A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative Trait Variation

M.R. Nelson, S.L.R. Kardia, R.E. Ferrell, et al.

Genome Res. 2001 11: 458-470
Access the most recent version at doi:10.1101/gr.172901

A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative Trait Variation

M.R. Nelson,[1,4] S.L.R. Kardia,[2] R.E. Ferrell,[3] and C.F. Sing[1,5]

[1]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109-0618, USA; [2]Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48109-2029, USA; [3]Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA

Recent advances in genome research have accelerated the process of locating candidate genes and the variable sites within them and have simplified the task of genotype measurement. The development of statistical and computational strategies to utilize information on hundreds — soon thousands — of variable loci to investigate the relationships between genome variation and phenotypic variation has not kept pace, particularly for quantitative traits that do not follow simple Mendelian patterns of inheritance. We present here the combinatorial partitioning method (CPM) that examines multiple genes, each containing multiple variable loci, to identify partitions of multilocus genotypes that predict interindividual variation in quantitative trait levels. We illustrate this method with an application to plasma triglyceride levels collected on 188 males, ages 20–60 yr, ascertained without regard to health status, from Rochester, Minnesota. Genotype information included measurements at 18 diallelic loci in six coronary heart disease–candidate susceptibility gene regions: *APOA1–C3–A4*, *APOB*, *APOE*, *LDLR*, *LPL*, and *PON1*. To illustrate the CPM, we evaluated all possible partitions of two-locus genotypes into two to nine partitions (~10^6 evaluations). We found that many combinations of loci are involved in sets of genotypic partitions that predict triglyceride variability and that the most predictive sets show nonadditivity. These results suggest that traditional methods of building multilocus models that rely on statistically significant marginal, single-locus effects, may fail to identify combinations of loci that best predict trait variability. The CPM offers a strategy for exploring the high-dimensional genotype state space so as to predict the quantitative trait variation in the population at large that does not require the conditioning of the analysis on a prespecified genetic model.

Recent advances in genome research have accelerated the process of locating candidate genes and the variable sites within them and have simplified the task of genotype measurement. In spite of such advances, the task of identifying both the genes and the variable loci within them that influence interindividual variation in quantitative traits that are measures of health in human populations has emerged as one of the most difficult challenges facing geneticists (Risch and Merikangas 1996; Clark et al. 1998; Terwilliger and Weiss 1998). This realization may be contrasted with the many successes over the last century that have characterized the genetic basis for thousands of inborn errors of metabolism that segregate in a predictable Mendelian fashion (OMIM 2000). Early models of the genetics of quantitative traits (Nilsson-Ehle 1909; Fisher 1918) suggest why: Continuous phenotypic variation is influenced by variation among genotypes defined by many genetic loci and variation in exposures to many environmental agents.

The possible complexity of the genotype–phenotype relationship was later emphasized by Sewall Wright (1923), who argued for the importance of epistasis (the interaction between two or more genetic loci) and genotype-by-environment interaction in the mapping of genetic variability into phenotypic variability. Most biologists accept the basic tenet that the influence of variation in a particular gene depends on the context defined both by other genes and by exposures to environments, both internal and external to the individual, over the life cycle (Lewontin 1992). The importance of context in the analysis and interpretation of quantitative variation in risk factors for coronary heart disease (CHD) has been documented by our group (Reilly et al. 1991; Sing et al. 1996; Zerba et al. 1996, 2000; Lussier-Cacan et al. 1999; Nelson et al. 1999) as well as others (Cobb et al. 1992; Taimela et al. 1996; Jarvik et al. 1997) in studies of the impact of variation in the gene coding for the apolipoprotein E molecule on quantitative measures of lipid metabolism.

Despite the reality of these underlying biological

[4]Present address: Esperion Therapeutics, Ann Arbor, Michigan, 48108 USA.
[5]Corresponding author.
E-MAIL csing@umich.edu; FAX (734) 763-5277.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.172901.

11:458–470 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

complexities, a review of the literature reveals that the goal of most genetic studies of human quantitative traits is to identify and characterize the effects of individual loci that influence interindividual variability. There is a need to develop analytical methods to identify combinations of variable loci that may exhibit epistatic effects on quantitative traits. The analytical strategies that are typically employed make the implicit assumption that such interacting loci can each be identified through their independent, marginal contribution to trait variability. This simplified approach ignores the possibility that the effects of multilocus functional genetic units play a larger role than do single-locus effects in determining trait variability (Franklin and Lewontin 1970; Templeton et al. 1976; Templeton 2000).

The goal of this article is to present the combinatorial partitioning method (CPM) for identifying partitions of multilocus genotypes that predict variation in quantitative trait levels. Each set of partitions is evaluated for the phenotypic similarity of individuals within partitions and dissimilarity of the partition means. The identification and interpretation of biological interactions between alleles at each locus (dominance) as well as biological interactions between nonalleles (epistasis) is not constrained by their representation as parameters in a linear statistical genetic model. We illustrate this method with an application to plasma triglycerides and 18 variable loci from six candidate CHD susceptibility gene regions measured in 188 adult males. The results of these analyses using the CPM provide statistical evidence for biological interactions between loci. These findings demonstrate that methods relying on single-locus marginal effects to identify variable loci influencing quantitative trait variability may overlook those loci whose contribution is revealed only when considered in combination with other loci.

Box 1. Glossary of Notation

Symbol	Definition
L	Set of all measured loci
l	Number of loci in set of loci L
M	Subset of measured loci L
m	Number of loci in subset of loci M
G_M	Set of m-locus genotypes defined by the subset of loci M
g_M	Number of m-locus genotypes defined by the subset of loci M
K	Set of partitions of multilocus genotypes
k	Number of partitions in set K
n	Sample size
SS_W	Sum of squared differences among individuals within partitions of set K
MS_W	Mean squared differences among individuals within partitions of set K
SS_K	Sum of squared differences among partition means for set K
s_K^2	Bias-corrected estimate of the variance among partition means for set K
pv_K	Proportion of phenotypic variability explained by differences among partition means for set K
CV	Appended to the above statistical symbols to indicate the statistic is based on repeated 10-fold cross validation

CPM Definitions

The objective of the CPM is to identify sets of partitions of multilocus genotypes that predict quantitative trait variability. We introduce the following notation for defining the sets of genotypic partitions (see Box 1). Let L be the set of variable loci that are measured for a sample, where the number of loci in L is given as l. Let M be a subset of L loci, and the number of loci in M be m. For a particular subset M, the set of observed m-locus genotypes is denoted as G_M with size g_M. We define a genotypic partition as a partition that includes one or more of all possible genotypes from the set G_M. A set of genotypic partitions, denoted K with size k, is a collection of two or more disjoint genotypic partitions. In the CPM, every one of the possible m-locus genotypes is included in one, and only one, of the disjoint genotypic partitions that make up a set, K. The collection of all possible sets of partitions of the G_M genotypes for all subsets M of L total loci into genotypic partitions defines the state space that is evaluated by the CPM. A measure (or vector of measures) of phenotypic characteristics computed for each set represents a point on the surface of points that span all possible sets of partitions of genotypes that make up the state space.

Traditional genetic analyses of variation in trait levels begin by assuming that the number of genotypic partitions, which we denote as k, to be equal to the number of genotypes g_M on M. An analysis of variance among genotype means is followed by a posteriori comparisons to identify genotypes that have significantly different trait means. For example, consider a quantitative trait that is influenced by a diallelic variation at locus A, where the influence of the A allele is dominant to the influence of the a allele. A test of significant differences among the means of the three unique genotypes, AA, Aa, and aa, may be followed by all possible pair-wise comparisons that would establish that, for example, the mean of AA is not significantly different from the mean of Aa. The objective of the CPM is to simultaneously identify the A locus as predicting trait variability and group genotypes that are phenotypically similar into genotypic partitions as $\{AA, Aa\}$ and $\{aa\}$, emphasizing similarity among genotypes within partitions as well as differences between partitions. A posteriori comparisons have limited utility for characterizing the partitions of genotypes when g_M is large. To overcome this limitation, the CPM searches over the possible partitions of the g_M m-locus

genotypes observed on the subset of loci, M, into $2 \leq k \leq g_M$ partitions.

The application of the CPM to identify the subset of $m \leq l$ loci that divide g_M genotypes into k partitions that are similar within and most dissimilar between partitions for the mean of a quantitative trait can be broken down into three steps. These steps are diagrammed in Figure 1. The first step is to conduct the primary evaluation of the state space of sets of genotypic partitions for statistical measures of phenotypic characteristics of the k partitions of genotypes. In this presentation, we consider the estimation of the genetic variance measured by variation among the means of the k partitions of the g_M genotypes. The sets of genotypic partitions that each predict more than a prespecified level of trait variability are retained for further analysis. The second step is to validate each of the retained sets of genotypic partitions by cross validation methods. The third step is to select the best sets of genotypic partitions, on the basis of the results of the cross-validation from step 2, and proceed to draw inferences about the combinations of variable loci and the relationships between the distribution of phenotypic variability and the distribution of the g_M multilocus genotypes for these sets of genotypic partitions. In the following sections, we describe each of these steps in greater detail.

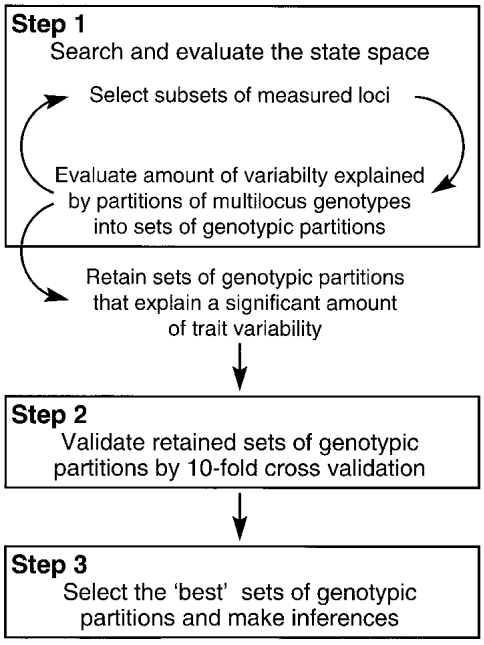

Figure 1 The three steps that constitute the combinatorial partitioning method.

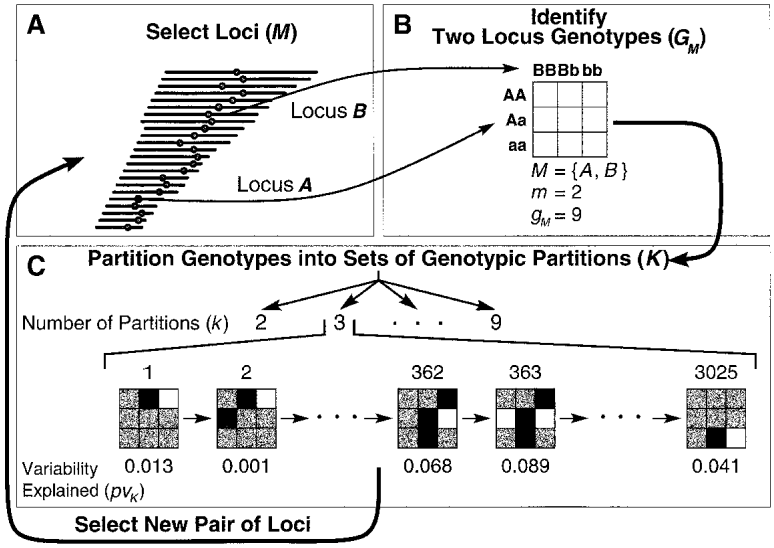

Figure 2 A depiction of the combinatorial partitioning method applied two variable loci at a time ($m = 2$) over a range of k.

Step 1: Search and Evaluation of the State Space

Searching the state space of all possible ways to partition m-locus genotypes into sets of genotypic partitions for all subsets of l loci can be separated into two nested combinatorial operations, illustrated in Figure 2 for $m = 2$. The first operation, described in Figure 1, Step 1, and illustrated in Figure 2A, consists of systematically selecting all possible subsets of variable loci for the desired values of m, of which there are $\binom{l}{m}$ ways. For each subset M, the m-locus genotypes are identified, illustrated in Figure 2B by the 3×3 grid for two diallelic loci A and B. The second operation, depicted in Figure 2C, is to evaluate the possible sets of genotypic partitions over the desired range of k. The number of ways to partition g_M genotypes into a set of k genotypic partitions is known as a Stirling number of the second kind (Comtet 1974), computed from the sum

$$S(g_M,k) = \frac{1}{k!}\sum_{i=0}^{k-1}(-1)^i\binom{k}{i}(k-i)^{g_M}.$$

For diallelic loci, the maximum number of m-locus genotypes that could be observed is $g_M = 3^m$. For two diallelic loci with $m = 2$ and $g_M = 9$, there are 21,146 ways to partition G_M into $k = 2, 3, \ldots, 9$ partitions. Combining these two combinatorial operations (choosing M from L and enumerating all K possible from G_M) in the example application below where the number of variable loci $l = 18$, if $m = 2$ and $g_M = 9$ for all M in L, results in 3,235,338 possible sets of genotypic partitions. At this size it is practical to search the two-locus state space exhaustively. However, by adding even one more variable locus such that $g_M = 27$, and we partition G_M into $k = 2, 3, \ldots, 27$, the number of sets of genotypic partitions for a single combination of three loci is ~10^{25}, clearly out of the range of what can

be exhaustively enumerated. The computational feasibility of the CPM is further addressed in the discussion below. In this article, to illustrate the CPM, we restrict our application to the two-locus case, and evaluate directly all possible sets of genotypic partitions. This complete enumeration of the state space provides additional advantages. As we have no a priori knowledge of how the state space of sets of genotypic partitions, as defined above, relates to the measured phenotypic characteristics, enumeration in low dimensions allows us to examine this relationship directly. Also, observing the nature of this relationship over the entire state space can aid in the development of algorithms for searching higher-dimensional spaces.

To evaluate the state space, we must select a statistical function that provides a measure of the value of each set of genotypic partitions. This statistical function is often referred to as the objective function. The sum of the squared deviations of the trait means of the partitions from the overall sample mean (SS_K) is a natural choice for an objective function of a set of genotypic partitions. The value of this measure increases as the individuals within genotypic partitions become more phenotypically similar and as the phenotypic differences between partitions increase. A disadvantage of the partition sum of squares is that it will tend to increase as k increases, favoring the division of genotypes into a greater number of partitions. The bias-corrected estimate of genotypic variance (Boerwinkle and Sing 1986) is used to compensate for this bias. It can be written as

$$s_K^2 = \sum_{i=1}^{k} \frac{n_i(\overline{Y}_i - \overline{Y})^2}{n} - \frac{(k-1)}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(Y_{ij} - \overline{Y}_i)^2}{n-k}$$
$$= \frac{SS_K}{n} - \frac{(k-1)}{n} MS_W$$

where n is the total sample size, $\overline{Y}$ is the sample grand mean, n_i is the sample size and $\overline{Y}_i$ is the mean of partition i, Y_{ij} is the phenotype of the jth individual in the ith partition, and MS_W is the mean squared estimate of the phenotypic variability among individuals within genotypic partitions. This statistic was derived using the expected mean squares from a standard one-way analysis of variance model to obtain an unbiased estimator of the measured genotypic variance. This correction has the effect of penalizing the scaled partition sum of squares by a quantity that increases with k if the estimate of MS_W does not decrease as additional partitions are considered. For comparative purposes, the proportion of variability explained by a set of genotypic partitions is preferred for general usage over s^2_K. The proportion, denoted pv_K, is computed as

$$pv_K = \frac{s_K^2}{s_P^2} = \frac{s_K^2}{s_K^2 + MS_W}.$$

Note that the pv_K statistic, which is a function of unbiased estimators, is not an unbiased estimator of the proportion of variability explained by set K (Wijsman and Nur 2001). However, it is appropriate for our purpose of comparing sets of genotypic partitions.

For most traits and combinations of variable loci, most of the sets evaluated will explain very little phenotypic variability and will not be useful for predictive purposes. For this reason, some filter is needed to decide which sets of genotypic partitions should be retained for further consideration in Step 2 of the CPM approach (Fig. 1). Many criteria could be used in setting this filter, including criteria based on the significance level of an F-test (e.g., $MS_K/MS_W > F[\alpha = 0.005; k-1, n-k]$), biological significance (e.g., $pv_K > 0.05$), or some proportion of all of the sets considered (e.g., the top 1%, the top 100, or simply the best). In the application described below, we use a cutoff based on the distribution of the F statistic.

In addition to the proportion of variability that a set is predicting, we include an additional criterion in the evaluation of each set of genotypic partitions. Variable loci that have alleles of low relative frequency are expected to result in some multilocus genotypes being represented by only a few individuals in any given sample. It follows that some of the sets will also have few individuals in a genotypic partition. Because our method of evaluating sets is based on both within- and between-partition sums of squares, sparse partitions with only one or a few individuals do not have sufficient degrees of freedom to reliably estimate both partition means and partition sums of squares. To avoid these cases, we set a lower bound on the number of individuals observed for a partition to be included in the CPM evaluation. The value of this lower bound is dependent on the judgement of the investigator. Setting this bound too low could lead to spurious results, as could be the case if a partition contained one or a few outliers. However, setting this bound too high could result in failing to consider sets that could be useful for predicting phenotypic variability and providing insights into the trait biology. In the application described in this article, we set a lower bound of five individuals in a valid partition.

Step 2: Validating the Selected Sets of Genotypic Partitions

The next step in the CPM is to validate the retained sets of genotypic partitions (Fig. 1, Step 2). A very large number of sets are considered in the process of searching and evaluating the state space of all sets of genotypic partitions. This problem is common to the methods usually applied in the fields of data mining, pattern recognition, machine learning, and artificial intelligence. These loosely related fields have long struggled to find appropriate ways to validate the model or collection of models that were constructed as a conse-

quence of considering the large number of possible relationships that are present within a particular data set (Ripley 1998, section 2.7). One of the most common methods of model validation is multifold cross-validation (Stone 1978), where the number of folds is typically 10 (Kohavi 1995). This method simulates the process of going back into the population of inference and collecting an independent sample with which to validate the constructed models. Briefly, 10-fold cross validation is carried out by randomly dividing the sample into 10 approximately equal-sized groups. The first group is removed from the sample, and the remaining nine are used to estimate the parameters of the model, which in the case of the CPM are the means of the k genotypic partitions of a set. Then the one group that was withheld is used to compute a portion of the predicted within partition sum of squares. This process is repeated 10 times, with each of the 10 groups being withheld from the parameter estimation and used to compute a portion of the predicted within partition sum of squares exactly once for each of the 10 groups. The 10 portions of the predicted within partition sum of squares are then summed to provide an estimate ($SS_{W,CV}$) of the cross-validated within partition sum of squares. To reduce the possibility that a particular random assignment of the sample into 10 cross validation groups might favor one set over another, the random 10-fold random division of the sample and cross validation is repeated 10 times and the resulting cross-validated within-partition sums of squares are averaged for each set of genotypic partitions.

This cross-validated estimate ($SS_{W,CV}$) of the trait variability within each genotypic partition can then be used to judge the predictive ability of a given set of genotypic partitions. For consistency and comparability, we use $SS_{W,CV}$ and $SS_{K,CV} = SS_{Total} - SS_{W,CV}$ in place of SS_W and SS_K in the equations above to calculate a cross-validated proportion of variability explained, denoted as $pv_{K,CV}$. The larger $pv_{K,CV}$ becomes, the more predictive the set K is said to be. Note that $SS_{W,CV}$ must be $\geq SS_W$, which implies that $pv_{K,CV}$ will be $\leq pv_K$.

Step 3: Select the 'Best' Sets of Genotypic Partitions and Make Inferences about the Relationship between Phenotypic and Genotypic Variation

Steps 1 and 2 provide a strategy for identifying sets of genotypic partitions that predict variation in quantitative trait levels. The third and final step in this approach (Fig. 1) is to select some subset of the validated sets of genotypic partitions on which inferences can be made. The utility of information about these sets depends on the goals of the investigator and the questions that are being asked. One use could be to identify the overall "winner," for eaxmple, the set of partitions that best predicts phenotypic variation. However, many other goals cannot be addressed by considering only one "best" set. For instance, if we are interested in enhancing our understanding of the relationship between phenotypic variability and genotypic variability at multiple loci, there may be information about the penetrance function or norm of reaction to be gained by scrutinizing a subgroup of multiple sets of genotypic partitions that may be almost as predictive as the overall best predictive set. When making inferences about such a group of multiple sets, we may ask the following questions: How much trait variability does each of the selected sets explain? Which combinations of loci are involved in the selected subgroup of all possible sets? If there are multiple combinations of loci represented in these selected sets, how do the genotype–phenotype relationships in one combination compare with another? Are the selected loci acting additively on the variance of the trait? How do the genotype–phenotype relationships observed in the selected sets compare with the relationships expected using traditional multilocus models? What criteria should be used in making the subjective decision of how many sets should be selected for drawing inferences about genotype–phenotype relationships? An application of the CPM to illustrate it's utility in providing analytical results to deal with these questions is presented next.

Example Application

Sample

To illustrate the combinatorial partitioning method, we present an application to identify sets of genotypic partitions of two-locus combinations of 18 diallelic loci (16 single nucleotide polymorphism [SNPs] and two insertion-deletions [InDels]) located in six candidate CHD susceptibility gene regions (Table 1) that predict interindividual variability in plasma triglycerides (Trig) levels, a known CHD risk factor. The sample used here comes from the Rochester Family Heart Study (RFHS), a population-based study of multigeneration pedigrees sampled without regard to health status from the Rochester, Minnesota, population (Turner et al. 1989). Because the genetic influence on variation in most quantitative trait levels is influenced by age (Reilly et al. 1992; Zerba et al. 1996; Jarvik et al. 1997; Nelson et al. 1999), we restricted our analyses to adults ages 20–60 yr. From 281 pedigrees, we selected 188 males for this sample application. All individuals were within the prescribed age window and had Trig measurements and complete genotypes for all 18 variable loci. In this

Table 1. Variable Diallelic Loci Used in the Example Application

Locus	Relative frequency[a]	*pv*	Location	Reference
APOB			2p24	
InDel	0.683	0.000	Exon 1	Boerwinkle and Chan 1989
*Xba*I RFLP	0.525	0.002	Exon 26	Berg et al. 1986
*Msp*I RFLP	0.899	0.009	Exon 26	Priestley et al. 1985
*Eco*RI RFLP	0.871	0.000	Exon 29	Shoulders et al. 1985
PONI			7q22	
L54M	0.629	0.000	Exon 3	Humbert et al. 1993
R192N	0.718	0.010	Exon 6	Humbert et al. 1993
LPL			8p22	
*Pvu*II RFLP	0.526	0.000	Intron 6	Li et al. 1988
*Hin*dIII RFLP	0.721	0.012	Intron 8	Heizmann et al. 1987
*Bst*NI RFLP	0.703	0.000	Intron 9	Funke et al. 1988
APOAI-C3-A4			11q23	
InDel	0.548	0.010	5′ of APOAI	Coleman et al. 1986
*Pst*I RFLP	0.943	0.010	3′ of APOAI	Kessling et al. 1985
*Sst*I RFLP	0.907	0.000	Exon 4, APOC3	Rees et al. 1983
LDLR			19p13	
*Taq*I RFLP	0.911	0.017	Intron 4	Yamakawa et al. 1987
*Stu*I RFLP	0.955	0.021	Exon 8	Kotze et al. 1986
*Hin*cII RFLP	0.556	0.000	Exon 12	Leitersdorf and Hobbs 1987
*Ava*II RFLP	0.552	0.000	Exon 13	Hobbs et al. 1987
APOE			19p13.2	
C112R	0.833	0.000	Exon 4	Rall et al. 1982
R158C	0.922	0.000	Exon 4	Rall et al. 1982

[a]For the most common allele.
Loci are shown with the proportion of lnTrig variability (*pv*) that they explain independently.

sample, the distribution of Trig was right skewed (skewness = 2.7) and leptokurtic (kurtosis = 9.6). In accordance with common practice in the study of lipid metabolism, we applied a natural logarithm transformation (lnTrig). This transformation greatly reduced both the skewness (0.83) and the kurtosis (1.04). Finally, to reduce the influence of age and body size on lnTrig variability, we adjusted lnTrig using a linear regression model that included age, height, and weight (each up to third-order polynomial terms) and waist-to-hip ratio and body mass index.

Analysis

The single-locus contribution to adjusted lnTrig (henceforth referred to as lnTrig) variability (*pv*) is shown in Table 1 for each of the 18 diallelic loci. The bias-corrected estimate of the proportion of lnTrig variability explained ranges from 0.000 for nine of the loci to 0.021 for the *Stu*I locus in the *LDLR* gene. The variability among single-locus genotypes was not statistically significant at the level of $\alpha = 0.01$ for any of the 18 variable loci.

Step 1: Evaluation of the State Space of Sets of Genotypic Partitions

As discussed above, we limited the state space in this application to all possible sets of genotypic partitions (K: $k = 2, 3, \ldots, 9$) defined by all possible pairs ($m = 2$) of the $l = 18$ diallelic loci. We set five as the lower bound for the number of individuals that must be present to constitute a valid partition. Sets containing a partition with fewer than five individuals were excluded from further evaluation. To filter the potentially millions of valid sets of genotypic partitions for validation in Step 2, we selected a criterion based on the test-wise significance of each set. All sets with an F statistic that exceeded the 0.995 quantile of the F distribution corresponding to each k considered, that is, $F(\alpha = 0.005; k - 1, n - k)$, were retained for further validation. This combination of criteria for $k = 2, 3, \ldots, 9$ resulted in the consideration of 794,699 sets and the retention of 7710 sets (1.0%).

The number of sets considered is much lower than the 3,235,338 possible sets associated with all two-locus combinations of 18 diallelic loci. There are two reasons for this. The first is that for most of the pairwise combinations of loci, fewer than nine two-locus genotypes were observed, such that the total number of enumerable sets of genotypic partitions for this sample is 1,001,270. The second reason is that 206,571 sets were not evaluated because they contained partitions with fewer than five individuals. The proportion of variability explained by each of the retained sets is displayed in Figure 3. Each line in the plot corresponds to the retained sets for a different number of partitions ($k = 2, 3, \ldots, 8$). The retained partitions for each k are sorted by the proportion of variability explained for each set, with the sets that explain the least at the left

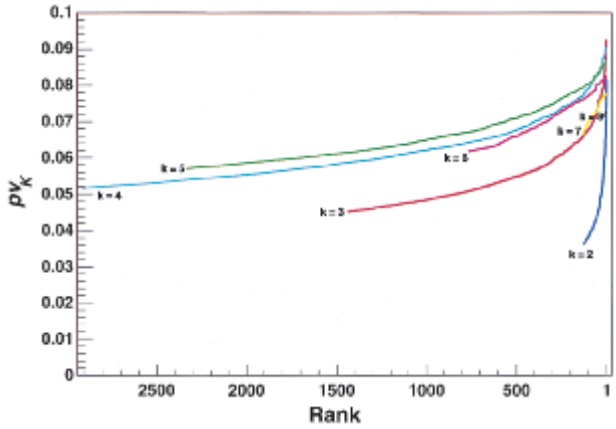

Figure 3 Plot of the proportion of variability explained by the 7710 retained sets of genotypic partitions. The sets are sorted by the proportion of variability explained and connected by a colored line corresponding to the number of partitions in each set.

of the plot and the sets that explain the most at the right.

There are a few notable features of the plot in Figure 3. First, the number of sets retained varies greatly by k, ranging from 11 ($k = 8$) to 2911 ($k = 4$). Second, the sets that explain the greatest proportion of variability for each k range from $pv_K = 0.073$ ($k = 8$) to $pv_K = 0.093$ for ($k = 3$). Recall that the retained sets that explain the least proportion of variability in lnTrig correspond to the minimum established by the F distribution with the appropriate degrees of freedom for each k. The third notable feature in Figure 3 is the shape of the lines. There are many sets of genotypic partitions near the cutoffs that explain roughly the same proportion of variability, and relatively few that explain substantially more. For example, for $k = 3$, 1357 (94%) of the retained sets explain between 0.045 and 0.069 (lower 50% of the range in pv_K), while 81 (6%) explained between 0.070 and 0.093 (upper 50% of the range in pv_K), and only 12 (0.8%) explained within the upper 25% of the range.

While Figure 3 provides a useful description of the proportion of variability being explained by the retained sets of genotypic partitions, it contains no information about which pairs of loci are involved. This information is presented in Figure 4 for $k = 2, 3, 4, 5$. The remaining k are not presented because of limitations in space and, as it will be shown by cross validation in the next section, all of the sets for $k = 6, 7, 8$ lack the ability to predict lnTrig variability. In Figure 4, there is one panel corresponding to each k. The Y-axis

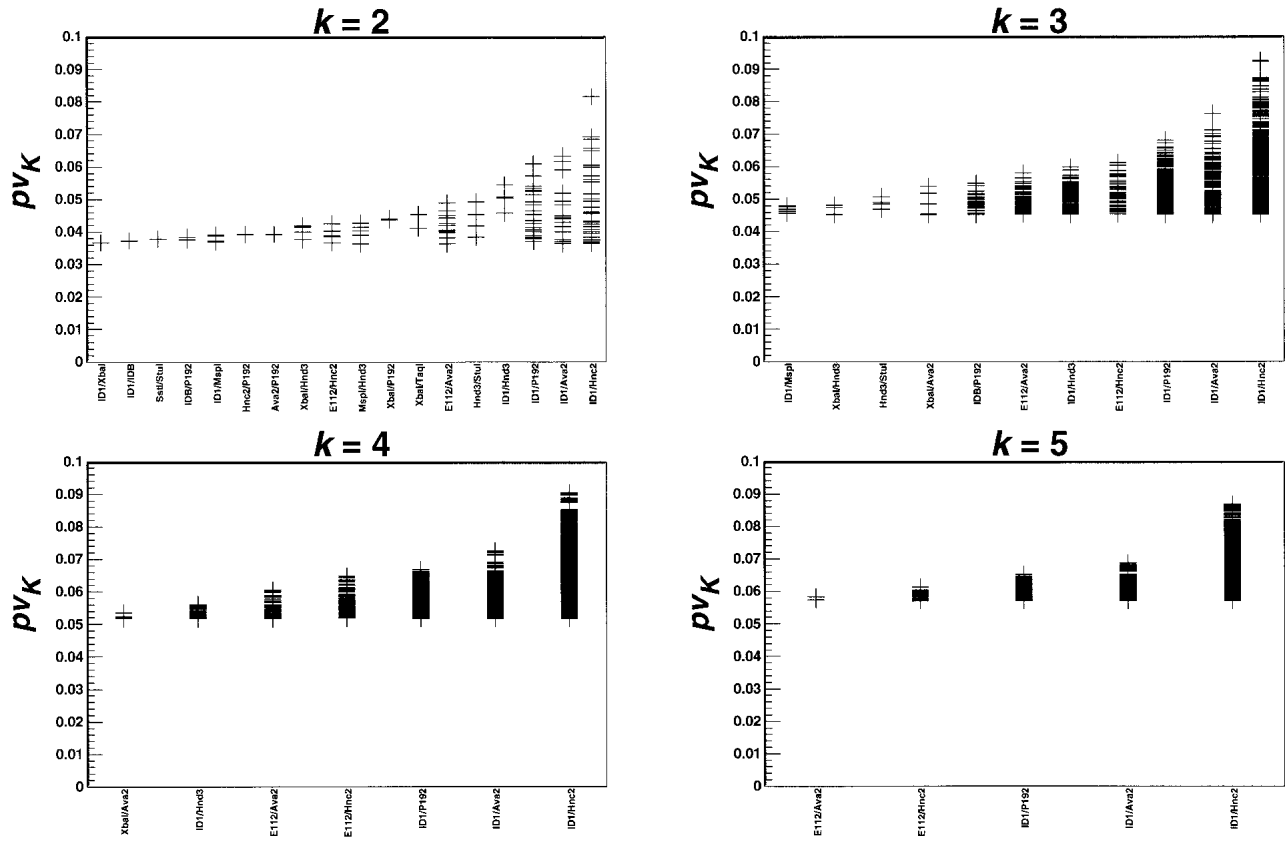

Figure 4 Plot of the proportion of variability explained by the same sets shown in Figure 3, after grouping by the pairs of variable loci included in each set and sorting groups by the proportion of variability explained by the best partition for each group.

corresponds to the proportion of variability as in Figure 3. The *X*-axis corresponds to each pair of loci with a retained set of genotypic partitions, sorted on the basis of the maximum value of pv_K for each pair. The greatest proportion of variability is explained by a set with $k = 3$ defined by the pair {InDel (*APOA1-C3-A4*), *Hin*cII (*LDLR*)}. The set of genotypic partitions that explains the greatest proportion of ln Trig variability not involving the InDel and *Hin*cII pair of loci also corresponds to $k = 3$ and is defined by {InDel (*APOA1-C3-A4*), *Ava*II (*LDLR*)} with $pv_K = 0.076$. Note that *Hin*cII and *Ava*II are both located within *LDLR*, are separated by a single intron, and have previously been shown to be in very strong linkage disequilibrium in this population (Zerba et al. 1998). As a result, the most explanatory sets defined by these variable loci have very similar assignments of genotypes (and individuals) to those of partitions. The next most explanatory set also corresponds to $k = 3$ defined by the pair {InDel (*APOA1-C3-A4*), R192N (*PON*)}, where $pv_K = 0.068$. In addition to the most explanatory set overall, the pair {InDel (*APOA1-C3-A4*), *Hin*cII (*LDLR*)} identified the most explanatory set of genotypic partitions for all *k*. The most explanatory set in $k = 4$ explains nearly as much as the best set overall ($pv_K = 0.090$ versus $pv_K = 0.093$). The most explanatory sets with two and five partitions had pv_K of 0.082 and 0.087, respectively.

Step 2: Validate Retained Sets of Genotypic Partitions

The 7710 sets of genotypic partitions retained in Step 1 were validated using the repeated 10-fold cross-validation method discussed above. The cross-validation was carried out on each of the retained sets 10 times. Using the predicted within-partition sum of squares ($SS_{W,CV}$) obtained from averaging the 10 cross-validation results, $pv_{K,CV}$, a measure of the predictive ability of each retained set was obtained. The sets were then sorted by $pv_{K,CV}$ and plotted in Figure 5, which, analogous to Figure 3, plots a nondecreasing line connecting the ranked sets for each number of partitions with $pv_{K,CV}$ on the Y-axis. The values of $pv_{K,CV}$ range

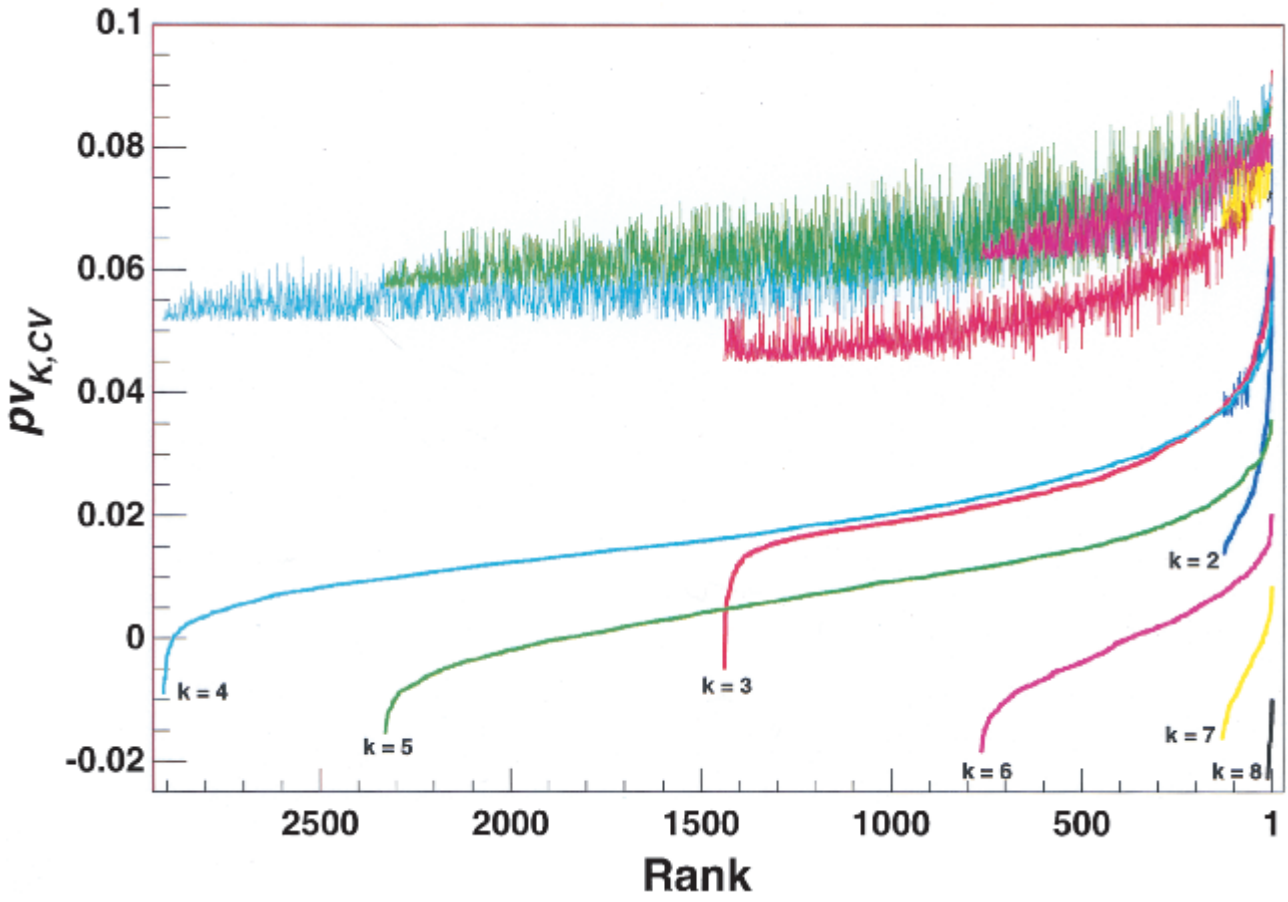

Figure 5 Plot of the cross-validated proportion of variability explained by the 7710 retained sets of genotypic partitions. The sets are sorted by the proportion of variability explained and connected by a colored line corresponding to the number of partitions in each set (smooth, nondecreasing lines). The proportion of variability explained for each set before cross-validation is shown by the jagged lines of corresponding colors.

from −0.023 (k = 8) to 0.067 (k = 3). All of the sets for k = 6, 7, 8 fall below 0.020, and on the basis of our inspection of Figure 5, we chose to consider these sets as not predictive. The greatest difference between the lines displaying $pv_{K,CV}$ in Figure 5 and those in Figure 3 is that instead of slowly sloping toward the lower bound, there is a small proportion for each k that is noticeably less predictive than the rest, shown by the sharp drop-off at the left side of each line. As in Figure 3, a majority of the sets fall within a very narrow range of the line. This is particularly pronounced for k = 3, 4, 5. The right-hand portion of the plot showing the best sets for each k again shows that they are outliers from the rest. This is a useful feature for the selection of a subset of sets of genotypic partitions on which to make inferences, focusing attention on just a few sets from among the many possibilities.

To display the relationship between pv_K and $pv_{K,CV}$, the values of pv_K for each retained set are also included in Figure 5 (jagged lines), where the color of the line is used to differentiate among k. Though the two measures are correlated, the relationship between them is not one to one. There are many sets with high values of pv_K that do not have correspondingly high values of $pv_{K,CV}$. In this particular case, the set with the highest overall value of pv_K also had the highest overall value of $pv_{K,CV}$.

The predictive ability of each of the retained sets of genotypic partitions sorted according to combinations of variable loci, analogous to Figure 4, is shown in Figure 6 for k = 2, 3, 4, 5. In comparison to Figure 4, the ranking of the pairs of variable loci is unchanged among the top three combinations. As noted previously, for each k, the top three combinations of loci are {InDel (*APOA1-C3-A4*), *Hin*cII (*LDLR*)}, {InDel (*APOA1-C3-A4*), *Ava*II (*LDLR*)}, and {InDel (*APOA1-C3-A4*), R192N (*PON*)}.

Step 3: Select the Best Sets of Genotypic Partitions and Make Inferences about the Relationship between Phenotypic and Genotypic Variation

Here we are faced with the challenge of restricting our attention to some subset of the 7710 retained sets with which we can make inferences. Our focus for this task is on the results plotted in Figures 5 and 6, based on the cross-validated proportion of variability explained (predictive ability). Clearly, sets that take on negative values for $pv_{K,CV}$, which occurs when $SS_{W,CV}$ is greater than SS_{Total}, have no predictive ability and would not be useful for making inferences. Rather than select

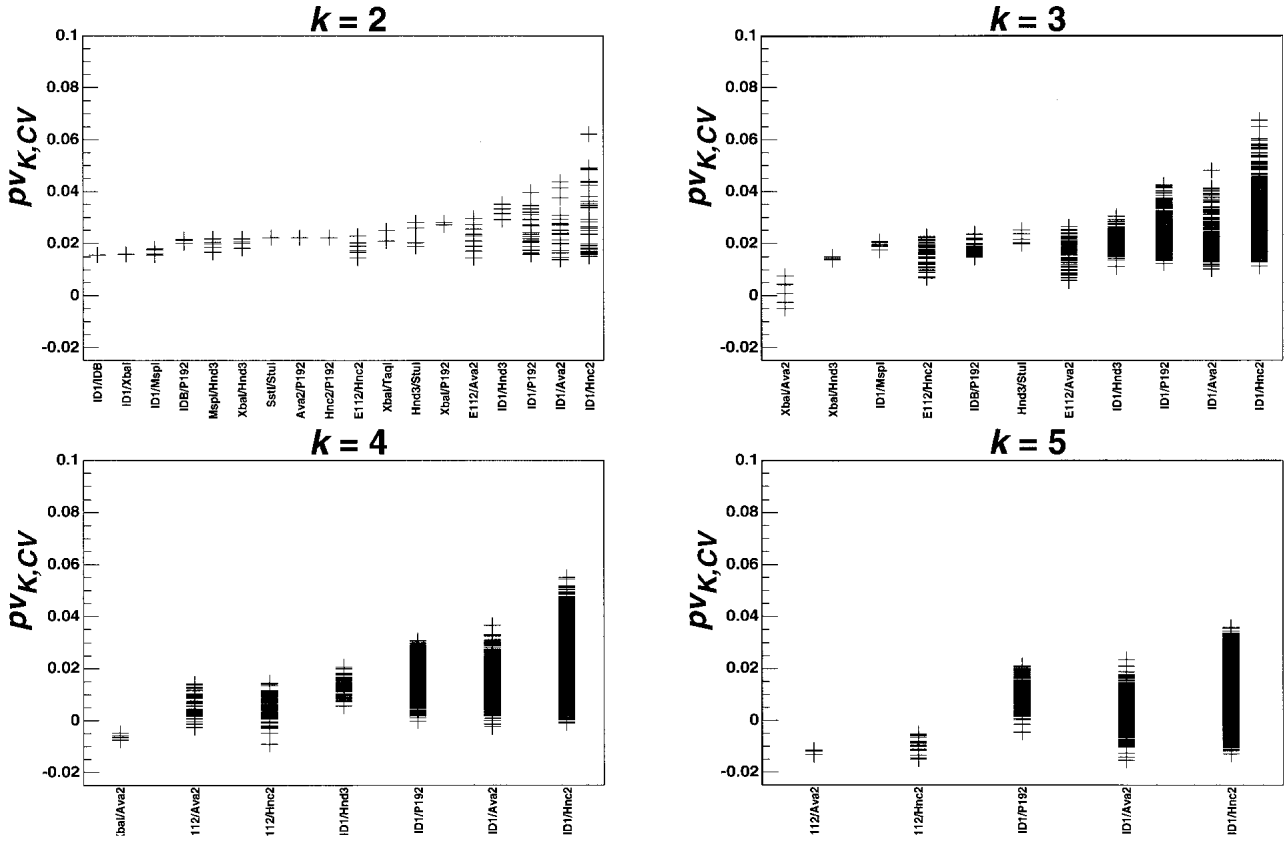

Figure 6 Plot of the proportion of cross-validated variability explained by the same sets shown in Figure 5, after grouping by the pairs of variable loci included in each set and sorting groups by the proportion of variability explained by the best partition for each group.

some predetermined level of $pv_{K,CV}$ as being adequately predictive for inferential purposes, we instead can benefit from the information contained in the distribution of the cross-validated results. Combining the information contained in Figures 5 and 6, we see that there are three sets of genotypic partitions that stand out above the remaining 7707 sets. The values of $pv_{K,CV}$ for these three best sets are 0.062 (k = 2), 0.065 (k = 3), and 0.067 (k = 3), with corresponding values of pv_K of 0.081, 0.092, and 0.093, respectively. Each of these sets are partitions of nine two-locus genotypes defined by {InDel (*APOA1-C3-A4*), *Hinc*II (*LDLR*)}. In fact, there are 57 sets defined by these two loci that predict lnTrig variability better than the best set defined by the next best pair of loci, {InDel (*ApoA1-C3-A4*), *Ava*II (*LDLR*)}.

The partitioning of the multilocus genotypes into the three best sets is shown in Figure 7. The genotypes are represented by a 3 × 3 grid, and the assignment of each genotype to a partition is indicated by shade, with the lightest shade for the partition with the lowest mean lnTrig and the darkest shade for the partition with the greatest mean lnTrig. In all three sets, the partition corresponding to the lightest shade (lowest mean) is unchanged. Also unchanged is the presence of (*DD*,− −) and (*ID*,+−) in the same partition as well as(*ID*,++) and (*II*,+−). The only difference between the two sets with k = 3 is the change of(*II*,− −) from the darkest partition to the intermediate partition. This change has only a minor effect on both pv_K and $pv_{K,CV}$. It is not surprising that this two-locus genotype is the least frequent of the nine, containing seven individuals. For the set with k = 2, the multilocus genotypes within the intermediate and high mean ln Trig in the k = 3 sets are combined into a single partition, which results in a moderate decrease in both pv_K and $pv_{K,CV}$.

To obtain an estimate of the statistical significance of the most predictive set, we performed a permutation test (Edgington 1995). The objective of a permutation test is to produce the distribution of the test statistic of interest under the null hypothesis, which in this case is proportion of variability explained by the most predictive set of genotypic partitions identified by the CPM. The test was carried out by disassociating the genotypes and phenotypes in the data set, randomly reassigning the phenotypes to genotypes via sampling without replacement, and then performing the CPM on this randomized data. This process was repeated 1000 times, and for each permutation, the set that explained the greatest proportion of variability was obtained. The proportion of permutations with results greater than the observed estimate of 0.093 for the most predictive set was 0.14.

DISCUSSION

The CPM is being developed to simultaneously identify variable loci and model the statistical relationship between interindividual variability in quantitative trait levels and the selected loci. The CPM addresses three limitations of traditional statistical genetic methods. It considers, first, a model free strategy for evaluating combinations of a large number of variable loci; second, the detection of nonadditivity of locus effects, even in the absence of marginal effects of the loci being considered; and third, relationships between phenotypic variability and genotypic variability that are not constrained by a priori genetic models.

The example application presented in this article illustrates the fundamental steps involved in the CPM (Fig. 1) and demonstrates several of its advantages. First, the CPM identified combinations of loci with small marginal effects that combined to predict the greatest amount of trait variation. The best partition of the genotypes defined by the InDel and *Hinc*II loci predicted 9% of trait variability (Fig. 7), but each locus predicted <1% variability when considered separately (see Table 1). Second, the CPM identified multiple predictive sets of genotypic partitions. The best three sets of genotypic partitions, each predicting >8% of trait variability, were defined by the InDel and *Hinc*II loci. Third, comparison of the best sets of genotypic partitions enhances our understanding of the relationship between phenotypic vari-

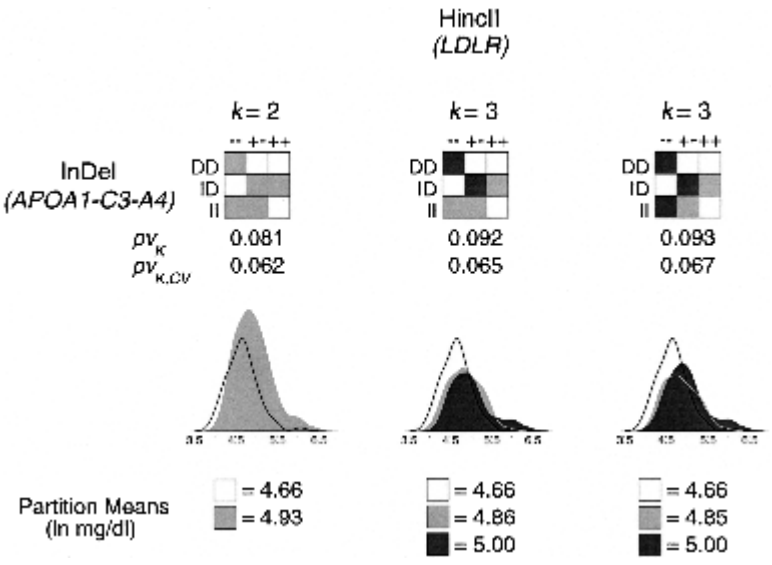

Figure 7 The three selected sets of genotypic partitions with the greatest proportion of cross-validated variability explained are represented by a 3 × 3 grid of nine two-locus genotypes with shading to represent the partition each genotype belongs to. Below each partition is a smoothed histogram showing the ln Trig distribution within each partition (indicated by shading) and the mean and sample size of each partition.

ability and genotypic variability at multiple loci. The best three sets of partitions identified by the CPM in the example application suggest that the dominance effects of the *Hin*cII alleles are dependent on the context defined by genotypes at the InDel locus.

None of the loci considered predicted >2% of trait variability, yet many two-locus combinations predicted >5% of trait variation. The implication of the kind of nonadditivity observed here is that methods for identifying variable loci that influence quantitative trait variability that rely on independent marginal effects will underestimate the contribution of genetic variation to determining the genetic architecture of a quantitative trait. If such gene-by-gene interaction proves to be a common theme, as we expect (Templeton 2000), the common practice of searching for influential loci one at a time will overlook the contribution of many important variable loci.

The example application also brings to attention issues that must be addressed to take full advantage of the CPM. An investigator is faced with the subjective decision of establishing the criteria for the size of the selected subgroup of sets of partitions that will be considered in making inferences. This decision becomes more complicated when the best sets of genotypic partitions are defined by multiple independent pairs of variable loci. This was not an issue in the example application presented here but has appeared in other data sets to which we are applying this method. Partition comparison methods such as the Rand index (Hubert and Arabie 1985) that compare the distributions of individuals among partitions between sets may prove useful in resolving this issue. Further work on methods of validating the selected sets of genotypic partitions is needed to guard against artifacts that may influence the 10-fold cross-validation strategy. Work is ongoing in our group to apply alternative methods of model validation, particularly randomization tests (Edgington 1995; Efron and Tibshirani 1997) and comparing the inferences drawn using the different methods such as CART (Breiman et al. 1984) and neural networks (Ripley 1998).

The full potential of the CPM to investigate epistasis and genotype-by-environment interaction will not be achieved if only pairs of variable loci (or other discrete variables) are considered. It is therefore of great interest to extend the CPM to larger numbers of loci (l), combinations of loci (m), and sample sizes (n). At what point is this method no longer computable? As discussed above, the two operations in Figure 1, Step 1 involve two levels of combinatorics that we need to be concerned about. The first level involves choosing subsets M of set L. For each combination of loci M, various sums of squares and sample statistics must be updated, which require $O(n)$ time to accomplish (for a review of big-O notation, see Knuth 1997, pp. 107–111). The second level of combinatorics involves searching over the state space of sets of genotypic partitions for each choice of M for those sets we wish to retain for further validation. With the initial sums of squares and other statistics computed for a particular M, calculating the necessary statistics to judge the proportion of variability explained by a particular set K using efficient updating algorithms is accomplished in $O(g_M)$ time, which is short for the relatively small values of m that would be of interest. However, the number of possible K for each combination of variable loci becomes very large very quickly. Heuristic optimization methods such as simulated annealing (Kirkpatrick et al. 1983) and genetic algorithms (Goldberg 1989) might prove effective for finding the sets of genotypic partitions that are best for predicting quantitative trait variability.

Most traits of interest in human genetics have a complex multifactorial etiology. Variation in such traits is influenced by variation in many factors, including multiple interacting genetic loci and environmental exposures. However, most studies of the genetic basis of such traits ignore this reality, relying on methods that have been effective in the arena of simpler Mendelian traits that consider the relationship between trait variability and genotypic variability on a locus-by-locus basis, assuming context invariance. This general tendency toward oversimplification was clearly stated by Franklin and Lewontin (1970):

The models of population genetics, which have remained almost unchanged for forty years, are most commonly criticized for ignoring the "natural" unit of selection, the genotype, in favor of the gene. This criticism is really an attack on one of the basic assumptions of population genetics theory, namely that the genotypic array in a random mating population, and evolutionary changes in that array, can be described in terms of gene frequencies at the individual loci.

This criticism continues to be relevant to the conceptual models of genotype–phenotype relationships that most current analytical methods are based on. Efforts to advance the available methods for studying the relationships between phenotypic variability and genotypic variability must consider the entire genotypic array, however complex, and identify the combinations of variable loci that define multilocus genotypes that predict quantitative trait variability. The CPM is one step in the direction toward the development of methods that embrace a more realistic perspective on the role that genotypic variability plays in interindividual variation in quantitative traits that have a complex multifactorial etiology.

ACKNOWLEDGMENTS

We thank Steven T. Turner at the Mayo Clinic for his supervision of the collection of the pedigrees in the Rochester

Family Heart Study. We thank Ken Lange at UCLA, Andy Clark, and Ken M. Weiss at Pennsylvania State University, and Jim Cheverud at Washington University for constructive conversations on the presented method. We also thank Tracy Fuller, Paul Kopec, Ken G. Weiss, Richard Merkle, Lynn Illeck, Debbie Theodore, and Sara Hamon at the University of Michigan for their excellent support and technical assistance. Funding for this research was provided by the National Institutes of Health grants HL39107 and HL58240.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Berg, K., Powell, L.M., Wallis, S.C., Knott, T.J., and Scott, J. 1986. Genetic linkage between the antigenic group (Ag) and the apolipoprotein B gene: Assignment of the Ag locus. *Proc. Natl. Acad. Sci.* **83:** 7367–7370.

Boerwinkle, E. and Chan, L. 1989. A three codon insertion/deletion polymorphism in the signal peptide region of the human apolipoprotein B (ApoB) gene directly typed by the polymerase chain reaction. *Nucleic Acids Res.* **17:** 4003.

Boerwinkle, E. and Sing, C.F. 1986. Bias of the contribution of single-locus effects to the variance of a quantitative trait. *Am. J. Hum. Genet.* **39:** 137–144.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and regression trees*. Wadsworth & Brooks/Cole, Monterey, CA.

Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Buchanan, A., Stengård, J., Salomaa, V., Vartianainen, E., Perola, M., Boerwinkle, E., et al. 1998. Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63:** 595–612.

Cobb, M., Teitlebaum, H., Risch, N., Jekel, J., and Ostfeld, A. 1992. Influence of dietary fat, apolipoprotein E phenotype, and sex on plasma lipoprotein levels. *Circulation* **86:** 849–857.

Coleman, R.T., Gonzalez, P.A., Funke, H., Assmann, G., Levy-Wilson, B., and Frossard, P.M. 1986. Polymorphisms in the apolipoprotein AI-CIII gene complex. *Mol. Biol. Med.* **3:** 213–228.

Comtet, L. 1974. *Advanced combinatorics: The art of infinite expansions*. Reidel, Boston.

Edgington, E.S. 1995. *Randomization tests*. 3rd ed. Dekker, New York.

Efron, B. and Tibshirani, R. 1997. Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Stat. Assn.* **92:** 548–560.

Fisher, R.A. 1918. The correlations between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.* **52:** 399–433.

Franklin, I. and Lewontin, R.C. 1970. Is the gene the unit of selection? *Genetics* **65:** 707–734.

Funke, H., Reckwerth, A., Stapenhorst, D., Schulze, M., Beiering, D., Jansen, M., and Assmann, G. 1988. A *Bst*NI (*Eco*RII) RFLP in the lipoprotein lipase gene (LPL). *Nucleic Acids Res.* **16:** 2741.

Goldberg, D.E. 1989. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, Reading, MA.

Heizmann, C., Ladias, J., Antonarakis, S., Kirchgessner, T., Schotz, M., and Lusis, A.J. 1987. RFLP for the human lipoprotein lipase (LPL) gene: *Hin*dIII. *Nucleic Acids Res.* **15:** 6763.

Hobbs, H.H., Esser, V., and Russell, D.W. 1987. *Ava*II polymorphism in the human LDL receptor gene. *Nucleic Acids Res.* **15:** 379.

Hubert, L. and Arabie, P. 1985. Comparing partitions. *J. Classification* **2:** 193–218.

Humbert, R., Adler, D.A., Disteche, C.M., Hassett, C., Omiecinski, C.J., and Furlong, C.E. 1993. The molecular basis of the human serum paraoxonase activity polymorphism. *Nat. Genet.* **3:** 73–76.

Jarvik, G.P., Goode, E.L., Austin, M.A., Auwerx, J., Deeb, S., Schellenberg, G.D., and Reed, T. 1997. Evidence that the apolipoprotein E-genotype effects on lipid levels can change with age in males: A longitudinal analysis. *Am. J. Hum. Genet.* **61:** 171–181.

Kessling, A.M., Horsthemke, B., and Humphries, S.E. 1985. A study of polymorphisms around the human apolipoprotein AI gene in hyperlipidaemic and normal individuals. *Clin. Genet.* **28:** 296–306.

Kirkpatrick, S., Gelatt Jr., C.D., and Vecchi, M.P. 1983. Optimization by simulated annealing. *Science* **220:** 671–680.

Knuth, D.E. 1997. *The art of computer programming*. Vol. 1. *Fundamental algorithms*. 3rd ed. Addison-Wesley, Reading, MA.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence*, Montreal, PQ. pp. 1137–1143.

Kotze, M.J., Retief, A.E., Brink, P.A., and Weich, H.F. 1986. A DNA polymorphism in the human low-density lipoprotein receptor gene. *S. Afr. Med. J.* **70:** 77–79.

Leitersdorf, E. and Hobbs, H.H. 1987. Human LDL receptor gene: Two *Apa*LI RFLPs. *Nucleic Acids Res.* **15:** 2782.

Lewontin, R.C. 1992. Genotype and phenotype. In *Keywords in evolutionary biology* (eds. E. Keller, E.A. Lloyd), pp. 137–144. Harvard University Press, Cambridge, MA.

Li, S., Oka, K., Galton, D., and Stocks, J. 1988. *Pvu*II RFLP at the human lipoprotein lipase(LPL) gene. *Nucleic Acids Res.* **16:** 2358.

Lussier-Cacan, S., Xhignesse, M., Kessling, A.M., Davignon, J., and Sing, C.F. 1999. Sources of variation in plasma lipid and lipoprotein traits in a sample selected for health. *Am. J. Epidemiol.* **150:** 1229–1237

Nelson M.R., Kardia, S.L.R., Ferrell, R.E., and Sing, C.F. 1999. Influence of apolipoprotein E genotype variation on the means, variances, and correlations of plasma lipids and apolipoproteins in children. *Ann. Hum. Genet.* **63:** 311–328.

Nilsson-Ehle, H. 1909. *Kreuzunguntersuchungen an Hafer und Weisen*. Lund, Lunds Univ. Aerskr., NF.

OMIM (Online Mendelian Inheritance in Man). 2000. Center for Medical Genetics, Johns Hopkins Univerisity and National Center for Biotechnology Information, National Library of Medicine, http://www.ncbi.nlm.nih.gov/omim.

Priestley, L., Knott, T., Wallis, S., Powell, L., Pease, R., and Scott, J. 1985. RFLP for the human apolipoprotein B gene. IV. *Msp*I. *Nucleic Acids Res.* **13:** 6792.

Rall, S.C., Weisgraber, K.H., and Mahley, R.W. 1982. Human apolipoprotein E: The complete amino acid sequence. *J. Biol. Chem.* **257:** 4171–4178.

Rees, A., Shoulders, C.C., Stocks, J., Galton, D.J., and Baralle, F.E. 1983. DNA polymerphism adjacent to human apoprotein A-1 gene: Relation to hypertriglyceridaemia. *Lancet* **8322:** 444–446.

Reilly, S.L., Ferrell, R.E., Kottke, B.A., Kamboh, M.I., and Sing, C.F. 1991. The gender-specific apolipoprotein E genotype influence on the distribution of lipids and apolipoproteins in the population of Rochester, MN. I. Pleiotropic effects on means and variances. *Am. J. Hum. Genet.* **49:** 1155–1166.

Reilly, S.L., Ferrell, R.E., Kottke, B.A., and Sing, C.F. 1992 The gender-specific apolipoprotein E genotype influence on the distribution of lipids and apolipoproteins in the population of Rochester, MN. II. Regression relationships with concomitants. *Am.J. Hum. Genet.* 51:1311–1324.

Ripley, B.D. 1998. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge.

Risch, N. and Merikangas, K. 1996. The future of genetic studies of complex human diseases. *Science* **273:** 1516–1517.

Shoulders, C.C., Myant, N.B., Sidoli, A., Rodriguez, J.C., Cortese, C., Baralle, F.E., and Cortese, R. 1985. Molecular cloning of human LDL apolipoprotein B cDNA. *Atherosclerosis* **58:** 277–289.

Sing, C.F., Haviland, M.B., and Reilly, S.L. 1996. Genetic architecture of common multifactorial diseases. In *Variation in the Human Genome. Ciba Foundation Symposia* (eds. D. Chadwick, G. Cardew), pp. 211–232. Wiley, Chichester.

Stone, M. 1978. Cross-validation: A review. *Math. Operationsforsch. Statist. Ser. Statistics* **9:** 127–129.

Taimela S., Lehtimäki, T., Porkka, K.B.K., Räsänen, L., and Viikari, J.S.A. 1996. The effect of physical activity on serum total and

low-density lipoprotein cholesterol concentrations varies with apolipoprotein E phenotype in male children and young adults: The Cardiovascular Risk in Young Finns Study. *Metabolism* **45:** 797–803.

Templeton, A.R. 2000. Epistasis and complex traits. In *Epistasis and the evolutionary process* (eds. M. Wade, B. Brodie III, and J. Wolf). Oxford University Press, Oxford.

Templeton, A.R., Sing, C.F., and Brokaw, B. 1976. The unit of selection in *Drosophila mercatorum*. I. The interaction of selection and meiosis in parthenogenetic strains. *Genetics* **82:** 349–376.

Terwilliger, J.D. and Weiss, K.M. 1998. Linkage disequilibrium mapping of complex disease: Fantasy or reality? *Curr. Opin. Biotech.* **9:** 578–594.

Turner, S.T., Weidman, W.H., Michels, V.V., Reed, T.J., Ormson, C.L., Fuller, T., and Sing, C.F. 1989. Distribution of sodium-lithium counter-transport and blood pressure in Caucasians five to eighty-nine years of age. *Hypertension* **13:** 378–391.

Wijsman, E.M. and Nur, N. 2001. On estimating the proportion of variance attributable to a measured locus. *Hum. Hered.* **51:**

Wright, S. 1923. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc. 6th Int. Congress Genet.* **1:** 356–366.

Yamakawa, K., Okafuji, T., Iwamura, Y., Russell, D.W., and Hamaguchi, H. 1987. *Taq*Ipolymorphism in the human LDL receptor gene. *Nucleic Acids Res.* **15:** 7659.

Zerba, K.E., Ferrell, R.E., and Sing, C.F. 1996. Genotype-environment interaction: Apolipoprotein E (*ApoE*) gene effects and age as an index of time and spatial context in the human. *Genetics* **143:** 463–478.

Zerba, K.E., Ferrell, R.E., and Sing, C.F. 1998. Genetic structure of five susceptibility gene regions for coronary artery disease: Disequilibria within and among regions. *Hum. Genet.* **103:** 346–354.

———. 2000. Complex adaptive systems and human health: The influence of common genotypes of the apolipoprotein E (*ApoE*) gene polymorphism and age on the relational order within a field of lipid metabolism traits. *Hum. Genet.* **107:**466–475.

Received November 27, 2000; accepted in revised form January 2, 2001.