form the combination are genotyped altogether only for 212 patients and 312 controls.

While progress continues to be made in identifying key genetic variants associated with diseases, the explosion of data brought about by rapid changes in technology and the human genome project will overwhelm classical analysis methods. We have presented here a novel approach that has the potential to scale to the problems generated by high-dimensionality data. Linking this work to progress in the definition of biochemical mechanisms, to deep understanding of cellular biology, and to modeling of targets and development of therapeutics can lead to progress against complex genetic diseases.

This study was supported by the Russian Foundation for Basic Research (grants 04-04-49601 and 05-04-48982), the Russian Academy of Science Programs "Molecular and Cellular Biology," project 10; by the Howard Hughes Medical Institute (55000309); by Ludwig Institute of Cancer Research (U.S. Civilian Research and Development Foundation RBO-1268); by the National Institutes of Health [National Cancer Institute (NCI) CA06927, R01CA105090, and NCI P30CA06973 grants]; by the National Science Foundation (NSF034211); by the Pennsylvania Department of Health; and by the Pew Foundation.

LITERATURE CITED

Beaumont, M. A., and B. Rannala, 2004 The Bayesian revolution in genetics. Nat. Rev. Genet. **5:** 251–261.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B **57:** 289–300.

Boiko, A. N., E. I. Gusev, M. A. Sudomoina, A. D. Alekseenkov, O. G. Kulakova *et al.*, 2002 Association and linkage of juvenile MS with HLA-DR2(15) in Russians. Neurology **58:** 658–660.

Clyde, M. A., G. Parmigiani and B. Vidakovic, 1998 Multiple shrinkage and subset selection in wavelets. Biometrika **85:** 391–402.

Di Bucchianico, A., 1996 *Combinatorics, Computer Algebra and Wilcoxon-Mann-Whitney Test.* Memorandum COSOR 96–24, Eidhoven University of Technology, Eidhoven, The Netherlands.

Dinneen, L. C., and B. C. Blakesley, 1973 A generator for the sampling distribution of the Mann-Whitney U statistic. Appl. Stat. **22:** 269–273.

Efron, B., and R. Tibshirani, 2002 Empirical Bayes methods and false discovery rates for microarrays. Genet. Epidemiol. **23:** 70–86.

Favorova, O. O., T. V. Andreewski, A. N. Boiko, M. A. Sudomoina, A. D. Alekseenkov *et al.*, 2002 The chemokine receptor CCR5 deletion mutation is associated with MS in HLA-DR4-positive Russians. Neurology **59:** 1652–1655.

George, E. I., and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88:** 881–889.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

Gusev, E., M. Sudomoina, A. Boiko, T. Deomina and O. Favorova, 1997 TNF gene polymorphisms: associations with multiple sclerosis susceptibility and severity, pp. 35–41 in *Frontiers in Multiple Sclerosis,* edited by O. Abramsky and H. Ovadia. Dunitz Martin, London.

Hahn, L. W., M. D. Ritchie and J. H. Moore, 2003 Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics **19:** 376–382.

International HapMap Consortium, 2003 The international HapMap project. Nature **426:** 789–796.

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

Kooperberg, C., I. Ruczinski, M. L. LeBlanc and L. Hsu, 2001 Sequence analysis using logic regression. Genet. Epidemiol. **21**(Suppl. 1): S626–S631.

Liu, J. S., 2001 *Monte Carlo Strategies in Scientific Computing.* Springer, New York.

Mann, H. B., and D. R. Whitney, 1947 On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18:** 50–60.

McPherson, J. D., M. Marra, L. Hillier, R. H. Waterston, A. Chinwalla *et al.*, 2001 A physical map of the human genome. Nature **409:** 934–941.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1091.

Müller, P., G. Parmigiani, C. Robert and J. Rousseau, 2004 Optimal sample size for multiple testing: the case of gene expression microarrays. J. Am. Stat. Assoc. **99:** 990–1001.

Parmigiani, G., E. Garrett, R. Anbazhagan and E. Gabrielson, 2002 A statistical framework for expression-based molecular classification in cancer. J. R. Stat. Soc. B **64:** 717–736.

Pratt, J., and J. Gibbons, 1981 *Concepts of Nonparametric Theory*. Springer-Verlag, New York.

Priebe, C. E., and L. J. Cowen, 1999 A generalized Wilcoxon-Mann-Whitney statistic. Communications in Statistics. Part A **28:** 2871–2878.

Rannala, B., 2001 Finding genes influencing susceptibility to complex diseases in the post-genome era. Am J Pharmacogenomics **1:** 203–221.

Robert, C., and G. Casella, 1999 *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan *et al.*, 2001 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. **29:** 308–311.

Tabor, H. K., N. J. Risch and R. M. Myers, 2002 Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. Nat. Rev. Genet. **3:** 391–397.

Thornton-Wells, T. A., J. H. Moore and J. L. Haines, 2004 Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet. **20:** 640–647.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science **291:** 1304–1351.

Wilcoxon, F., 1945 Individual comparisons by ranking methods. Biometrics **1:** 80–83.

Communicating editor: P. J. Oefner