Genetic Epidemiology 26: 11–21 (2004)

Principal Component Analysis for Selection of Optimal SNP-Sets That Capture Intragenic Genetic Variation

Benjamin D. Horne[1,2] and Nicola J. Camp[1]

[1]*Genetic Epidemiology Division, Department of Medical Informatics, University of Utah, Salt Lake City, Utah*
[2]*Cardiovascular Department, LDS Hospital, Salt Lake City, Utah*

Candidate gene association studies often utilize one single nucleotide polymorphism (SNP) for analysis, with an initial report typically not being replicated by subsequent studies. The failure to replicate may result from incomplete or poor identification of disease-related variants or haplotypes, possibly due to naive SNP selection. A method for identification of linkage disequilibrium (LD) groups and selection of SNPs that capture sufficient intra-genic genetic diversity is described. We assume all SNPs with minor allele frequency above a pre-determined frequency have been identified. Principal component analysis (PCA) is applied to evaluate multivariate SNP correlations to infer groups of SNPs in LD (LD-groups) and to establish an optimal set of group-tagging SNPs (*gt*SNPs) that provide the most comprehensive coverage of intragenic diversity while minimizing the resources necessary to perform an informative association analysis. This PCA method differs from haplotype block (HB) and haplotype-tagging SNP (*ht*SNP) methods, in that an LD-group of SNPs need not be a contiguous DNA fragment. Results of the PCA method compared well with existing *ht*SNP methods while also providing advantages over those methods, including an indication of the optimal number of SNPs needed. Further, evaluation of the method over multiple replicates of simulated data indicated PCA to be a robust method for SNP selection. Our findings suggest that PCA may be a powerful tool for establishing an optimal SNP set that maximizes the amount of genetic variation captured for a candidate gene using a minimal number of SNPs. *Genet Epidemiol* 26:11–21, 2004.

Key words: group-tagging SNP (*gt*SNP); haplotype-tagging SNP (*ht*SNP); linkage disequilibrium; haplotype block; association analysis

Grant sponsor: NIH; Grant numbers: CA99844, CA90752, CA89600, GM31575, and HL073117.
*Correspondence to: Benjamin D. Horne, Genetic Epidemiology, University of Utah, 391 Chipeta Way, Suite D, Salt Lake City, UT 84112.
E-mail: benjamin@genepi.med.utah.edu
Received 12 May 2003; Accepted 7 August 2003
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.10292

INTRODUCTION

Despite the recent explosion in knowledge about the human genome and tools for its evaluation, identification of genetic determinants of many diseases remains clouded. The search for genes involved in common, complex, multifactorial diseases is proving challenging. Recent evidence suggests association studies may provide the best means to study these diseases [Lander, 1996; Risch and Merikangas, 1996; Collins et al., 1997]. Although association studies may provide greater ability to discern the independent contribution of multiple environmental and genetic factors, the wide variability in findings between studies evaluating particular candidate genes and comparable disease states suggests the need for further refinement of analytic techniques [Risch, 2000].

The variability between studies may, in part, result from a failure to adequately account for the patterns of linkage disequilibrium (LD), or intragenic correlation, between a disease-causing variant and the single nucleotide polymorphism (SNP) used as the marker of that variant [Tabor et al., 2002]. It has been suggested that DNA fragments are inherited as a limited number of ''haplotype blocks'' (HB) wherein recombination is rare [Gabriel et al., 2002], and that these HBs may be adequately characterized by selection of a limited set of optimal ''haplotype-tagging'' (*ht*) SNPs from among the vast array of sequence

variants that are commonly found within a gene [Johnson et al., 2001; Daly et al., 2001]. These *ht*SNPs are then utilized for evaluation of the gene's association to disease endpoints.

An assumption that underlies the HB framework is that SNPs in LD will be on contiguous DNA fragments, with the disequilibrium decaying as the distance between SNPs increases. While this may be a reasonable assumption when dealing with extended chromosomal regions and very common SNPs (minor allele frequency >0.30, say), with a higher resolution or rarer SNPs, such as variants found within a gene, their existence may be more likely due to recent mutation rather than recombination. In this situation, SNPs in high LD do not necessarily lie on contiguous DNA fragments or "blocks." Considering this, we frame the problem with a subtle difference. First we determine LD-groups (which need not be contiguous and for which a single SNP may occur in more than one group), and then introduce the concept of group-tagging SNPs (*gt*SNPs) selected as the SNP-set that captures the desired proportion of the total genetic diversity. We utilize Principal Component Analysis (PCA) to perform both steps. PCA is a well-established method for determining the contributions of individual variables to a set of independent, indirectly observed factors. PCA calculates a factor loading for each variable, or component, within each factor. When squared, this factor loading represents a multivariate r^2. Concurrently but in parallel with another group [Meng et al., 2003], we have developed a novel application of PCA to determine LD-groups of SNPs and select the optimal set of *gt*SNPs for use in candidate gene association analyses. Using data simulated as part of Genetic Analysis Workshop (GAW) 12, we demonstrate the application of this PCA method to 100 replicates of data, provide a detailed illustration of an association analysis using *gt*SNPs selected by this novel application for one replicate, and compare the results to those of published HB/*ht*SNP methods.

MATERIALS AND METHODS

PRINCIPAL COMPONENT ANALYSIS

PCA determines the linear combinations that have maximum variances for factors $Y_1, Y_2, \ldots, Y_p$, where p factors ($Y_i = \mathbf{a}'_{i\bullet}\mathbf{X}$) each have the principal components $\mathbf{X} = \Sigma\mathbf{a}_{i\bullet}$ (with $\mathbf{X}'=[X_1, X_2,\ldots, X_p]$) having factor loadings $\mathbf{a}'_{i\bullet}= [a_{i1}, a_{i2}, \ldots, a_{ip}]$ and a covariance matrix Σ of eigenvalues $\lambda_1 \ldots \lambda_p$ where $\mathrm{Var}(Y_i)=\lambda_i$ [Johnson and Wichern, 1999]. In PCA, factors are fixed as orthogonal: $\mathrm{Cov}(Y_i,Y_i)=0$. Factor loadings ($\mathbf{a}_{i\bullet}$), or coefficients, are measures of the multivariate correlation of principal components within a factor, with any a_{ij}^2 being interpreted as the square of a multivariate correlation coefficient (similar to the bivariate r^2 or Δ^2).

DETERMINING LD-GROUPS AND gtSNPS

A two-stage PCA protocol (as opposed to the one-step method of Meng et al. [2003]) was developed to determine a gene's LD-groups and to identify the *gt*SNPs that were most parsimonious yet best characterize each group's intragenic variance. PCA was used to examine the SNP haplotypes as the factors within a gene and SNPs as the constituent components of each factor. For estimation of haplotypes and their frequencies, multiple algorithms have been proposed [Clark, 1990; Excoffier and Slatkin, 1995; Zhao et al., 1999], and we utilized Clayton's SNPHAP package that employs the Expectation-Maximization algorithm with "trimming" of improbable assignments.

Stage 1. In the first stage of our protocol, LD-groups were determined from a PCA considering only SNP haplotypes with frequency of at least 0.01. While convention suggests that factors should be extracted only when the initial eigenvalue >1.0, to capture a greater amount of the intragenic variance we used the more liberal criterion of >0.7, as suggested by Jolliffe [1986]. The extracted factors then underwent oblique rotation (using the Oblimin method) to remove inter-factor correlations. Although Meng et al. [2003] proposed the use of the orthogonal "Varimax" rotation method, genetic theory suggests that groups of SNPs within a gene will have some degree of inter-relatedness due to origin on a common ancestral haplotype (or on a limited set of haplotypes). Thus, since oblique rotation can account for such relatedness, it likely is a more appropriate method for genetic analysis.

The Scree plot and the cumulative percent of variance explained by LD-groups (ordered by rotated eigenvalue) were used to determine the number of important LD-groups within the gene, with a goal of accounting for at least 90% of the variation of the extracted factors (more comprehensive coverage than the criteria of 80% used by Patil et al. [2001] for haplotype coverage).

For LD-group determination, SNPs within the retained factors with $|a_{ij}| > 0.4$ [Stevens, 1992] were considered to constitute an LD-group.

Stage 2. In the second stage, for selection of *gt*SNPs, each LD-group was considered separately using PCA. For these analyses, all SNP haplotypes, regardless of frequency, were used in the analysis. The number of factors within each separate LD-group analysis indicates the number of *gt*SNPs necessary to represent each LD-group, with each *gt*SNP chosen as the SNP that provided the best characterization of the variance within each factor, as measured by its factor loading (i.e., for *gt*SNP [j] for LD-group [i], $a_{ij} > a_{ik}$ for all $k \propto$ [1, …, j−1, j+1, …, p]). The optimal number of *gt*SNPs for an LD-group was considered to be the number of extracted factors (eigenvalue > 0.7) that explain $> 90\%$ of the group's variance where possible (only one *gt*SNP is needed if PCA extracts just one factor). Multi-group analyses may also be performed if an inability to completely rotate out all dependence between factors is an issue, or because one SNP belongs to multiple LD-groups, although this usually will simply confirm the initial findings. Single SNP factors do not require the second stage of the PCA protocol, while SNPs with equal factor loadings in the second stage may require consideration of other data, possibly including the functionality of the SNP, the physical location relative to other *gt*SNPs, the results of the first stage of PCA, and any differential cost or ease in genotyping of each. In some cases, several SNPs may be found in stage 2 to have such very similar factor loadings that they may be considered to be equally loaded.

DESCRIPTION OF SOURCE DATA

For the Genetic Analysis Workshop (GAW) 12, a simulated data set including 50 replicates of 23 extended pedigrees was created for a general ''continental'' population, with 1,000 living individuals per replicate for whom genetic data were available [Almasy et al., 2001]. The full sequences of seven candidate genes were simulated. Gene lengths were between 13,000–20,000 base pairs and each gene had 12–40 diallelic SNPs. Genes were segregated through 120,000–200,000 generations with simulated population level LD, recombinant hot spots, and mutation. Phenotypic data were also simulated, including age, gender, environmental variables E1 and E2, and disease affection. The prevalence of the disease in the overall population was 25%.

SNP DISCOVERY

We chose to analyze data from *GENE6*, a 17,000-bp-length candidate gene encoded with a causative variant influencing several quantitative traits and disease affection. A program to perform *in silico* sequencing of *GENE6* was distributed with the data. We used the option that allowed identification of all variants with at least 2 copies of the minor allele occurring in the sequencing of 50 individuals (i.e., approximate minor allele frequency > 0.02).

We chose 50 independent individuals with genotype data from replicates 1, 16, and 42, with no more than one individual from any of the 69 pedigrees (23 pedigrees, 3 replicates). Replicate 42 was chosen because it was suggested to be the ''best'' replicate because it represented the generating parameters most closely, and replicate 1 was said to be a ''random'' replicate [Thomas et al., 2001]; replicate 16 was chosen at random.

SELECTION OF INDIVIDUALS FOR gtSNP DETERMINATION

The SNPs identified in the discovery phase were then evaluated by the PCA protocol among 384 unrelated individuals selected randomly and without regard to affection status to estimate SNP haplotypes, LD-groups, and *gt*SNPs. Individuals were selected from the 1,081 pedigrees available in the remaining 47 GAW12 replicates not used in the SNP discovery stage (23 pedigrees $\times$ 47 replicates), with no more than one individual selected per pedigree to assure unrelatedness. This was repeated 100 times, without replacement, to generate 100 replicates of 384 independent individuals on which we performed our PCA method. A total of 38,400 unique individuals were evaluated. The sample size of 384 is similar to that in other studies of LD: 129 offspring from trios [Daly et al., 2001] and 384 unrelated individuals [Johnson et al., 2001], and simulates the maximum number of samples in one laboratory micro-well plate.

Further, to evaluate our PCA protocol in a smaller sample size, we repeated our analyses on 100 individuals in 10 of the PCA replicates. The PCA protocol was run on each of these 10 reduced-size replicates and results compared to the PCA results from the corresponding larger (384 individual) analyses.

COMPARISON OF PCA LD-GROUP/gtSNP METHOD TO HB/htSNP METHODS

For comparison to established methods, we also determined haplotype blocks (HBs) and/or

haplotype-tagging SNPs (*ht*SNPs) in four alternate software packages: Clayton's *ht*SNP "STATA utility" [Johnson et al., 2001] and the program "tagSNPs" [Stram et al., 2003] were run in a PC environment; "SNPtagger" [Ke and Cardon, 2003] and "HaploBlockFinder" [Zhang and Jin, 2003] were used on-line.

EVALUATION OF ASSOCIATION OF gtSNPS TO DISEASE ENDPOINTS

Each of the 23 SNPs evaluated in the five *gt*SNP/*ht*SNP methods were then evaluated in an association analysis. The primary endpoint considered was disease affection. We chose a study population of 1,150 unrelated individuals for this analysis, selected at random from among the 50 GAW12 replicates of the simulated GAW12 data, discounting any individuals that had been used at the SNP discovery or LD-group/*gt*SNP determination stages. No disease or demographic restriction was placed on the selection of the study population, and this sample size reflected that of each of the 50 GAW12 replicates. The chi-square test was used to evaluate the association of each *gt*SNP with disease, and logistic regression was performed for multivariate analyses of association with adjustment for age, sex, and the two environmental risk factors.

RESULTS

SNP DISCOVERY AND HAPLOTYPES

Twenty-three SNPs were identified by *in silico* SNP discovery in 50 individuals within *GENE6* (Table I). Allele frequencies for these 23 SNPs estimated from the first replicate population of 384 unrelated individuals are also shown in Table I. Sixty-three haplotypes were identified, 15 having frequency greater than 1% (Table II).

TABLE I. Allele frequencies for SNPs discovered by *in silico* sequencing of *GENE6* among 50 unrelated individuals, and frequencies as estimated from 384 individuals in the replicate demonstrating the LD-Groups and *gt*SNPs

SNP bp position	SNP Discovery (n=50) Minor allele frequency	LD-groups/*gt*SNPs (n=384) Minor allele frequency
993	0.16	0.23
1748	0.16	0.23
1987	0.16	0.24
4411	0.16	0.23
4848	0.38	0.33
5007	0.11	0.11
5782	0.10	0.11
6805	0.10	0.11
7073	0.10	0.11
7332	0.10	0.12
8067	0.10	0.12
8226	0.10	0.12
9616	0.24	0.23
9954	0.07	0.02
10054	0.24	0.23
10955	0.24	0.24
11782	0.27	0.28
11981	0.12	0.12
12408	0.08	0.11
13869	0.27	0.29
14425	0.07	0.05
14544	0.05	0.10
15021	0.37	0.24

DETERMINATION OF LD-GROUPS AND SELECTION OF gtSNPS

Results for PCA among all 23 SNPs in the 15 haplotypes with frequency >0.01 are found in Table III for the first replicate. These are generally representative of the results of each of the 100 PCA replicates (see below). For the first replicate, seven factors with eigenvalues >0.7 were found, explaining 97% of the variance. Six LD-groups were identified to satisfy the $>90\%$ threshold criterion (explaining 93%). Only 4 LD-groups would be required to cover a less stringent $>80\%$ threshold. Interestingly, LD-group 4 was the seventh factor extracted in most replicates, but when rotated its eigenvalue had higher ranking (4th) and would be included as an LD-group even for the lower $>80\%$ threshold.

PCA analyses for each LD-group separately are shown in Table IV for the first replicate. LD-group 4 is not shown in Table IV as this group only contains a single SNP. For LD-group 1, the analysis extracted 2 factors and hence the two SNPs with highest factor loadings within these were chosen as *gt*SNPs {7332, 14544}. For LD-groups 2 and 3, one factor was extracted with *gt*SNPs {9616} and {1987}, respectively. LD-group 4 had only one SNP with a factor loading >0.4 and thus being its *gt*SNP: {15021}. These five *gt*SNPs accounted for 83% of the intragenic variance: {1987, 7332, 9616, 14544, 15021}. To account for more than 90%, an additional two LD-groups and *gt*SNPs were needed, resulting in a 7-SNP set {1987, 7332, 9616, 12408, 14425, 14544, 15021}. In these two additional groups, there existed two SNPs with equal factor loadings. The *gt*SNPs were chosen based on their superior loadings in stage 1 of the PCA protocol.

TABLE II. Haplotypes (≥0.01 frequency) determined from 23 SNPs among the first replicate of 384 unrelated individuals

Haplotype																							Frequency
993	1748	1987	4411	4848	5007	5782	6805	7073	7332	8067	8226	9616	9954	10054	10955	11782	11981	12408	13869	14425	14544	15021	
1	1	2	2	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	0.22
1	1	2	2	1	1	1	1	1	1	1	2	2	1	2	2	2	1	1	2	1	1	1	0.12
2	2	1	1	1	2	2	2	2	2	2	1	2	1	2	2	2	1	1	2	1	1	1	0.10
2	2	1	1	1	1	1	1	1	1	1	2	2	1	2	2	2	1	1	2	1	2	1	0.07
1	1	2	2	1	1	1	1	1	1	1	2	2	1	2	2	2	2	1	2	1	1	1	0.07
1	1	2	2	2	1	1	1	1	1	1	2	2	1	2	2	2	1	2	2	1	1	1	0.06
1	1	2	2	2	1	1	1	1	1	1	2	2	1	2	2	2	1	1	2	1	1	1	0.05
1	1	2	2	1	1	1	1	1	1	1	2	2	1	2	2	2	1	1	2	1	1	2	0.05
1	1	2	2	1	1	1	1	1	1	1	2	2	1	2	2	1	1	1	1	1	1	1	0.03
1	1	2	2	1	1	1	1	1	1	1	2	2	1	2	2	2	2	1	2	1	1	2	0.03
1	1	2	2	1	1	1	1	1	1	1	2	2	1	2	2	2	1	1	2	2	1	1	0.02
2	2	1	1	1	1	1	1	1	1	1	2	2	1	2	2	2	1	1	2	1	2	2	0.02
1	1	2	2	1	1	1	1	1	1	1	2	2	1	2	2	2	1	2	2	1	1	2	0.02
1	1	2	2	1	1	1	1	1	1	1	2	2	1	2	2	2	1	1	2	2	1	2	0.01
1	1	2	2	1	1	1	1	1	1	1	2	2	2	2	2	2	1	1	2	1	1	2	0.01

The complete *gt*SNP sets are shown in Figure 1. If all LD-groups with eigenvalues >0.7 were included (97% of variance explained), 7 LD-groups are defined, and an 8-SNP set, additionally including SNP 9954.

COMPARISON TO OTHER METHODS

Figure 1 also demonstrates the results from four HB/*ht*SNP software packages. SNPtagger and HaploBlockFinder identify both HBs and *ht*SNPs, and require contiguous HBs. The STATA utility and tagSNPs program identify only *ht*SNPs, and do not supply any information about HB structure.

In SNPtagger, using 0.9 coverage of haplotypes, 13 HBs were identified. Finding *ht*SNPs by this method, however, simply used the first SNP on a block regardless of the relative merits of the SNPs, and resulted in a set of 13 *ht*SNPs {993, 1987, 4848, 5007, 8226, 9616, 9954, 11782, 11981, 12408, 14425, 14544, 15021}. For HaploBlockFinder, using >90% chromosomal coverage, HaploBlockFinder found five HBs but chose 15 *ht*SNPs to cover these {993, 1782, 4848, 5007, 8226, 9954, 10054, 10955, 11782, 11981, 12408, 13869, 14425, 14544, 15021}.

The STATA utility and tagSNPs provide no data on HBs or the optimal size for the *ht*SNP set. For these programs, the user chooses the size of the SNP set and the algorithms determine the best set of that selected size. In our PCA analyses, we found 5 LD-groups to cover >80% of variance and 7 covering >90%. Hence, we identified the best 5-SNP and 7-SNP sets from both programs. The STATA utility using default options showed the best 5-SNP subset as {1987, 4848, 13869, 14544, 15021}. This utility also determines each SNP's diversity rating, and the five selected *ht*SNPs were among those with the highest individual diversity ratings (locus and haplotype diversity are measures of the total number of differences between all of the pair-wise locus or haplotype comparisons [see Johnson et al., 2001]). Seven-*ht*SNP sets simply added two more SNPs to the previous five: {11981, 12408}. Each of these *ht*SNP subsets were the same when measured either by Clayton's residual mean diversity and percent of diversity explained. For tagSNPs, the best 5-SNP set was {4848, 5007, 11782, 12408, 13869}, while the 7-SNP set had several different *ht*SNPs included {4411, 5007, 11981, 12408, 13869, 14454, 15021}.

PCA REPLICATES (384 INDIVIDUALS)

We repeated our PCA analyses in 100 replicates. As found in the first replicate, 7 factors were extracted with initial eigenvalue>0.7 in 87% of the replicates, with 6 being extracted in the other 13%. The failure to extract 7 factors in those 13 replicates occurred when the initial eigenvalue of the seventh group was below 0.7; however, without exception and regardless of the number of LD-groups extracted, all eigenvalues were >1.0 after rotation in all replicates. The number of *gt*SNPS selected from the 6 or 7 LD-groups, was 7, 8, 9, or 10 in 3, 41, 55, and 1 replicate(s), respectively, and accounted for an average 95% of the intragenic variance (range: 92–97%). To

TABLE III. Principal component analysis solution for LD-groups in *GENE6*[a]

	LD groups						
bp position	1	2	3	4	5	6	7
993	0.38		−0.78				
1748	0.38		−0.78				
1987	−0.38		0.78				
4411	−0.38		0.78				
4848		−0.63			0.56		
5007	0.99						
5782	0.99						
6805	0.99						
7073	0.99						
7332	0.99						
8067	0.99						
8226	−0.99						
9616		1.0					
9954			0.11	0.13	−0.11	0.11	0.94
10054		1.0					
10955		1.0					
11782		0.92					
11981		0.21	0.32	0.27	−0.39	0.44	−0.34
12408		0.24			0.93		
13869		0.92					
14425		0.12	0.14	0.12	−0.13	−0.92	−0.13
14544	−0.48		−1.00				
15021				0.99			0.10
Eigenvalue:	9.0	6.1	5.4	2.0	1.6	1.1	1.1
Cumulative variance explained (%) (goal: >90%):	33	56	76	83	89	93	97

[a]Factor loadings are presented for each SNP, with ≥0.40 considered necessary for inclusion in an LD-group (values below 0.10 are suppressed). LD-groups are ordered according to the rotated eigenvalue.

TABLE IV. Principal component analysis solution for *gt*SNP Selection in *GENE6*[a]

		*gt*SNP factors	
Group	SNP	1	2
1	5007	0.991	
	5782	0.991	
	6805	0.994	
	7073	0.991	
	7332*	0.995*	
	8067	0.992	
	8226	−0.971	
	14544*		1.0*
Cum. Var. Explained:		*86%*	*98%*
2	4848	−0.74	
	9616*	0.970*	
	10054	0.94	
	10955	0.966	
	11782	0.92	
	13869	0.92	
Cum. Var. Explained:		*83%*	
3	993	0.980	
	1748	0.982	
	1987*	−0.984*	
	4411	−0.97	
	14544	0.68	
Cum. Var. Explained:		*86%*	
5	4848	0.80	
	12408*	0.80*	
Cum. Var. Explained:		*63%*	
6	11981	0.74	
	14425*	−0.74*	
Cum. Var. Explained:		*54%*	

[a]The highest factor loading indicates the SNP selected (*). LD-groups with more than one factor require multiple *gt*SNPs (factor loadings <0.10 are suppressed). LD-groups 4 and 7 contained only a single SNP and hence did not require further PCA.

explain >80% of the variance, though, only the first four LD-groups were required in 98 replicates, as was illustrated for the first replicate (see Table III).

Table V shows the frequency of each SNP being the highest loaded (or tied for the highest loaded SNP) on the LD-groups extracted across multiple replicates. The LD-groups noted in Table V are from the first replicate. LD-groups 1, 2, and 3 (ordered by rotated eigenvalue) remained largely the same in all replicates, with the only exception being that SNPs 993, 1748, 1987, and 4411 also loaded on LD-group 1 in 43 replicates (with loadings of 0.40–0.70). However, their inclusion in LD-group 1 in stage 1 of the PCA protocol had no effect on the *gt*SNPs chosen at stage 2 in any of those replicates. Further, examination across replicates of the cut-off of 0.40 for loading an SNP onto an LD-group indicated that decreasing the loading threshold may lead to inclusion of increased SNPs in the LD-groups, but would have little effect on the *gt*SNPs subsequently chosen. Increasing the 0.4 threshold would, however, have excluded SNP 14544 from a majority of analyses of LD-group 1 since its factor loadings usually ranged 0.40–0.60, and thus would have also excluded 14544 from selection as a *gt*SNP in almost all instances. The remaining 3 or 4 LD-groups were not in constant order, but were effectively representative of the same groups of SNPs. Interestingly, use of a one-stage PCA would have selected SNP 14544 as a *gt*SNP and excluded SNPs 993, 1748, 1987, and 4411 from selection in

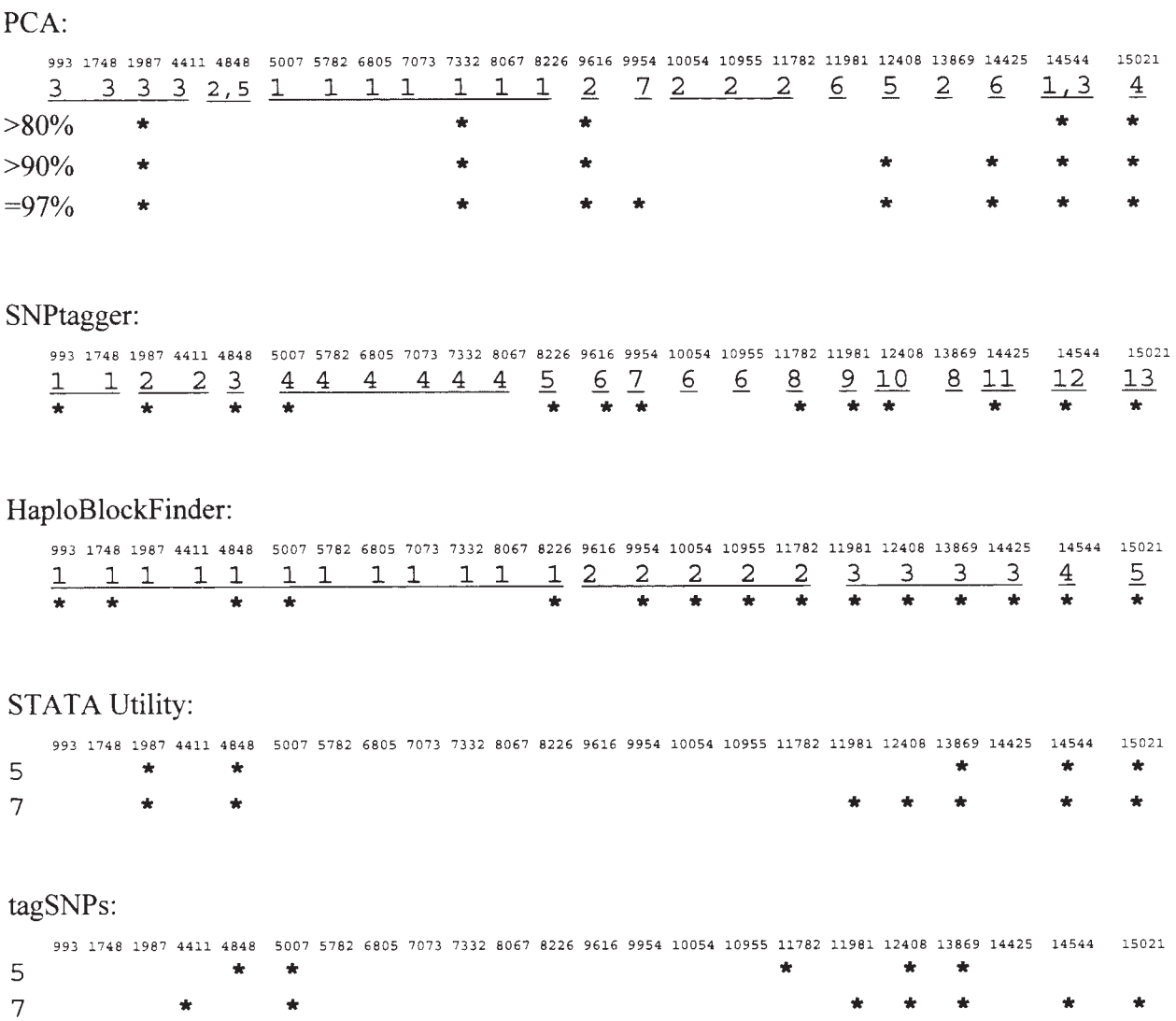

Fig. 1. LD-groups/HBs (numbered and underlined) identified by PCA, SNPtagger, and HaploBlockFinder, and the *gt*SNPs/*ht*SNPs (*) identified by each of the five methods. For PCA, accounting for >80% of the intragenic variation required 5 *gt*SNPs, while 7 were needed to account for >90% of the variation, and 8 if all factors with eigenvalue>0.7 are included (97% of the variation). For the STATA utility and tagSNPs, because they did not provide HBs, subsets of 5 and 7 *ht*SNPs were evaluated to compare to PCA.

all replicates; use of the two-stage PCA, however, selected both 14544 and one of those other four, more fully accounting for the variance in LD-groups 1 and 3.

PCA REPLICATES OF 100 INDIVIDUALS

Comparison of the smaller sample size of 100 in 10 of the replicates showed essentially the same results as in the 384 individuals. Overall, the number of LD-groups and the proportion of the variance explained by the LD-groups in the 100 individual analyses were consistently lower (average 93%). This is expected given data from fewer chromosomes, although the loss in variance explained was nominal given that the sample size here was about 25% the original size. The higher ordered LD-groups were essentially the same as in PCA of 384 individuals, although there was less power to identify the remaining factors, and fewer LD-groups were identified, with 5, 6, or 7 LD-groups extracted in 4, 3, and 3 replicates, respectively.

Based on the reduced number of LD-groups, fewer *gt*SNPs were required to tag all LD-groups (8 or 9 *gt*SNPs in 7 and 3 replicates, respectively). Some increased variation was seen in which SNP loaded the highest (Table V), but results generally

TABLE V. Association to disease affection of *gt*SNPs/*ht*SNPs after adjustment for age, sex, and two environmental factors (Relative risks [RR] are for carriage of a minor allele) in 1,150 independent individuals[a]

				No. of times chosen as *gt*SNP for >90% (or >80%) variance coverage in:	
SNP	RR	*P* value	1st Replicate's PCA LD-group	100 reps./384 ind.	10 reps./100 ind.
993	1.40	0.02	3	1	3
1748	1.37	0.03	3	16	4
1987	**1.41**	**0.02**	**3**	**90**	**7**
4411	1.44	0.01	3	10	4
4848	1.04	0.81	2,5	40 (0)	6
5007	2.09	0.00003	1	9	5
5782	**2.14**	**0.00002**	**1**	**62**	**5**
6805	2.16	0.00002	1	44	3
7073	2.13	0.00002	1	15	4
7332	2.24	0.000004	1	32	6
8067	2.33	0.000001	1	18	5
8226	2.21	0.000005	1	0	3
9616	1.14	0.37	2	22	4
9954	**2.63**	**0.22**	**7**	**96 (0)**	**9 (0)**
10054	1.19	0.23	2	0	3
10955	**1.15**	**0.33**	**2**	**84**	**3**
11782	1.01	0.92	2	0	1
11981	**0.85**	**0.36**	**6**	**94 (0)**	**10 (0)**
12408	**0.97**	**0.86**	**5**	**100 (0)**	**8 (0)**
13869	1.07	0.66	2	2	4
14425	**1.09**	**0.75**	**6**	**98 (0)**	**10 (0)**
14544	**0.92**	**0.65**	**1,3**	**88**	**8**
15021	**1.59**	**0.003**	**4**	**96**	**9**

[a]LD-group is shown for the first replicate (384 individuals), and while generally correct for most analyses was not strictly the same in all PCAs. The number of times that each SNP was chosen as a *gt*SNP in simulations of 100 replicates (for 384 individuals) or 10 replicates (for 100 individuals) when accounting for >90% (or >80%, if different) of the observed intragenic variance are also shown. SNPs that were loaded the highest in their groups for >60% of the replicates are indicated in **bold**.

mirrored those of the larger sample size analyses. LD-group 2 was the exception where SNP 10955 was not chosen as often (only 3 out of 10) and instead SNPs 4848, 9616, and 13869 were selected 6, 4, and 4 times, respectively.

EVALUATION OF DISEASE ASSOCIATIONS

The 1,150 individuals in the association analyses had an average age of 46 years and 51% were female. No differences in age, sex, or E2 were found for any *gt*SNP, and only two (4848 and 10955) had significant differences across the environmental variable E1. Overall, 26% (n=295) of the study population was affected.

Table V illustrates the association analysis results after adjustment in regression for age, sex, and the environmental variables. The true causal variant in the simulated data was the SNP at position 5782. This SNP exhibited a significant association with affection ($P=0.00005$), and belongs to LD-group 1. As shown in Table V, among the 100 replicates the PCA method chose SNP 5782 as a *gt*SNP 62 times, more than any other SNP in the same LD-group. Although they did not choose 5782, other methods did choose *ht*SNPs in the same LD-group as 5782, for which significant evidence was also found (5007, $P=0.00009$ chosen by tagSNPs; 8226, $P=0.00002$ chosen by SNPtagger and HaploBlockFinder). SNP 15021 in LD-group 4 ($P=0.004$, chosen by PCA and all 4 other methods), and the SNPs in LD-group 3 (1987, $P=0.02$ chosen by PCA and STATA), also indicated weak significance after adjustment for multiple comparisons. Covariates age, gender, and E1 (all $P<0.0001$) were associated with affection, but not E2. No significant intra-genic interactions were detected using the method of Cordell and Clayton [2002], as was expected, since none were simulated.

DISCUSSION

Determination of LD-groups/HBs and *gt*SNPs/*ht*SNPs is a burgeoning area of study, with only

very recent developments. In 1998, Cox et al. illustrated the complex LD patterns for multiple SNPs across a 430-kb region for three IL-1 cytokine genes on chromosome 2 and suggested interactive models across this region for association with inflammatory disease risk [Cox et al., 1998]. Recent evidence supports this use of multiple SNPs, including a study of the HLA region in insulin-dependent diabetes [Cordell and Clayton, 2002] and a study evaluating 122 SNPs across regions containing 9 different genes [Johnson et al., 2001]. In this latter study, Johnson discovered that subsets of SNPs were adequate to tag the most common haplotypes, with only 34 *ht*SNPs needed to fully characterize the HBs, instead of the originally typed 122 SNPs. An application of this method for disease association demonstrated significant association of a tagged haplotype to Crohns disease [Rioux et al., 2001]. With the increasing interest in candidate genes and association studies in complex, common diseases, and with the failure of single SNP studies to deliver on risk associations, the determination of informative sets of SNPs to genotype is crucial.

This study demonstrates a method for determination of LD-groups and the selection of group-tagging SNPs, *gt*SNPs, for use in evaluating the association of a candidate gene to a disease phenotype. In the first replicate, examining 23 common SNPs identified in *GENE6*, 5–8 *gt*SNPs were chosen from 4–7 LD-groups (dependent on the percentage of the genetic variance required to be covered), and several of these *gt*SNPs detected the simulated association with disease affection. The proportion of the intragenic variance desired to be covered by the selected LD-groups/*gt*SNPs had considerable impact on the number of LD-groups and *gt*SNP selection. Given this, it is important that investigators describe the coverage of their selected SNP-set when reporting association results, both positive and negative. Our results suggest that at least 80% of the variation covered should be required, although choices on SNP selection may be partly governed by cost of genotyping in a study. In our example, $>80\%$ was sufficient to find significant association to multiple variants in *GENE6* in every replicate, and included selection of the causative variant in 62% of replicates. However, our example was simplistic in the sense that a single casual variant existed in the simulated data such that associations of independent SNPs successfully extracted the association evidence. Further, the fact that our method selected the causal variant in the majority of replicates is most likely due to chance, or the way the data were simulated, since the *gt*SNPs were selected without respect to affection status. In more complex, and perhaps more realistic, situations where intragenic epistatic effects or allelic heterogeneity exist, the definition of the lower order LD-groups and SNP haplotypes may be more crucial in extracting the underlying associations. Therefore, where costs allow, increasing the threshold to $>90\%$ of the genetic variance would be recommended.

Our two-stage PCA method was applied to 100 replicates of simulated data from GAW12 to identify *gt*SNPs in *GENE6* for use in evaluating the association of that gene to disease affection. The number and selection of LD-groups and *gt*SNPs was found to be generally robust across replicates. The 100 replicates of 384 unrelated individuals showed that this proposed two-stage PCA protocol with thresholds for eigenvalues >0.7 for LD-group extraction and factor loading >0.4 for SNP inclusion in a group were appropriate. Evaluation of this protocol in a set of 10 of the replicates with a reduced sample size (100 individuals) also indicated that the method is robust to the smaller sample size of 100 individuals.

This PCA method provides several advantages over other HB and *ht*SNP methods, while performing comparably to those methods on their strong points. Advantages of PCA include that it is a well-established method for determining the contributions of individual variables to a set of independent, indirectly observed factors and is readily available in most commonly used statistical packages (e.g., SAS, SPSS, STATA). The theory of PCA is also similar to that of traditional measurement of LD between SNPs, in that the calculated factor loadings for each SNP within each LD-group, when squared, represents a multivariate r^2.

The PCA method evaluates the degree to which the genetic variation within a gene or within an LD-group is accounted for by each component group or SNP, respectively, and allows the determination of LD-groups and the selection of the best *gt*SNPs from within those LD-groups. In PCA, the LD pattern is relatively easy to visualize (indicating the LD-groups and *gt*SNPs), and allows the investigator insight of the genes LD structure (including recombination and mutation), which may aid design or interpretation of subsequent association analyses and results. In other

methods evaluated in this study, it was more difficult to visualize the haplotype blocks (HBs), or they did not provide this information at all (STATA utility and tagSNPs). The methods that did provide haplotype block determination (SNPtagger and HaploBlockFinder) required these to be contiguous, which resulted in a loss of information to the investigator of the true LD structure. In addition, this led to an increase in the number of HBs, since these were broken into their contiguous subsets, and subsequently an increased number of *ht*SNPs chosen, which is inefficient.

Another strength of PCA that is not provided by other methods is its ability to determine the optimal number of the *gt*SNPs needed to efficiently describe the complete intragenic variance. The PCA method first determines the number of LD-groups (factors) needed to explain the intragenic variance (we used 90%, or 80%, as our threshold). Second, the method is repeated within groups to determine the number of gtSNPs (intra-group factors) needed to explain the intra-group variance. This is not to say that the other methods did not also select useful SNP-sets in our example. However, they are more limited in the information available for determining the appropriate number of SNPs necessary for a comprehensive association analysis, and some methods required the user to define this value (STATA utility and tagSNPs).

Other considerations also exist for choosing the number of, or specific, SNPs to genotype in a candidate gene. This might include the proportion of the variance that an investigator desires to account for, the consideration of SNPs reported in the literature, the ease of genotyping certain SNPs, and the financial costs of genotyping multiple SNPs. In such cases, methods such as the STATA utility or tagSNPs are useful since the user specifies the number of SNPs, and may also "force" certain SNPs into the selection. PCA may also be applied to such situations, though, wherein a situation of limited finances may, for example, be approached by genotyping only one or a few SNPs that best tag the largest, most encompassing LD-groups (such as LD-group 1 in this study), or selecting the one or few SNPs that travel with multiple LD-groups (such as 14544 or 4848 in this study). A systematic *a priori* approach to reducing the number of SNPs used or required for detecting a disease association is, however, not straightforward.

The value of this PCA method has been further highlighted by the recent publication of a similar PCA method developed in parallel by Meng and associates [Meng et al., 2003]. Several differences exist between the two approaches to PCA. Our method employs a two-step procedure to determine LD-groups first and then *gt*SNPs, providing the investigator with useful information on the LD structure and what the SNPs are tagging, which may indicate likely regions of the gene that harbor causal variants. Further, multiple *gt*SNPs may be chosen to represent a single LD-group, whereas the one-step Meng method is more limited in extracting SNPs that tag variance within the gene but not necessarily all variance within all LD-groups. For example, SNP 14544 may not have been extracted as a *gt*SNP using Meng's method. Additionally, Meng proposed use of the Varimax orthogonal rotation for PCA, whereas, as noted above in Materials and Methods, an oblique rotation such as the Oblimin method may be more appropriate for genetic analysis.

CONCLUSIONS

The promise of genomics is in the complete and accurate characterization of genetic risk for common diseases. Such diseases, however, are primarily complex chronic diseases and typically the genetic risk for these results from many small-effect genes. To comprehensively assess associations of disease with genetic variants in candidate genes, identification of the optimal markers for analysis is essential. By identification and use of optimal sets of SNPs, through methods such as those of this study, clear genetic association signals may become detectable. This should enable replication, and increase comparability across studies. The selection of *gt*SNPs, independent of phenotype, allows hypotheses to be prospectively tested and, thus, improves the validity of study findings (with only a small increase in genotyping price and statistical multiple-testing cost).

The use of PCA in this study to determine LD-groups and select *gt*SNPs has performed with much promise. In the application here, the PCA method was able to identify 4–7 LD-groups that explained the vast majority (83–97%) of the total genetic variation within a gene. Further, the *gt*SNPs selected were successful in extracting highly significant evidence for the simulated association with disease. The use of PCA for determination of LD patterns, SNP selection, and quantifying the importance of LD-groups and

*gt*SNPs to the total genetic variation within a gene clearly warrants further exploration.

ACKNOWLEDGMENTS

This study was supported in part by NIH grants CA99844, CA90752, and CA89600 (N.J.C.), and HL073117 (B.D.H., N.J.C.). Simulation of the GAW12 data was supported by NIH grant GM31575.

ELECTRONIC DATABASE INFORMATION

URLs for the haplotype-estimation program and comparison HB/*ht*SNP packages used in this study are: SNPHAP program and htSNP STATA utility, http://www-gene.cimr.cam.ac.uk/clayton/software TagSNPs, http://www-rcf.usc.edu/~stram/SNPtagger, http://www.well.ox.ac.uk/~xiayi/haplotype/HaploBlockFinder, http://cgi.uc.edu/cgi-bin/kzhang/haploBlockFinder.cgi

REFERENCES

Almasy L, Terwilliger JD, Nielsen D, Dyer TD, Zaykin D, Blangero J. 2001. GAW12: simulated genome scan, sequence, and family data for a common disease. Genet Epidemiol 21(Suppl 1): S332–S338.

Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122.

Collins FS, Guyer MS, Charkravarti A. 1997. Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581.

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70:124–141.

Cox A, Camp NJ, Nicklin MJH, di Giovine FS, Duff GW. 1998. An analysis of linkage disequilibrium in the interleukin-1 gene cluster, using a novel grouping method for multiallelic markers. Am J Hum Genet 62:1180–1188.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. Nat Genet 29:229–232.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237.

Johnson RA, Wichern DW. 1999. Applied multivariate statistical analysis, 4th ed. Upper Saddle River, NJ: Prentice-Hall, Inc.

Jolliffe IT. 1986. Principal component analysis. New York: Springer-Verlag.

Ke X, Cardon LR. 2003. Efficient selective screening of haplotype tag SNPs. Bioinformatics 19:287–288.

Lander ES. 1996. The new genomics: global views of biology. Science 274:536–539.

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG. 2003. Selection of genetic markers for association analysis, using linkage disequilibrium and haplotypes. Am J Hum Genet 73:115–130.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.

Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, Kulbokas EJ, O'Leary S, Winchester E, Dewar K, Green T, Stone V, Chow C, Cohen A, Langelier D, Lapointe G, Gaudet D, Faith J, Branco N, Bull SB, McLeod, Griffiths AM, Bitton A, Greenberg GR, Lander ES, Siminovitch KA, Hudson TJ. 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29:223–228.

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

Risch NJ. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856.

Stevens JP. 1992. Applied multivariate statistics for the social sciences, 2nd ed. Hillsdale, NJ: Erlbaum.

Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC. 2003. Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered 55:27–36.

Tabor HK, Risch NJ, Myers RM. 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev 3:391–397.

Thomas DC, Boreki IB, Thomson G, Weiss K, Almasy L, Blangero J, Nielsen D, Terwilliger J, Zaykin D, MacCluer J. 2001. Evolution of the simulated data problem. Genet Epidemiol 21(Suppl 1):S325–S331.

Zhang K, Jin L. 2003. HaploBlockFinder: haplotype block analyses. Bioinformatics 19:1300–1301.

Zhao H, Pakstis AJ, Kidd JR, Kidd KK. 1999. Assessing linkage disequilibrium in a complex genetic system I. Overall deviation from random association. Ann Hum Genet 63:167–179.