Original Article

Identification of *PVT1* as a Candidate Gene for End-Stage Renal Disease in Type 2 Diabetes Using a Pooling-Based Genome-Wide Single Nucleotide Polymorphism Association Study

Robert L. Hanson,[1] David W. Craig,[2] Meredith P. Millis,[2] Kimberly A. Yeatts,[2] Sayuko Kobes,[1] John V. Pearson,[2] Anne M. Lee,[2] William C. Knowler,[1] Robert G. Nelson,[1] and Johanna K. Wolford[2]

To identify genetic variants contributing to end-stage renal disease (ESRD) in type 2 diabetes, we performed a genome-wide analysis of 115,352 single nucleotide polymorphisms (SNPs) in pools of 105 unrelated case subjects with ESRD and 102 unrelated control subjects who have had type 2 diabetes for ≥10 years without macroalbuminuria. Using a sliding window statistic of ranked SNPs, we identified a 200-kb region on 8q24 harboring three SNPs showing substantial differences in allelic frequency between case and control pools. These SNPs were genotyped in individuals comprising each pool, and strong evidence for association was found with rs2720709 ($P = 0.000021$; odds ratio 2.57 [95% CI 1.66–3.96]), which is located in the plasmacytoma variant translocation gene *PVT1*. We sequenced all exons, exon-intron boundaries, and the promoter of *PVT1* and identified 47 variants, 11 of which represented nonredundant markers with minor allele frequency ≥0.05. We subsequently genotyped these 11 variants and an additional 87 SNPs identified through public databases in 319-kb flanking rs2720709 (~1 SNP/3.5 kb); 23 markers were associated with ESRD at $P < 0.01$. The strongest evidence for association was found for rs2648875 ($P = 0.0000018$; 2.97 [1.90–4.65]), which maps to intron 8 of *PVT1*. Together, these results suggest that *PVT1* may contribute to ESRD susceptibility in diabetes. *Diabetes* 56: 975–983, 2007

Diabetic nephropathy is the leading cause of end-stage renal disease (ESRD) in developed countries (1). In the Pima Indians of Arizona, 95% of ESRD cases occur in diabetic subjects, and among individuals with diabetes, 97% of ESRD is attributable to diabetic nephropathy (2). In this population, diabetic offspring are three times as likely to develop nephropathy if both parents have proteinuria than if neither parent has proteinuria (3). Although the causes of diabetic nephropathy are not fully understood, familial aggregation of the disease and disproportionate prevalence among specific ethnic minority groups suggest that genetic factors may influence the risk of developing the disease (4–9). Segregation analysis of diabetic kidney disease in Pimas supports a major genetic effect for disease prevalence after accounting for duration of diabetes (10).

Historically, linkage analysis and candidate gene investigation have been commonly used to identify genes that increase susceptibility to complex disease. Compared with association studies, linkage analysis can be less effective for detecting alleles conferring modest risk (11). Recently, novel technologies and emerging statistical tools have enabled the use of genome-wide association (GWA) studies for mapping susceptibility genes for complex disease (12). GWA studies are attractive because they do not rely on a priori knowledge of protein function or disease relevance. However, GWA studies are costly to undertake and require large numbers of carefully phenotyped case and control subjects; consequently, they are presently beyond the fiscal reach of most research groups (13). To circumvent the costs associated with GWA studies, several groups have explored the feasibility of genome-wide scans using pooled genomic DNA. Multiple studies have assessed the accuracy of allelic frequency predictions using the Affymetrix 10K or 100K platforms (14–22). In these studies, fairly good agreement in allele frequency differences was observed in genotyping between pools of individuals, suggesting that this approach may be appropriate for identification of susceptibility alleles for complex disease.

To identify genes with major effects on the development of diabetic ESRD, we sought to apply new approaches for gene discovery in susceptible individuals. The goal of this study was therefore to combine the emerging technology of high-density single nucleotide polymorphism (SNP) microarrays with a pooled genomic DNA design to identify novel loci for genes predisposing to ESRD in Pima Indians with type 2 diabetes.

RESEARCH DESIGN AND METHODS

All subjects in this study are participants in a longitudinal study of type 2 diabetes and its complications conducted in the Gila River Indian Community since 1965, in which individuals are invited to have a health examination every

From the [1]Diabetes Epidemiology and Clinical Research Section, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Phoenix, Arizona; and the [2]Translational Genomics Research Institute, Phoenix, Arizona.

Address correspondence and reprint requests to Johanna K. Wolford, PhD, Translational Genomics Research Institute, 445 North Fifth St., Phoenix, AZ 85004. E-mail: jwolford@tgen.org.

Received for publication 1 August 2006 and accepted in revised form 26 November 2006.

ACR, urinary albumin-to-creatinine ratio; ESRD, end-stage renal disease; GWA, genome-wide association; LD, linkage disequilibrium; LDU, linkage disequilibrium unit; RAS, relative allele signal; SNP, single nucleotide polymorphism.

DOI: 10.2337/db06-1072

2 years (23); urinary albumin-to-creatinine ratio (ACR) has been assessed at each examination since 1982. Some of these individuals (33%) participated previously in genome-wide linkage studies for diabetes and diabetic nephropathy (24,25). A detailed description of the study participants, including clinical diagnosis of kidney dysfunction, can be found elsewhere (26). Briefly, the study group included 105 case subjects with ESRD and 102 control subjects with diabetes duration >10 years and a maximum ACR observed in the longitudinal study of <300 mg/g. All individuals were of full Pima (Akimiel O'odham) or Tohono O'odham (a closely related tribe) heritage, and none were first-degree relatives of another individual in the sample. The mean ages (±SD) of case and control subjects were 55.9 ± 8.9 and 58.0 ± 9.7 years, respectively. There were no statistically significant differences between case and control subjects in the mean age of diabetes onset (±SD) (35.5 ± 8.7 [case] versus 38.1 ± 9.7 years [control]) or diabetes duration (±SD) (20.4 ± 7.1 [case] versus 20.7 ± 5.5 years [control]). At their last examination, 47% of control subjects had ACR <30 mg/g, while 53% had ACR 30–299 mg/g. The studies were approved by the institutional review boards of NIDDK and the Translational Genomics Research Institute. All subjects provided informed consent for participation in the study.

Pool construction. Genomic DNA concentrations were determined using the NanoDrop ND-1000 spectrophotometer according to the manufacturer's instructions (NanoDrop Technologies, Wilmington, DE). DNA integrity was verified by gel electrophoresis. No samples showed evidence of DNA degradation, indicated by band smearing during electrophoresis. One hundred nanograms of each genomic DNA sample, corresponding to a volume of 1.5–2.5 μl, was used in the appropriate pool. Before SNP genotyping, each pool was diluted to 50 ng/μl with TE buffer (10 mmol/l TrisHCl, 0.1 mmol/l EDTA, pH 8.0). Pools were constructed in duplicate for case and control subjects, and each pool was genotyped on three replicate chips.

Whole-genome SNP genotyping. DNA pools were genotyped using the Affymetrix 100K Human Mapping set according to the manufacturer's protocol (Affymetrix, Santa Clara, CA).

Analysis of SNP microarray data. For each SNP, data from six replicate array sets for both case and control subjects were available. Probe intensity data were exported from the Affymetrix GType 4.008 software and converted to two relative allele signal (RAS) values using a previously developed Perl script (21). RAS values correspond to the ratio of the major allele probe to the sum of the major and minor allele probes and provide a quantitative index of allele frequencies in pooled DNA (18). Because both sense and antisense directions are probed, two RAS values are generated, RAS_s (sense) and RAS_a (antisense), which yield independent measures of different hybridization events and are consequently treated as individual data points. Because suitable data for calibration of signal intensities from sense and antisense probes were not available, the directly measured intensities were analyzed. Differences in RAS_a and RAS_s between case and control pools were quantified using the silhouette statistic. This silhouette statistic (27), which represents the mean of the distance of a point to all other points in its class (e.g., case pool) relative to points in the other class (e.g., control pool), was used to rank all SNPs. Silhouette statistics range from 1, where complete separation between pools has been achieved, to −1, where allele frequencies do not meaningfully distinguish between case and control pools. The calculation for a silhouette statistic is shown in equation 1:

$$S = \frac{\sum_{i=1\ldots N} s(i)}{N},\ s(i) = \frac{b(i) - a(i)}{\max[a(i),b(i)]}$$

where the overall silhouette statistic, S, is the average of all of the individual silhouette values and $s(i)$ is for each of the class comparisons, N refers to the number of replicate measures, i represents the replicate array within a class, $a(i)$ refers to the average Euclidian distance of RAS_a and RAS_s for replicate i of a class to all other replicates within its class, and $b(i)$ refers to the average distance of a replicate (i) to all replicates not within its class. In this study, the overall silhouette statistic of one SNP is the average of $N = 12$ $s(i)$ values.

Reproducibility of pooled data on Affymetrix arrays was assessed by calculation of the Pearson correlation coefficient for RAS values on replicate case and control arrays; overall, arrays correlated extremely well with one another, exhibiting correlation coefficients of 0.98–0.99. Following this quality check, SNPs were ranked by silhouette statistic, ranging from the highest score of 1 to the lowest score of 115,352. With each sample ranked by silhouette statistic, we calculated a sliding window statistic of the mean rank for 2–31 consecutively neighboring SNPs. By using information from a number of markers simultaneously, the sliding window method identifies regions where neighboring SNPs consistently show differences in allele frequency between case and control subjects; this minimizes spurious differences arising from technical anomalies in the analysis of pooled DNA. Although a range of window sizes was used, a window size of 25 consecutive SNPs provided the strongest evidence for allelic frequency differences. The evidence of association for a given window was expressed as a P value, which was calculated by randomly permuting the SNP ranks 10,000 times and recalculating the sliding window statistic. All calculations, including computation of the silhouette statistic, were completed using MATLAB 7.0 (MathWorks, Natick, MA).

SNP detection. Because the above analyses identified associations in *PVT1*, sequencing studies were initiated to identify variants in this gene. The *PVT1* genomic sequence (chromosome 8: 128,875,522–129,182,684, build 35.1) was obtained from the UCSC Genome Browser (v.129; http://genome.ucsc.edu) and used to design primers for SNP detection. All exons, exon-intron boundaries, and 3 kb of 5′ flanking sequence, comprising a total of 21 kb, were screened by direct sequencing using genomic DNA obtained from 36 Pima individuals (18 ESRD case subjects and 18 diabetic control subjects). DNA was amplified and sequenced as previously described (28), and sequences were resolved on the 3730xl sequence analysis system (Applied Biosystems). Sequencing chromatograms were analyzed using Mutation Surveyor software version 2.61 (SoftGenetics, State College, PA).

An additional 87 markers spanning a 319-kb interval around *PVT1* were also selected from the HapMap database (http://www.hapmap.org) based on physical position and a minor allele frequency ≥0.05 in the genotyped families of each ethnic group available in the HapMap resource. Genotype data from the CEU (CEPH Utah) population were compiled and downloaded into the Tagger program (http://www. broad.mit.edu/mpg/tagger/) to identify a parsimonious set of markers for genotyping; pairwise $r^2 \geq 0.80$ was considered indicative of redundancy in determining the genotyping panel.

SNP genotyping in individuals. All SNPs genotyped in individual samples were assayed with the iPLEX assay in conjunction with the MassARRAY platform (Sequenom, La Jolla, CA). Primers and multiplex conditions were designed using the Assay Design version 3.0 software (Sequenom). DNA amplification and the iPLEX primer extension assay were performed according to the manufacturer's protocol. Reaction products were dispensed onto a 384-element SpectroCHIP bioarray (Sequenom) using a MassARRAY Nanodispenser and assayed on the MassARRAY platform. The MassARRAY Workstation version 3.3 software was used to process and analyze iPLEX SpectroCHIP bioarrays.

Statistical analyses. The extent to which observed genotype frequencies for each SNP deviated from that expected under Hardy-Weinberg equilibrium (HWE) was assessed (χ^2 with 1 d.f.); none of the markers varied significantly from HWE. In addition, encrypted samples were used to assess data quality. The statistical evidence for association and the strength of the association between genotypes and affection status, as determined by the odds ratio (OR) and the corresponding 95% CI, were calculated by logistic regression. For these analyses, an "additive" model was used in which the genotype was coded as a numeric variable representing the number of risk alleles; thus, the OR designates the odds for ESRD associated with each copy of the risk allele.

Haplotype frequencies for multiple loci within *PVT1* were estimated using the EH program (29), and linkage disequilibrium (LD) was quantified using the measures D′ and r^2, which respectively represent the strength of the allelic association and the degree of concordance (30). The LDMAP program was also used to construct an LD map for all 101 genotyped SNPs (31), wherein each SNP was assigned a position in LD units (LDUs), which reflect historic recombination among markers given their physical order. The k-means algorithm (32) was used to assign markers to LD "clusters" based on their position within this map. To analyze the association of common haplotypes within each of these clusters with ESRD status, a set of SNPs informative for such haplotypes was defined from analysis of the pairwise LD—pairs of SNPs for which $r^2 > 0.7$ were considered redundant, and only one of the SNPs was selected for inclusion in the analysis. Common haplotypes within each of these clusters were tested for association with ESRD by a modification of the zero-recombinant haplotyping method (33). The MLINK program (34) was used to assign each individual a probability of carrying one or two copies of a given haplotype, based on their genotypes for the markers considered and the haplotype frequencies. These probabilities were then used as predictor variables in a logistic regression model in a fashion analogous to the analysis of SNPs. An "exhaustive" analysis was conducted for all common haplotypes observed in all combinations of nonredundant SNPs in a cluster.

Pairs of SNPs were also analyzed to examine the association of ESRD with genotypes at one marker conditioned on the association at the other. The Mantel extension test (35) was used to assess the significance of the association between genotypes at one SNP and ESRD stratified by genotypes at a second SNP.

RESULTS

Results of the sliding window analysis for the genome-wide association in pools are shown in Fig. 1. The single

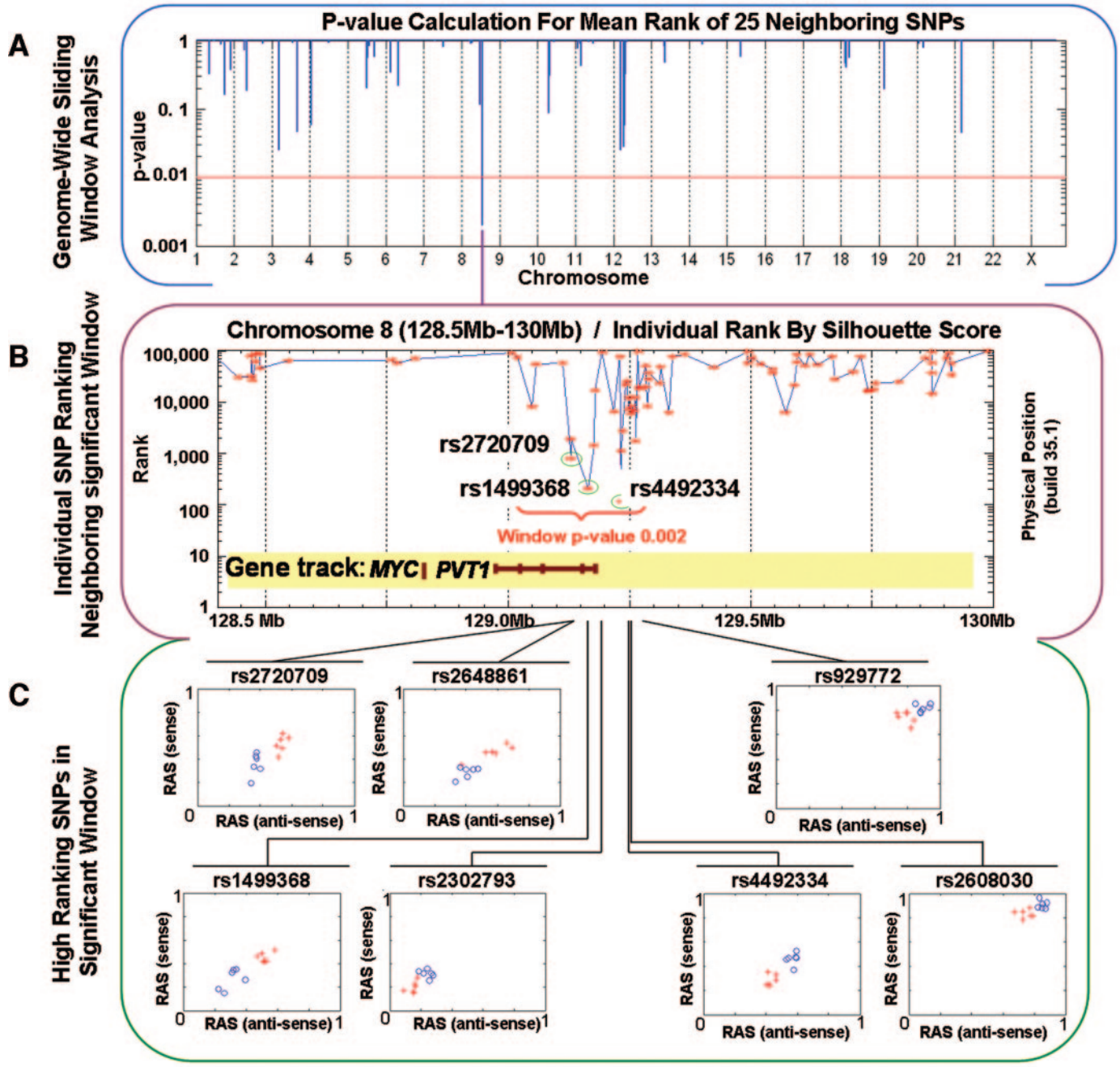

FIG. 1. Ranking of genome-wide allele frequency differences and identification of highly ranked markers on 8q24. SNPs were ranked from 1 to 115,352 by a silhouette test statistic. A sliding window of mean rank for 25 consecutive SNPs was calculated for all SNPs ordered by chromosome and position. *A*: *P* values were calculated by permuting SNP order through 10,000 iterations and recalculating genome-wide sliding window statistics. The region of highest significance and lowest mean rank was on chromosome 8q24.21. *B*: Individual ranks for SNPs in and neighboring 8q24.21 were found to overlap *PVT1*. *C*: The RAS value for the sense (Y-axis) and anti-sense (X-axis) probes are plotted for the seven highest ranking SNPs in the most significant window. In each plot, crosses and circles represent data from the six replicate arrays for case and control pools, respectively.

best window was found at 8q24.21, which was primarily driven by seven SNPs with ranks in the top 3% of all markers (Fig. 1*B* and *C*). The physical position of these SNPs spanned ~250 kb from chromosome 8, position 129.01–129.26 Mb (build 35.1). This window encompassed one gene, the plasmacytoma variant translocation gene PVT1, which spans 128.87–129.18 Mb. The nearest gene to this locus is MYC, which is located ~53 kb centromeric to PVT1; however, no genes are located within 1 Mb telomeric to PVT1. SNPs on the Affymetrix 100K Mapping array closest to MYC were not highly ranked (all were <10,000), suggesting that PVT1 is the major gene encompassed by the region.

To assess whether differences in allele frequency between case and control pools in SNPs located on chromosome 8 were artifactual, we genotyped three of the highest ranking markers, rs2720709, rs1499368, and rs4492334, in the individuals comprising each pool. Each of these SNPs ranked in the top 1% of all markers assessed by silhouette statistic. The largest difference in allele frequency between case and control subjects was observed for rs2720709, where the frequency of the A allele was 0.54 in case

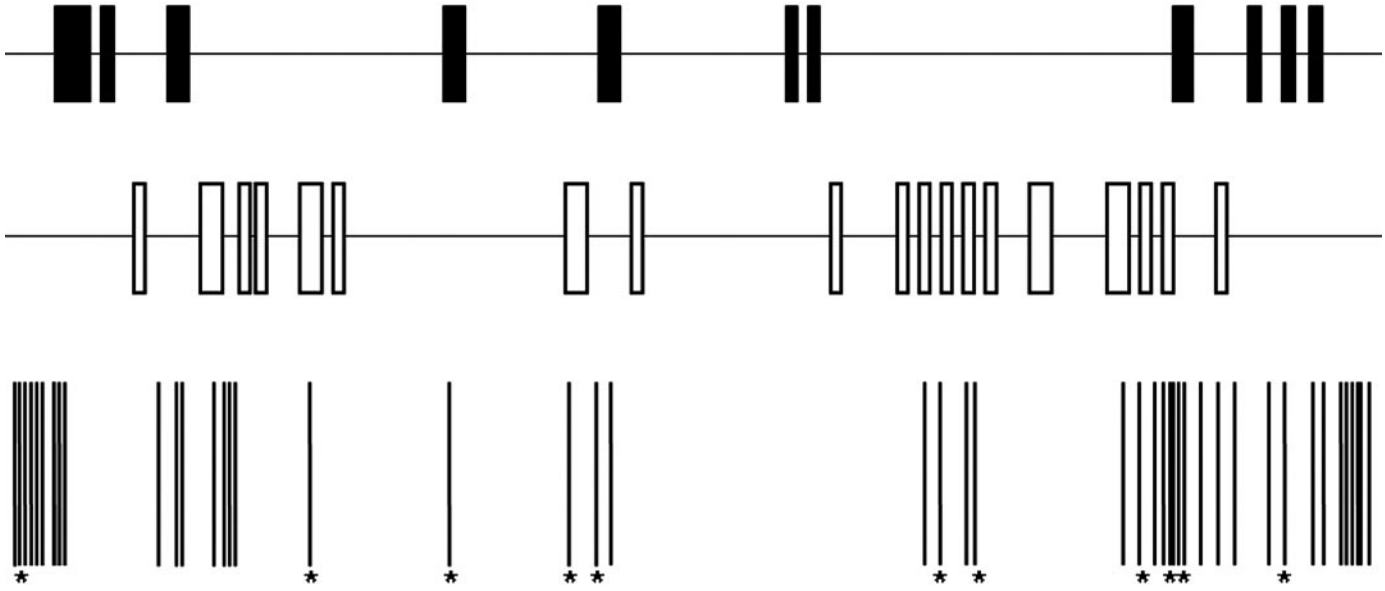

FIG. 2. Identification of variants within *PVT1*. The genomic organization of *PVT1* is shown in the upper figure; black rectangles represent exons and the thin horizontal lines designate noncoding sequence. Gene structure shown reflects the major eleven exons found in the largest and most common *PVT1* transcript. However, alternative splicing events in *PVT1* produce at least 27 isoforms, which differ by variable truncations of the 5′ and 3′ ends, inclusion of 1 or more of the 19 cassette exons available for splicing, shown in the middle, and differential splicing of common exons that yield different boundaries. Variants identified by direct sequencing are shown as thin vertical lines at the bottom of the figure. *Polymorphisms that were genotyped in the study sample.

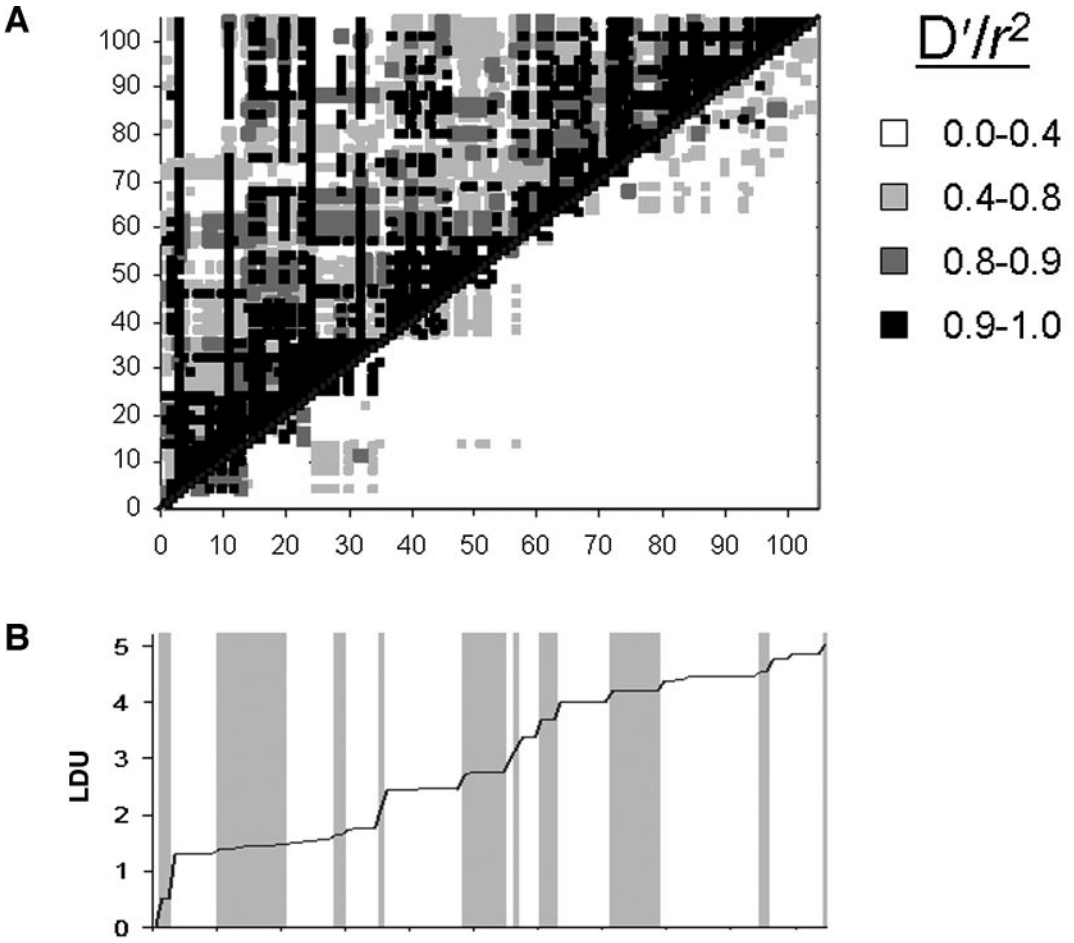

FIG. 3. LD relationships among PVT1 markers. *A*: Measures of LD between pairs of each of the 101 individually genotyped SNPs. D′ is shown above the diagonal, and r^2 is shown below the diagonal. *B*: The bottom panel shows the position of each of the markers on an LD map. Regions where the line is nearly horizontal indicate a low degree of historical recombination among markers, while areas where the slope changes rapidly indicate a region of historical recombination. The shaded areas represent clusters of SNPs among which there was a small difference in LDUs. Clusters were defined by the k-means algorithm (assuming a radius of 0.10 LDU).

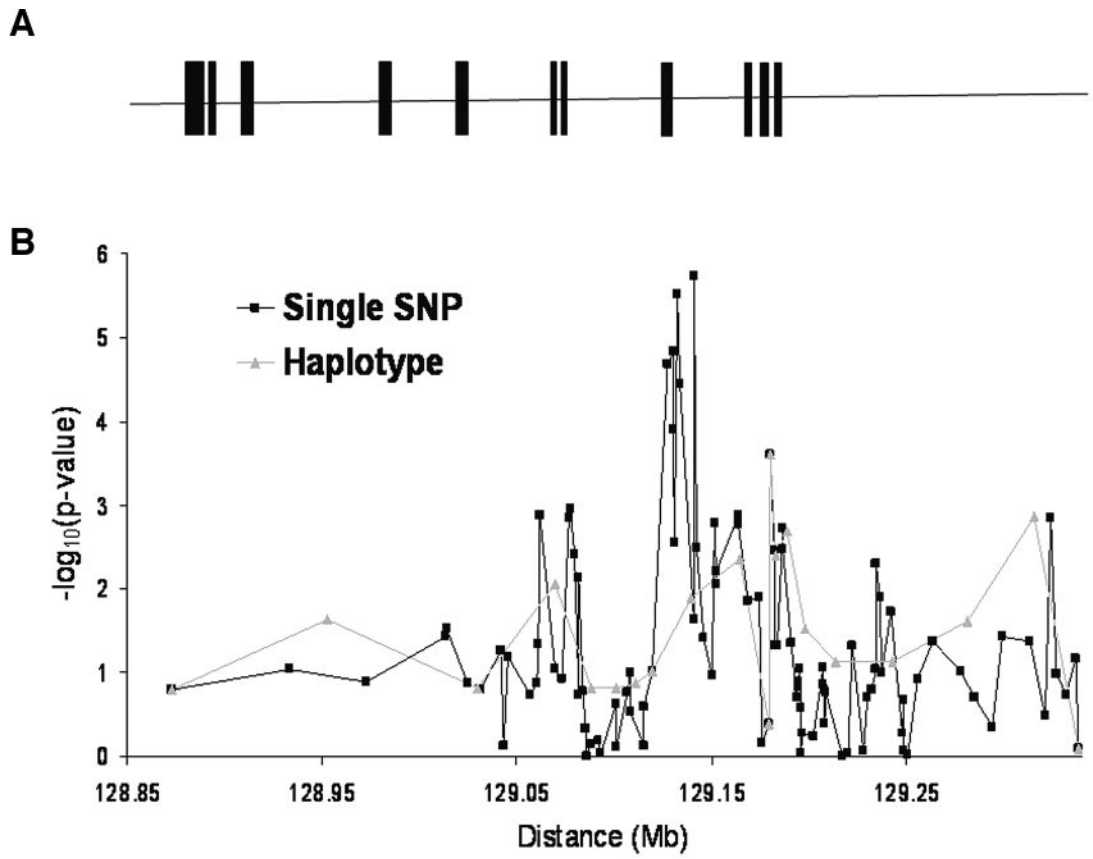

FIG. 4. Association of single markers and haplotypes with ESRD. *A*: Structure of the *PVT1* gene. *B*: *P* values were calculated for each SNP by logistic regression assuming an additive effect of number of alleles on the logarithm of the odds for ESRD. Results for haplotype analyses represent the lowest *P* value obtained for a common haplotype (frequency >0.01) in each LD cluster, with no correction for multiple comparisons.

subjects and 0.33 in control subjects ($P = 2.1 \times 10^{-5}$). SNP rs1499368 was also associated with ESRD; the frequency of the T allele was 0.63 in case subjects and 0.48 in control subjects ($P = 1.3 \times 10^{-3}$), while the association between ESRD and rs4492334 was less strong with a frequency of the A allele of 0.77 in case subjects and 0.70 in control subjects ($P = 0.09$). Concordance between the marker pairs was $r^2 = 0.17$ (rs2720709-rs1499368), $r^2 = 0.07$ (rs2720709-rs4492334), and $r^2 = 0.02$ (rs1499368-rs4492334).

Because rs2720709, rs1499368, and rs4492334 lie within *PVT1*, we screened this gene to identify all common variation within the locus for subsequent genotyping and assessment of association with ESRD. *PVT1* spans over 300 kb on 8q24.21. We identified 47 polymorphisms, including 40 SNPs and 7 insertion-deletion markers (Fig. 2). Thirty-four of the *PVT1* polymorphisms were present in the public SNP database dbSNP (http://www.ncbi.nlm.nih.gov/SNP), while the remaining markers may represent newly identified alleles. Of the 47 markers found in *PVT1*, 31 segregated into clusters of genotypic concordance, including 12 polymorphisms that were redundant with SNPs that were genotyped as part of the LD mapping of the 8q locus (see below) and 19 variants that were in 100% concordance with other markers identified during *PVT1* sequencing. Five markers had a minor allele frequency <0.01 and were not genotyped. In total, 11 nonredundant markers were identified by sequencing. These 11 variants, along with 87 markers ascertained from the HapMap database, were genotyped in the individuals comprising the case-control study group (in addition to the three SNPs genotyped as follow-up to the GWA study). The LD relationships among these 101 SNPs, along with their positions on the LD map, are shown in Fig. 3*A*. SNPs across the region were generally in strong LD (high D′ values), although the information for association provided by a given SNP was, in many cases, distinct from that provided by other SNPs, as indicated by moderate r^2 values. Analysis of the LD map identified 20 "clusters" of SNPs in a very high degree of LD with one another of size 1–15 (Fig. 3*B*). The common haplotypes within each of these clusters could be identified by fewer SNPs (1–5 SNPs per cluster).

Results of the analyses of association between each of the 101 SNPs, or the common haplotypes, with ESRD are shown in Fig. 4. Several SNPs showed a strong association with ESRD, with the lowest *P* values occurring in the vicinity of 129.14 Mb. Twenty-three SNPs were associated with ESRD at $P < 0.01$, and the results of analyses for each of these are shown in Table 1. The strongest evidence for association was observed for rs2648875, which is located in intron 8 of *PVT1*; the frequency of the A allele was 0.77 in case subjects and 0.53 in control subjects (OR 2.97 per copy of the A allele [95% CI 1.90–4.65], $P = 2.0 \times 10^{-6}$). In addition, rs2720662 and rs2720659, both of which were perfectly concordant with rs2648875 ($r^2 = 1.00$), were also strongly associated with ESRD. Most of the SNPs showing the strongest evidence for association, including rs2648875, rs2720662, rs2720659, rs2720709, and rs1499373, which were highly concordant with rs2720709 ($r^2 = 0.99$), were located on a single LD cluster extending from 2.43 to 2.47 LDU (Fig. 4*B*). There was strong allelic association (D′ = 1.0) and moderate concordance ($r^2 = 0.42$) between rs2648875 and rs2720709. In the haplotype analyses, the haplotype

TABLE 1
SNPs in or near *PVT1* associated with diabetic ESRD

Marker	Position (Mb)	Position (LDU)	Genotype	*n* (%) case subjects	*n* (%) control subjects	OR (95% CI)	*P*
rs11993333	129.061669	1.4113	CC	63 (62)	41 (40)		
			CT	34 (33)	49 (48)		
			TT	5 (5)	13 (13)	2.10 (1.33–3.29)	0.001317
rs10808565	129.076594	1.4452	TT	84 (90)	59 (69)		
			TC	8 (9)	22 (26)		
			CC	1 (1)	4 (5)	3.34 (1.59–7.02)	0.001442
rs3815871	129.07776	1.4452	GG	87 (84)	66 (64)		
			GC	15 (15)	32 (31)		
			CC	1 (1)	5 (5)	2.75 (1.50–5.03)	0.001093
rs13447075	129.079772	1.4505	CC	91 (88)	67 (70)		
			CA	12 (12)	26 (27)		
			AA	1 (1)	3 (3)	2.67 (1.37–5.19)	0.003774
rs10087240	129.081756	1.4655	CC	85 (85)	58 (69)		
			CT	15 (15)	24 (29)		
			TT	0 (0)	2 (2)	2.57 (1.29–5.11)	0.00721
rs2720709	129.127538	2.4432	AA	30 (29)	9 (9)		
			AG	54 (51)	50 (49)		
			GG	21 (20)	44 (43)	2.57 (1.66–3.96)	0.000021
rs2720659	129.129986	2.4432	AA	56 (62)	26 (30)		
			AG	28 (31)	40 (45)		
			GG	7 (8)	22 (25)	2.73 (1.73–4.30)	0.000015
rs2720660	129.130424	2.4455	GG	72 (70)	44 (44)		
			GA	27 (26)	43 (43)		
			AA	4 (4)	14 (14)	2.49 (1.56–3.98)	0.000124
NA	129.130967	2.4477	CC	68 (68)	88 (85)		
			CA	30 (30)	15 (15)		
			AA	2 (2)	0 (0)	0.36 (0.19–0.71)	0.002845
rs2720662	129.132203	2.4477	TT	62 (60)	28 (27)		
			TC	34 (33)	53 (51)		
			CC	7 (7)	22 (21)	2.89 (1.85–4.51)	0.000003
rs1499373	129.133847	2.4600	CC	21 (21)	44 (44)		
			CG	50 (50)	46 (46)		
			GG	30 (30)	10 (10)	0.41 (0.26–0.62)	0.000036
rs2648875	129.141343	2.4600	AA	63 (61)	28 (27)		
			AG	33 (32)	54 (52)		
			GG	7 (7)	22 (21)	2.97 (1.90–4.65)	0.000002
rs2648876	129.142148	2.4603	GG	71 (76)	45 (54)		
			GA	19 (20)	30 (36)		
			AA	4 (4)	9 (11)	2.14 (1.29–3.55)	0.003225
rs2250888	129.15156	2.4603	TT	41 (41)	20 (21)		
			TC	46 (46)	55 (57)		
			CC	12 (12)	22 (23)	2.01 (1.30–3.11)	0.001623
rs2720666	129.152641	2.7158	AA	87 (84)	70 (70)		
			AG	16 (16)	27 (27)		
			GG	0 (0)	3 (3)	2.38 (1.25–4.53)	0.008602
rs2720667	129.152766	2.7493	AA	87 (84)	72 (69)		
			AG	16 (16)	29 (28)		
			GG	0 (0)	3 (3)	2.45 (1.29–4.66)	0.006183
rs1499368	129.163771	2.7493	TT	40 (38)	21 (20)		
			TC	52 (50)	58 (56)		
			CC	12 (12)	25 (24)	2.01 (1.31–3.08)	0.00131
rs1499367	129.16416	2.7493	AA	40 (39)	21 (21)		
			AG	50 (49)	57 (56)		
			GG	12 (12)	24 (24)	1.98 (1.29–3.04)	0.001705
rs3931283	129.179915	3.1601	CC	75 (75)	50 (52)		
			CT	24 (24)	38 (39)		
			TT	1 (1)	9 (9)	2.70 (1.59–4.59)	0.000252
rs4526320	129.182269	3.3900	CC	82 (80)	62 (62)		
			CG	20 (19)	32 (32)		
			GG	1 (1)	6 (6)	2.29 (1.31–4.00)	0.003499
rs4733595	129.186837	3.6957	GG	82 (79)	62 (60)		
			GA	21 (20)	36 (35)		
			AA	1 (1)	6 (6)	2.40 (1.38–4.16)	0.001827

Continued on following page

TABLE 1
Continued

Marker	Position (Mb)	Position (LDU)	Genotype	*n* (%) case subjects	*n* (%) control subjects	OR (95% CI)	*P*
rs2608030	129.234261	4.4565	CC	46 (45)	70 (67)		
			CT	49 (48)	28 (27)		
			TT	8 (8)	6 (6)	0.52 (0.32–0.82)	0.004991
rs7465157	129.324588	4.8513	GG	82 (79)	62 (60)		
			GA	22 (21)	38 (37)		
			AA	0 (0)	4 (4)	2.57 (1.44–4.59)	0.001414

Genotypes for each marker were assessed in 105 diabetic individuals with ESRD and 102 diabetic control subjects. Chromosomal position is based on Build 35.1. The number of individuals (*n*) per genotype is shown with the frequency (%) for each group. ORs shown were calculated using an analytical model that assumed an additive allele effect on the logarithm of the odds and is expressed per copy of the allele listed first.

with the strongest association was highly concordant with rs2648875 ($r^2 = 1.00$), which suggests that most of the information regarding association is contained within the genotypes of this individual SNP.

Because rs2648875 showed the strongest evidence for association with ESRD, conditional analyses were conducted for this SNP paired with each of the other 20 SNPs that were associated with $P < 0.01$ in single-marker analyses, excluding rs2720662 and rs2720659, which were too strongly concordant with rs2648875 for estimation of conditional association. When stratified by genotypes at each of these other SNPs, genotypes at rs2648875 remained associated with ESRD ($P < 0.005$ for each); however, none of the other SNPs were associated with ESRD when stratified by genotypes at rs2648875 (all $P > 0.10$). Results for the analysis of rs2648875 and rs2720709, the two most strongly associated nonredundant SNPs, are shown in Table 2. Within each genotype at rs2720709, there was a higher frequency of the A allele at rs2648875 in case subjects with ESRD than in control subjects (OR 2.26 per copy of the A allele controlled for genotypes at rs2720709 [95% CI 1.29–3.96], $P = 0.004$). There was also a tendency for case subjects to have an increased frequency of the A allele at rs2720709 within each genotype at rs2648875, but the association was not statistically significant (1.54 per copy of the A allele controlled for genotypes at rs2648875 [0.88–2.71], $P = 0.133$). These analyses suggest that the associations observed between diabetic ESRD and SNPs on 8q24.21 reflect in large part the association with alleles at rs2648875, or alleles at markers in strong LD with rs2648875.

DISCUSSION

Studies of familial aggregation and segregation analyses suggest the potential importance of genetic factors in the development of diabetic nephropathy in Pima Indians (3,10) and other populations (4–6), but the identities of specific susceptibility genes remain largely unknown. The present study strongly implicates a region near *PVT1* on chromosome 8q, or perhaps *PVT1* itself, as an ESRD susceptibility locus. While a previous genome-wide linkage study of diabetic nephropathy in the Pima Indians using nonparametric methods did not identify linkage in this region (24), a subsequent analysis using model-based methods found modest evidence for linkage (logarithm of odds 1.1) on 8q, ~7 Mb from *PVT1* (36). It is therefore possible that this result reflects effects of the locus detected in the present study.

Markers showing the strongest evidence for association with ESRD were located within the *PVT1* gene. While few studies have focused on the investigation of this gene or its protein product, *PVT1* is known to participate (2,8) in translocations found in some human Burkitt lymphomas (37). Further, *PVT1* is co-amplified with the transcription factor *MYC* and plays a role in cell cycle progression, apoptosis, and cellular transformation (38). Because disregulated cell growth, particularly mesangial cell expansion, is a hallmark of diabetic kidney disease, it is possible that *PVT1* impacts on this process by attenuating restraints controlling cell division. Importantly, PVT1 is expressed at high levels in the kidney (39), although its role in that tissue is not yet known.

Many of the markers in the present study have associations with ESRD that are very strong by conventional criteria, with some *P* values $<10^{-5}$. However, in genetic association studies, there is generally a low prior probability that any given marker is associated with disease, and multiple statistical tests are potentially conducted, particularly in GWA studies. For these reasons, very stringent thresholds are generally advocated for declaring statistical significance (11,40,41). While the appropriate thresholds continue to be debated, the *P* value of 2×10^{-6} observed with rs2648875 suggests that association is likely to be

TABLE 2
Association of genotypes at rs2648875 and rs2720709 with ESRD

rs2648875	rs2720709	*n* (%) case subjects	*n* (%) control subjects	OR (95% CI)
GG	GG	7 (7)	22 (21)	1.00 (reference)
AG	GG	9 (9)	19 (18)	1.49 (0.47–4.76)
AA	GG	5 (5)	3 (3)	5.24 (0.99–27.7)
AG	AG	24 (23)	34 (33)	2.21 (0.82–6.02)
AA	AG	28 (27)	16 (16)	5.50 (1.92–15.7)
AA	AA	30 (29)	9 (9)	10.5 (3.38–32.5)

OR represents the odds for ESRD associated with each genotype combination compared with the odds for those in the reference category (GG at rs2648875 and GG at rs2720709).

reproduced (41). However, these proposed thresholds require certain assumptions about unknown quantities, such as the prior probability of a true association of a given effect; thus, the present findings require empirical confirmation in additional groups of individuals to firmly establish these variants as markers for susceptibility to diabetic ESRD. Further experimental work is also necessary to examine the potential functionality of these alleles or other alleles in strong LD with these markers.

Although these findings present some of the strongest evidence for a nephropathy locus published to date, several issues should be considered when interpreting these data. It is widely recognized that association methods can be much more powerful than linkage studies, provided that markers strongly concordant with functional alleles are genotyped. However, in the present study genetic variation was not exhaustively captured. Based on data obtained using the HapMap resource, 30–40% of common variants have $r^2 > 0.80$ with a marker on this array in non-African populations (42); although they are not represented in the HapMap, surveys of LD in American Indian populations suggest that they are similar to other non-Africans in this respect (43). Thus, it is possible that additional variants with important effects on disease susceptibility were not identified in this study due to incomplete marker coverage. In addition, association studies require adequate statistical power to detect genetic loci with strong effects on disease susceptibility. In this study, statistical power was limited by the available sample size. However, because the present study sample was based upon selection of extremely discordant individuals, the power to detect genetic determinants of ESRD is expected to be augmented. For example, control subjects were selected based on a long duration of diabetes and no evidence of heavy proteinuria; as such, these individuals are presumably resistant to developing diabetic nephropathy. Based on the population incidence of ESRD and heavy proteinuria, we estimate that case subjects are derived from the upper 15% of the liability distribution for diabetic nephropathy and that control subjects are derived from the lower 40% (26). With these assumptions, we calculate that the power to detect an association at $P < 0.001$ with an allele accounting for 5% of the variance in liability (OR 2.3 for an allele with frequency 0.5) is 80% (44). Therefore, we expect that the sample size used here is adequate for detecting loci with major effects, such as the putative one on 8q.

Finally, beyond a few published studies (15–17,19,21, 22), the utility of pooling-based approaches using SNP microarrays remains relatively unknown. Imprecision in estimates of allele frequency in pools and lack of reproducibility are important concerns with this approach. With >115,000 SNPs and allelic frequency differences that encompass a range of 0–20%, even a 1% error in measurement of differences in allelic frequency between pools can alter SNP ranking considerably, leading to falsely positive or negative results. However, identifying several neighboring SNPs that are highly ranked within a specific window can, in some cases, reduce the potential of focusing on artifactual findings, albeit at the cost of possibly missing signals in isolated SNPs in low LD with surrounding markers. In contrast, the potential for false negatives, or markers with strongly significant associations in individual genotyping not detected in the pooled analyses, is more difficult to assess without extensive individual genotyping. Thus, while the present study has identified markers on 8q24.21 showing strong associations with diabetic ESRD, we recognize that there may be other markers with equally strong or stronger associations among those available on the array that remain undetected. Other regions that were highly ranked in the pooled study (e.g., chromosomes 3, 10, and 12) may also have markers with equally strong or stronger associations, and we are currently investigating these regions.

In summary, this study supports the use of pooling-based approaches for GWA studies and provides the first evidence supporting a potential locus for ESRD in diabetes within the *PVT1* gene. Replication of these results in other populations, as well as identification of potential functional variants for further characterization, will help to clarify the role of *PVT1* in the development of ESRD in diabetes.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award from the American Diabetes Association (to J.K.W.) and by the intramural research program of the National Institute of Diabetes and Digestive and Kidney Diseases.

We thank the members of the Gila River Indian Community for their continued participation in studies of diabetes and its complications.

REFERENCES

1. U.S. Renal Data System, USRDS Annual Data Report: *Atlas of End-Stage Renal Disease in the United States*. Bethesda, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2002, p. 296–322
2. Nelson RG, Newman JM, Knowler WC, Sievers ML, Kunzelman CL, Pettitt DJ, Moffett CD, Teutsch SM, Bennett PH: Incidence of end-stage renal disease in type 2 (non-insulin-dependent) diabetes mellitus in Pima Indians. *Diabetologia* 31:730–736, 1988
3. Pettitt DJ, Saad MF, Bennett PH, Nelson RG, Knowler WC: Familial predisposition to renal disease in two generations of Pima Indians with type 2 (non-insulin-dependent) diabetes mellitus. *Diabetologia* 33:438–443, 1990
4. Seaquist ER, Goetz FC, Rich S, Barbosa J: Familial clustering of diabetic kidney disease: evidence for genetic susceptibility to diabetic nephropathy. *N Engl J Med* 320:1161–1165, 1989
5. Quinn M, Angelico MC, Warram JH, Krolewski AS: Familial factors determine the development of diabetic nephropathy in patients with IDDM. *Diabetologia* 39:940–945, 1996
6. Freedman BI, Tuttle AB, Spray BJ: Familial predisposition to nephropathy in African-Americans with non-insulin-dependent diabetes mellitus. *Am J Kidney Dis* 25:710–713, 1995
7. Spray BJ, Atassi NG, Tuttle AB, Freedman BI: Familial risk, age at onset, and cause of end-stage renal disease in white Americans. *J Am Soc Nephrol* 5:1806–1810, 1995
8. Canani LH, Gerchman F, Gross JL: Familial clustering of diabetic nephropathy in Brazilian type 2 diabetic patients. *Diabetes* 48:909–913, 1999
9. Faronato PP, Maioli M, Tonolo G, Brocco E, Noventa F, Piarulli F, Abaterusso C, Modena F, de Bigontina G, Velussi M, Inchiostro S, Santeusanio F, Bueti A, Nosadini R: Clustering of albumin excretion rate abnormalities in Caucasian patients with NIDDM: the Italian NIDDM Nephropathy Study Group. *Diabetologia* 40:816–823, 1997
10. Imperatore G, Knowler WC, Pettitt DJ, Kobes S, Bennett PH, Hanson RL: Segregation analysis of diabetic nephropathy in Pima Indians. *Diabetes* 49:1049–1056, 2000
11. Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 273:1516–1517, 1996
12. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA: Mapping complex disease loci in whole-genome association studies. *Nature* 429:446–452, 2004
13. Wang WY, Barratt BJ, Clayton DG, Todd JA: Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6:109–118, 2005
14. Brohede J, Dunne R, McKay JD, Hannan GN: PPC: an algorithm for accurate estimation of SNP allele frequencies in small equimolar pools of

DNA using data from high density microarrays. *Nucleic Acid Res* 33:e142, 2005
15. Meaburn E, Butcher LM, Schalkwyk LC, Plomin R: Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. *Nucleic Acid Res* 34:e27, 2006
16. Meaburn E, Butcher LM, Liu L, Fernandes C, Hansen V, Al-Chalabi A, Plomin R, Craig I, Schalkwyk LC: Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics* 6:52, 2005
17. Craig I, Meaburn E, Butcher L, Hill L, Plomin R: Single-nucleotide polymorphism genotyping in DNA pools. *Methods Mol Biol* 311:147–164, 2005
18. Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'Donovan MC: Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics* 7:27, 2006
19. Craig I, Plomin R: Quantitative trait loci for IQ and other complex traits: single-nucleotide polymorphism genotyping using pooled DNA and microarrays. *Genes Brain Behav* 5 (Suppl. 1):32–37, 2006
20. Liu QR, Drgon T, Walther D, Johnson C, Poleskaya O, Hess J, Uhl GR: Pooled association genome scanning: validation and use to identify addiction vulnerability loci in two samples. *Proc Natl Acad Sci U S A* 102:11864–11869, 2005
21. Butcher LM, Meaburn E, Dale PS, Sham P, Schalkwyk LC, Craig IW, Plomin R: Association analysis of mild mental impairment using DNA pooling to screen 432 brain-expressed single-nucleotide polymorphisms. *Mol Psychiatry* 10:384–392, 2005
22. Butcher LM, Meaburn E, Knight J, Sham PC, Schalkwyk LC, Craig IW, Plomin R: SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children. *Hum Mol Genet* 14:1315–1325, 2005
23. Knowler WC, Bennett PH, Hamman RF, Miller M: Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. *Am J Epidemiol* 108:497–505, 1978
24. Imperatore G, Hanson RL, Pettitt DJ, Kobes S, Bennett PH, Knowler WC: Sib-pair linkage analysis for susceptibility genes for microvascular complications among Pima Indians with type 2 diabetes: Pima Diabetes Genes Group. *Diabetes* 47:821–830, 1998
25. Hanson RL, Ehm MG, Pettitt DJ, Prochazka M, Thompson DB, Timberlake D, Foroud T, Kobes S, Baier L, Burns DK, Almasy L, Blangero J, Garvey WT, Bennett PH, Knowler WC: An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *Am J Hum Genet* 63:1130–1138, 1998
26. Wolford JK, Yeatts KA, Red Eagle AR, Nelson RG, Knowler WC, Hanson RL: Variants in the gene encoding aldose reductase (AKR1B1) and diabetic nephropathy in American Indians. *Diabet Med* 23:367–376, 2006
27. Lovmar L, Ahlford A, Jonsson M, Syvanen AC: Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics* 6:35, 2005
28. Konheim YL, Wolford JK: Association of a promoter variant in the inducible cyclooxygenase-2 gene (PTGS2) with type 2 diabetes mellitus in Pima Indians. *Hum Genet* 113:377–381, 2003
29. Xie XO, Ott J: Testing linkage disequilibrium between a disease gene and marker loci. *Am J Hum Genet* 53:1107, 1993
30. Zondervan KT, Cardon LR: The complex interplay among factors that influence allelic association. *Nat Rev Genet* 5:89–100, 2004
31. Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, Morton NE: The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci U S A* 99:2228–2233, 2002
32. MacQueen J: Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA, University of California Press, 1965, p. 281–297
33. O'Connell JR: Zero-recombinant haplotyping: applications to fine mapping using SNPs. *Genet Epidemiol* 19 (Suppl. 1):S64–S70, 2000
34. Lathrop GM, Lalouel JM: Easy calculations of lod scores and genetic risks on small computers. *Am J Hum Genet* 36:460–465, 1984
35. Mantel N: Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *J Am Stat Assoc* 58:690–700, 1963
36. Imperatore G, Knowler WC, Nelson RG, Hanson RL: Genetics of diabetic nephropathy in the Pima Indians. *Curr Diab Rep* 1:275–281, 2001
37. Graham M, Adams JM: Chromosome 8 breakpoint far 3′ of the c-myc oncogene in a Burkitt's lymphoma 2;8 variant translocation is equivalent to the murine pvt-1 locus. *Embo J* 5:2845–2851, 1986
38. Shtivelman E, Bishop JM: The PVT gene frequently amplifies with MYC in tumor cells. *Mol Cell Biol* 9:1148–1154, 1989
39. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* 99:4465–4470, 2002
40. Colhoun HM, McKeigue PM, Davey Smith G: Problems of reporting genetic associations with complex outcomes. *Lancet* 361:865–872, 2003
41. Manly KF: Reliability of statistical associations between genes and disease. *Immunogenetics* 57:549–558, 2005
42. Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ: Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* 38:663–667, 2006
43. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK: A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 38:1251–1260, 2006
44. Hanson RL, Looker HC, Ma L, Muller YL, Baier LJ, Knowler WC: Design and analysis of genetic association studies to finely map a locus identified by linkage analysis: sample size and power calculations. *Ann Intern Med* 70:332–349, 2006