

The Restricted Partition Method for Detecting Epistatic Interactions Contributing to a Quantitative Trait

Robert Culverhouse, Tsvika Klein, Mary Relling, William Shannon

Washington University in St. Louis School of Medicine
660 S. Euclid Ave., Campus Box 8005
St. Louis, MO 63110

Phone/Fax: 314-454-8712/314-454-5113

Email: rob@ilya.wustl.edu, tklein@im.wustl.edu, mary.relling@stjude.org,
shannon@ilya.wustl.edu

Abstract

The restricted partition method (RPM) is a partitioning algorithm for examining multi-locus genotypes as (potentially non-additive) predictors of a quantitative trait. The motivating application was to develop a robust method to examine quantitative phenotypes for epistasis (gene-gene interactions), but the method can be applied without modification to gene-environment interactions. Simulation results indicate that the method provides an efficient way to identify loci contributing epistatically to a quantitative trait, even if the loci have no single locus effects. Statistical significance is estimated through permutation testing, and preliminary results indicate that the method is robust to allele frequency variation. An example using real data involving the metabolism of a chemotherapy drug is included for illustration. Although the examples in the paper involve 2-locus interactions, the RPM is computationally feasible for the analysis of more than two loci or factors.

This work was supported by National Institutes of Health grants R01-GM61218, GM61393, GM61374, GM069690-01, CA 51001, R01-HL71083, the Pharmacogenetics Research Network (GM63340), by a Monsanto/Washington University Grant, and by funds from the Barnes-Jewish Hospital Foundation.

Introduction

Most traditional methods of analyzing data for genetic association ignore or exclude the possibility of non-additive effects. Making this simplifying assumption has proven to be an effective strategy for rare, apparently Mendelian traits such as cystic fibrosis [Riordan et al., 1989], Huntington's disease [Gusella et al., 1983; Huntington's Disease Collaborative Research Group, 1993], and sickle cell anemia [Livingstone, 1967]. However, traditional approaches have thus far not proven themselves in the study of complex phenotypes, generally thought to arise as the result of interactions among multiple genes and environmental exposures. Diabetes [Cox et al., 1999; Elston et al., 1974; Hsueh et al., 2003], depression [Levinson et al., 2003; Marazita et al., 1997; Moldin et al., 1991], and schizophrenia [McGuffin et al., 2003] are known to have large genetic components, but traditional methods of genetic analysis have proven ineffective in explaining the majority of the genetic contribution to these phenotypes.

One factor contributing to this lack of success may be that a sizable portion of the genetic contribution to these phenotypes is due to non-additive interactions (epistasis) between a relatively small network of genes. If this is the case, an examination of multilocus genotypes may prove to be an important tool in understanding the etiology of both disease and drug response.

It is known that important epistatic effects can be present even when single-locus effects are minimal [Culverhouse et al., 2002]. Supporting the theoretical possibility of sizable epistatic interactions in the etiology of complex phenotypes, researchers have found important interaction effects contributing to phenotypes such as breast cancer [Ritchie et al., 2001], triglyceride levels [Nelson et al., 2001], Alzheimer's Disease [Zubenko et al., 2001], and even Sickle-Cell Anemia [Odenheimer et al., 1983; Sing et al., 1985; el-Hazmi et al., 1992]. For an excellent overview article on epistasis see [Templeton 2000]. Although some new methods, such as those used in the papers above, have been proposed to evaluate potential epistatic interactions, a need for methods development remains.

Our current focus is on epistatic models of *quantitative* traits. Our aim is to determine from a quantitative trait dataset if there are subgroups of the multilocus genotypes that correspond to different mean trait values. The straightforward approach of simply testing for any difference among all the multilocus genotypes suffers from several drawbacks including: (1) rejection of the null hypothesis does not suggest a genetic model for further investigation, (2) the p-value does not reflect a measure of the importance of the difference between groups, and (3) potentially low power due to small samples from individual genotypes.

CPM: Combinatorial Partition Method

The Combinatorial Partition Method (CPM) [Nelson et al, 2001] is a method that has been developed to address these issues. It assesses the amount of variation in the quantitative trait explained by each possible partitioning of the multilocus genotypes: the goal is to find a way to divide the multilocus genotypes into subgroups in such a way that the grouping explains a large portion of the overall trait variation.

The CPM comes with a heavy computational price. For instance, for every pair of biallelic loci, there are 21,146 ways to partition the genotypes. When there are n candidate loci involved, these 21,146 evaluations will need to be performed for each of the $n(n-1)/2$ ways to select two loci from the candidates. Although this involves a great number of tests, the fact that it can feasibly be done is demonstrated in the CPM paper [Nelson et al., 2001] where the authors, using serum triglyceride levels as a phenotype, detected clinically interesting interactions between loci that individually showed little or

no effect on the phenotype. Unfortunately, the number of partitions for each set of three biallelic loci is over 10^{21} . Clearly, if interactions involving more than 2 loci are to be analyzed, an alternative method needs to be employed.

The computational burden of the CPM is increased by orders of magnitude if one wants to assess statistical significance because the suggested method for evaluating the statistical significance of the models is permutation testing. This involves running the algorithm on many (the article suggests 1000) permutations of the data set for each pair of loci to generate null distributions.

Method: The Restricted Partition Method (RPM)

In an earlier paper, we suggested a computationally less expensive variant of the CPM called the Restricted Partition Method (RPM) [Culverhouse et al 2004] and tested it on simulated data from a variety of purely epistatic models wherein all the candidate loci had allele frequencies equal to 0.5. In this paper we will present the initial results of testing the RPM in simulated data wherein the candidate loci have a variety of allele frequencies. We also present the results of applying the method to a data set from a drug metabolism study that includes covariates other than genotypes and loci with more than two alleles.

The goal of the RPM, like that of the CPM, is to find partitions of multilocus genotypes that explain a “significant” proportion of the observed trait variation. The motivation for the RPM is the realization that much of the computational burden associated with the CPM is unnecessary. In contrast to the exhaustive approach of the CPM, the RPM algorithm attempts to find the most reasonable partition for evaluation, balancing maximization of the between group variation with minimization of the number of groups and the within group variation. As an example of a partition that does not need to be evaluated, consider a situation where each genotype falls into one of three categories: low, medium, or high trait value. Clearly, putting the lows and highs together in one group and the mediums in a second will not optimally explain the trait variation.

Description of the Restricted Partition Method

The RPM algorithm, described below, is an iterative search procedure for finding the “best” partition of the genotypes. Genotypes are sequentially merged based on the similarity of the mean values of their quantitative trait. Selection of which genotypes to merge at each step is based on statistical criteria from a multiple comparisons test. Initially, each multi-locus genotype forms its own group. The algorithm proceeds as follows:

1. A multiple comparisons test is performed to identify which (if any) groups have different mean quantitative trait values. The procedure halts if all groups have different means.
2. Pairs of genotype groups with means that are not significantly different from each other are ranked according to the difference in means between the two groups.
3. The pair with the smallest difference (i.e., most similar mean values) is merged to form a new group.
4. The algorithm returns to step 1.

As a measure of the importance of the final results, we estimate the R^2 value for the model of the quantitative trait value regressed on the final genotype group. If the genotypes are merged into a single group $R^2 = 0$, reflecting the lack of evidence for quantitative trait differences between the genotypes. The R^2 value is also used to derive a measure of statistical significance for the results (as described below).

The selection of the multiple comparison test is arbitrary. In the simulations and application to real data (presented below) we used a variant of Tukey's HSD multiple comparison method with $\alpha = 0.05$. Other methods we tested appear to perform similarly.

As a measure of the complexity of the algorithm, we note that each time the algorithm reaches Step 2, the number of groups is reduced by one. Thus, if there are initially n genotypes, the algorithm will always halt after no more than $n - 1$ iterations.

To illustrate, consider the case where the genotypes are derived from two biallelic loci. The individual data points will initially fall into one of nine groups (the 2-locus genotypes). If the first multiple-comparison test shows that all nine 2-locus genotypes have significantly different means, the procedure halts and R^2 is computed for the nine-factor model. If the algorithm proceeds to Step 2, two of the genotype groups will be merged and the new collection of eight groups will be assessed. Since there were initially nine genotype groups, the final partition will be determined after no more than eight iterations of the algorithm followed by one R^2 computation. In contrast, the CPM analysis of the same pair of loci would require R^2 to be calculated 21,146 times.

The computational difference is even more striking if biallelic loci are analyzed for possible 3-way interactions. The RPM would require at most 26 iterations to choose the partition of interest compared to more than 10^{21} partitions that would need to be evaluated using the CPM. For the RPM, given a sample size large enough to make the analysis meaningful, analyzing possible 4-way interactions (maximum of 80 iterations) or more would be computationally feasible.

Because the true model R^2 can vary widely, significance cannot be ascertained by the estimated R^2 alone. We estimate p-values for the R^2 using a permutation test. An empirical null distribution for the values of R^2 produced by the RPM are created by permuting the trait values in the data and running the RPM on the permuted data. Significance is estimated by the frequency with which the R^2 obtained from the original data exceeds the permuted R^2 values, correcting for the multiple tests using a simple Bonferroni correction.

Simulations

In [Culverhouse et al., 2004], we tested the method on a simulated data from ten two-locus purely epistatic models (models with no single-locus additive or dominance effects) with model R^2 values ranging from 1.5% to 81.8%. In these simulations, both alleles at each of the contributing loci were equally frequent. In addition, eight biallelic loci with equally frequent alleles that did not contribute to the trait were simulated as a test for false positives. For each model, the distribution of the quantitative trait in each genetic subgroup was normally distributed with the same variance for all genotypes. The only difference in trait distributions for the genotypes was the mean value. We found that for these models, the RPM had good power and that the empirical null distribution gave accurate nominal and corrected p-values when analyzing the null loci.

However, the question remained of how the RPM would perform in situations where the allele frequencies of the candidate loci were more natural. To answer this, we simulated data from six families of 2-locus purely epistatic models different allele frequencies in the contributing loci in order to discover how allele frequency variation would affect power. We also simulated non-contributing loci of various allele frequencies to see what effect this would have on the false-positive rate. The variety of allele frequencies in the simulated models is illustrated in Figure 1.

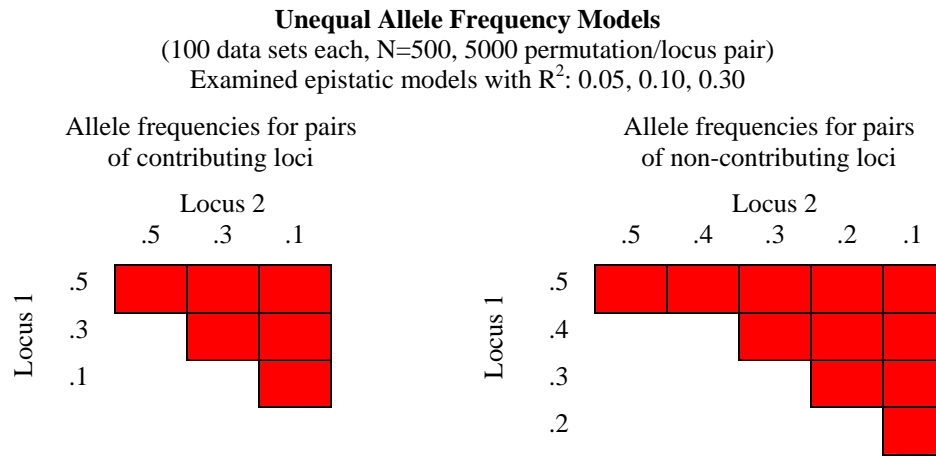


Figure 1

Etoposide Pharmacokinetics Data

Etoposide is a commonly used anticancer agent with a broad range of anti-tumor activity. Data provided to us by Dr. Mary Relling at St. Jude Children's Research Hospital in Memphis contained two pharmacokinetic assessments of etoposide metabolism. Predictor covariates consisted of genotypes from 8 candidate SNPs as well as the demographic factors of ethnicity and sex. Genotypes and phenotypes were available for 102 subjects. Details of phenotyping and genotyping methods used can be found in [Kishi et al., 2004].

The analysis of this small data set is included as a proof of principle that the RPM performs in the expected manner on real data: usually suggesting that there are no subgroups, sometimes detecting subgroups that are statistically non-significant, and occasionally picking out statistically significant substructure in the multi-locus genotype groups.

Results

Simulated Data

As we found in earlier simulations, the RPM showed excellent power with false positives kept at near the nominal level. Results from testing the RPM on six simulation models having true model $R^2 = 0.05$ are illustrated in Table 1. These results indicate that 500 subjects is a reasonable sample size for such models and additional testing indicated that smaller sample sizes would suffice if the 2-locus model R^2 is larger. The first two columns of Table 1 list the minor allele frequencies of the two loci whose genotypes contribute to the trait value. The next two columns indicate the power of the RPM using significance levels of 0.05 and 0.01 after a Bonferroni correction for 45 tests (the number of tests that would be performed if there were 10 candidate loci). To save computing time and to get the particular mix of allele frequencies we wanted to test in the null loci, we tested the 11 combinations of null allele frequencies listed in Figure 1 instead of the 45 combinations that would be tested if there really were 10 candidate loci. Nonetheless, we used corrected significance levels as if we had actually performed the 45 tests to make these results comparable to those in the first paper. We did not see any obvious trends of any particular allele frequency combinations leading to more or fewer false positives. The final three columns give the sum of the false positives detected from testing the 11 null locus pairs in the 100 data sets per model. The first of these columns uses an uncorrected

p-value of 0.05 as a test that the null distribution gives accurate nominal values. The count is followed by the percentage of the tested loci that had nominal p-values less than 0.05. The final two columns give the total number of null-locus combinations having corrected p-values less than the stated level. Since the level was set for 45 tests instead of the 11 actually performed, a rough estimate of the family-wise error rate for 10 candidate loci could be obtained by multiplying these values by 4. This would yield rates very close to the nominal level and are consistent with the more detailed analyses of our earlier models found in Culverhouse, et al. 2004.

Table I: Summary of RPM Results for Models with $R^2 = 0.05$

100 data sets for each model, 500 individuals for each data set
Null distributions based on 5000 permutations

Allele Freq	Contributing Loci (power)			Non-contributing Loci Combined (number of false positives)		
	$p_c < 0.05^1$	$p_c < 0.01$		$p_u < 0.05^2$	$p_c < 0.05$	$p_c < 0.01$
0.5	0.5	0.78	0.68	59 (5.4%)	1	0
	0.3	1.00	0.99	62 (5.6%)	3	1
	0.1	1.00	1.00	49 (4.5%)	1	1
0.3	0.3	0.85	0.71	55 (5.0%)	1	0
	0.1	1.00	1.00	46 (4.2%)	0	0
0.1	0.1	0.71	0.64	68 (6.2%)	1	0

¹ p_c indicates p-value after Bonferroni correction for 45 tests that would need to be performed if there were a total of 10 candidate loci.

² p_u indicates uncorrected p-values. The counts in the non-contributing loci section are totals for 100 trials, each having 11 tests of non-contributing pairs of loci. Since only 11 (instead of 44) pairs of non-contributing loci were tested in each data set, a more realistic estimate of the expected values for the last 2 columns¹ would be to multiply the values by 4.

Etoposide Pharmacokinetics Data

None of the predictors were significant in standard univariate statistical analyses. Because of known allelic heterogeneity between African American and Caucasian Americans, results of a combined analysis are suspect. Nonetheless, an example is included here merely as an illustration of some of the issues involved in performing an RPM analysis.

In our first analysis we combined the data from the two ethnic groups to maximize the sample size, but included ethnicity and sex as covariates (hence analyzing $C(10,2)=45$ pairs of covariates for their potential effect on 2 traits). In this analysis, neither ethnicity nor sex combined with any single SNP proved a significant predictor of trait variation. However, a combination of two SNPs did appear to have a significant effect, as illustrated in Figure 2. The RPM partitioned the 2-locus genotypes base on UGT1A1 and MCRC ex26 into 3 groups. This partition of the genotypes accounted for 26.6% of the variation in the trait 'Etoposide AUC' and resulted in an empirical p-value of 0.045 after a Bonferroni correction for 90 tests. In Figure 2, the cells are numbered by the group to which the corresponding genotype was assigned. The unnumbered cells correspond to genotypes for which there were no observations.

frequent alleles. The primary goal of that study was to explore how power and type I error were affected by changes in the amount of trait variation explained by the 2-locus genetic model. This seemed a reasonable first test, even though SNPs with two equally frequent variants are the exception rather than the rule.

For the tests reported in this paper, the simulated data was more realistic in that the candidate loci had a variety of allele frequencies. Preliminary analyses suggest that the RPM is robust with respect to variation of allele frequencies, demonstrating excellent power while the empirical null distributions appeared to provide an accurate estimate of the rate of false positives.

The application of the RPM to etoposide pharmacokinetics data should be viewed primarily a demonstration that the method performs as expected in real data as well as in simulated data. The analysis in the combined data (illustrated in Figure 2) provides an illustration that the method performs reasonably even if many of the potential multilocus genotypes are not observed. Although this is of technical interest, because of the known admixture of ethnic groups in that analysis, the particular result is of suspect biological merit.

The first result in Figure 3 illustrates that the covariates used by RPM to define subgroups do not need to be genotypes, but can be any qualitative covariate such as an environmental exposure or, as in this case, sex. It is interesting to note that if the GSTP genotype is not subdivided by sex, the '22' genotype corresponds to a mean trait value of 3.86, a value intermediate to that of the high and low groups. This helps explain why the GSTP genotype alone did not predict a significant proportion of the trait variation in the African American sample, even though GSTP and sex could account for over 60% of the variation. Again, because of the small sample sizes, both of the results in Figure 3 are probably most useful as illustrations of the behavior of the RPM rather than as true reflections of underlying biology.

These preliminary results have encouraged us to believe that the RPM is a useful tool for examining complex phenotype data for potential interacting covariates. Although the computational complexity is much less than that of the CPM, the permutation testing required for the RPM is still time consuming and it will generally not be practical to generate detailed empirical null distributions for every combination of candidate loci or to use a common null distribution to evaluate more than one multi-locus combination. To address this issue we are currently implementing a screening approach by which null distributions based on a small number of permutations can eliminate most of the multi-locus candidates quickly, requiring detailed empirical null distributions to be generated only for the few multi-locus combinations with suggestive results from the screening test. Although our particular method is ad hoc, the approach can be formalized in a sequential testing framework [Dixon, Massey, 1969].

As we continue to develop the RPM, we are beginning to explore the utility of generating larger null distributions and of varying the parameters in the multiple comparisons test used for merging groups. We have also begun investigating the performance of the RPM when the true model involves more than two loci: evaluating three-way RPM analyses as well as the effectiveness of using two-way analyses to identify contributing loci if the true model involves 3 loci.

Current difficulties that we will continue to address are an improvement in the method of correcting for multiple testing and a method to assess the robustness of the solution partition under perturbations in the data. In the near future we intend to extend the RPM to the analysis of qualitative traits.

Software implementing the RPM is available from the authors

References

- Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A. (1999). "Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans." *Nat Genet* **21**:213–215.
- Culverhouse R, Klein T, Shannon W. (2004). "Detecting epistatic interactions contributing to quantitative traits." *Gen Epi*. In press. Published online 26 April 2004.
- Culverhouse R, Suarez BK, Lin J, Reich T. (2002). "A perspective on epistasis: Limits of models displaying no main effect." *Am J Hum Genet* **70**:461-471.
- Elston RC, Namboodiri KK, Nino HV, Pollitzer WS. (1974). "Studies on blood and urine glucose in Seminole Indians: indications for segregation of a major gene." *Am J Hum Genet* **26**:13–34.
- Dixon WJ, Massey FJ. (1969). *Introduction to statistical analysis*. McGraw Hill, New York.
- Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, et al. (1983). "A polymorphic DNA marker genetically linked to Huntington's disease." *Nature* **306**:234-238.
- Hsueh WC, St Jean PL, Mitchell BD, Pollin TI, Knowler WC, Ehm MG, Bell CJ, Sakul H, Wagner MJ, Burns DK, Shuldiner AR. (2003). "Genome-wide and fine-mapping linkage studies of type 2 diabetes and glucose traits in the Old Order Amish: evidence for a new diabetes locus on chromosome 14q11 and confirmation of a locus on chromosome 1q21-q24." *Diabetes* **52**:550-7.
- Huntington's Disease collaborative Research Group. (1993). "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes." *Cell* **72**:971-983.
- Kishi S, Yang W, Boureau B, Morand S, Das S, Chen P, Cook EH, Rosner GL, Schuetz E, Pui CH, Relling MV. (2004). "Effects of prednisone and genetic polymorphisms on etoposide disposition in children with acute lymphoblastic leukemia." *Blood* **103**:76-72.
- Levinson DF, Zubenko GS, Crowe RR, DePaulo RJ, Scheftner WS, Weissman MM, Holmans P, Zubenko WN, Boutelle S, Murphy-Eberenz K, MacKinnon D, McInnis MG, Marta DH, Adams P, Sassoon S, Knowles JA, Thomas J, Chellis J. (2003). "Genetics of recurrent early-onset depression (GenRED): Design and preliminary clinical characteristics of a repository sample for genetic linkage studies." *Am J Med Genet* **119B**:118-30.
- Livingstone FB. (1967). *Abnormal Hemoglobins in Human Populations*. Aldine Publishing Co. Chicago, IL.
- Marazita ML, Neiswanger K, Cooper M, Zubenko GS, Giles DE, Frank E, Kupfer DJ, Kaplan BB. (1997). "Genetic segregation analysis of early-onset recurrent unipolar depression." *Am J Hum Genetics* **61**:1370-1378.
- McGuffin P, Tandon K, Corsico A. (2003). "Linkage and association studies of schizophrenia." *Curr Psychiatry Rep* **5**:121-7.
- Moldin SO, Reich T, Rice JP. (1991). "Current perspectives on the genetics of unipolar depression." *Behav Genet* **21**:211-242.
- Nelson MR, Kardina SL, Ferrell RE, Sing CF. (2001). "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation." *Genome Res* **11**:458-70.
- Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, Drumm ML, Iannuzzi MC, Collins FS, Tsui LC. (1989). "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA." *Science* **245**: 1066-1073.

- Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. (2001). "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer." Am J Hum Genet **69**:138-47.
- Templeton AR (2000). "Epistasis and complex traits." In: Epistasis and the Evolutionary Process (Wolf J, Brodie III B, Wade M, eds.) Oxford University Press 41-57.
- Westfall PH, Young SS. (1993). Resampling-based multiple testing. John Wiley and Sons, New York.
- Zubenko GS, Hughes HB 3rd, Stiffler JS. (2001). "D10S1423 identifies a susceptibility locus for Alzheimer's disease in a prospective, longitudinal, double-blind study of asymptomatic individuals." Mol Psychiatry **6**: 413-9.