European Journal of Human Genetics (2006) 14, 236–244

www.nature.com/ejhg

ARTICLE

Hidden population substructures in an apparently homogeneous population bias association studies

Mario Berger[1], Hans H Stassen[2], Karola Köhler[3], Vera Krane[1], Detlev Mönks[1], Christoph Wanner[1], Katrin Hoffmann[4], Michael M Hoffmann[5], Michael Zimmer[6], Heike Bickeböller[3], and Tom H Lindner*[,1,6,7]

[1]*Division of Nephrology, Department of Medicine, University of Würzburg, Würzburg, Germany;* [2]*Department of Psychiatry, University of Zurich, Zurich, Switzerland;* [3]*Department of Genetic Epidemiology, University of Göttingen, Göttingen, Germany;* [4]*Institute of Medical Genetics, Charité, Humboldt University Berlin, Berlin, Germany;* [5]*Division of Clinical Chemistry, Department of Medicine, Albert Ludwigs-University Freiburg, Freiburg, Germany;* [6]*Department of Clinical Biochemistry and Pathobiochemistry, University of Würzburg, Würzburg, Germany;* [7]*Department of Kidney Diseases and Hypertension, Medical Clinic IV, Hospital Nürnberg Süd, University of Erlangen-Nürnberg, Nürnberg, Germany*

Linkage- and association-based approaches have been applied to attempt to unravel the genetic predisposition for complex diseases. However, studies often report contradictory results even when similar population backgrounds are investigated. Unrecognized population substructures could possibly explain these inconsistencies. In an apparently homogeneous German sample of 612 patients with type 2 diabetic and end-stage diabetic nephropathy and 214 healthy controls, we tested for hidden population substructures and their possible effects on association. Using a genetic vector space analysis of genotypes of 20 microsatellite markers, we identified four distinct subsets of cases and controls. The significance of these substructures was demonstrated by subsequent association analyses, using three genetic markers (UCSNP-43,-19,-63; intron 3 of the calpain-10 gene). In the undivided sample, we found no association between individual SNPs or any haplogenotypes (ie the genotype combination of two multilocus haplotypes) and type 2 diabetes. In contrast, when analyzing the four groups separately, we found that there was evidence for association of the common C allele of UCSNP-63 with the trait in the largest group ($n = 547$ cases/101 controls; $P = 0.002$). In this subset haplotype 112 was more frequent in controls than in cases ($P = 0.006$; haplogenotype 112/121: odds ratio (OR) = 0.27, 95% confidence intervals (CI) = 0.13–0.57), indicating a protective effect against the development of type 2 diabetes. Our study demonstrates that unconsidered population substructures (ethnicity-dependent factors) can severely bias association studies.

European Journal of Human Genetics (2006) **14,** 236–244. doi:10.1038/sj.ejhg.5201546; published online 7 December 2005

Keywords: population substructure analysis; association study; type 2 diabetes mellitus; calpain-10; haplogenotypes; end-stage diabetic nephropathy

***Correspondence:** Dr TH Lindner, University of Erlangen-Nürnberg, Medical Clinic IV, Department of Kidney Diseases and Hypertension, Hospital Nürnberg Süd, Breslauer Str. 201, 90471 Nürnberg, Germany.
Tel: +49 911 9808140; Fax: +49 911 9808170;
E-mail: t-lindner@gmx.de
Received 12 May 2005; revised 15 September 2005; accepted 21 October 2005; published online 7 December 2005

Introduction

The phenotype of a complex disease such as diabetes or hypertension is the result of interactions between genetic and nongenetic (environmental) factors such as diet and physical activity. Genetically influenced factors differ in various populations, a fact which is partially due to natural

selection (which favors adaptations specific to the environment) but also helps to explain why complex traits occur with variable prevalences and varying subphenotypes in different regions of the world.

Both genetic and nongenetic factors can severely confound the results of any genetic study on a complex trait. Although advanced strategies on the phenotype level have been developed over the past decade, little attention has been paid to population stratification and genetic homogeneity of the study sample. In general, there are only a few reports in which researchers tried to address this issue. It has been argued that the effects of stratification can be eliminated simply by carefully matching cases and controls according to self-reported ancestry and geographical origin.[1] The argument was supported by studies using empirical methods such as STRUCTURE to detect stratification based on genotypes at unlinked markers.[2,3] In one of four studies involving genotyping of 44 unlinked markers, stratification was detected, but the signal was no longer apparent after more stringent matching of cases and controls based on the birthplaces of the individuals' grandparents.[4] This has been interpreted as evidence that stratification may be less of a concern than originally anticipated.

Although systematic differences in the ancestry of cases and controls can be a source of false-positive associations,[5,6] the fraction of published associations that is attributable to stratification is unknown.[7] Freedman *et al*[8] found no significant evidence for stratification with STRUCTURE by analyzing data from 24–48 SNPs in 11 association studies spanning a range of disease states and self-reported ancestries and three different epidemiological designs. However, after typing more SNPs and applying the method of Genomic Control to the data they found significant evidence for stratification ($P<0.0001$).[8] Even in the relatively homogeneous genetic isolate from Iceland, Helgason *et al*[9] found evidence for substructures, indicating that sampling strategies need to take account of this issue.

The search for type 2 diabetes (T2DM) associated SNPs provides a sobering example of contradictory association results for a complex trait where undetected population substructures might be responsible for the discrepancy. In an association-based follow-up of a genomewide linkage scan, Horikawa *et al*[10,11] identified three genetic markers in intron 3 of the calpain-10 gene that significantly contributed to diabetes susceptibility in a Mexican-American population. Haplogenotype 112/121 (UCSNP-43, -19, -63) defined the highest risk (original sample: odds ratio (OR) = 2.80, 95% confidence intervals (CI) = 1.23–6.34; replication sample: OR = 3.58, 95%CI = 1.43–8.92). The largest study to date, a meta-analysis by Weedon *et al*,[12] demonstrated a role for calpain-10 in T2DM susceptibility ($P = 0.0007$; OR = 1.17; 95%CI: 1.07–1.29). However, most studies that tried to replicate the findings have failed to show any association (Table 1).

Table 1 Tests for association of UCSNP-44 (if performed), -43, -19, -63 in intron 3 of the calpain-10 gene with T2DM in different populations

Population (reference)	*Cases/controls*	*UCSNP*				*Haplogenotype*
		-44	*-43*	*-19*	*-63*	
Afroamericans[39]	269/1159	NS	NS	NS	NS	NS
Botnia fins[40]	395/298	NS	0.011 (G)	NS	0.010 (T)	0.028 (1121/1121)
British[38]	153/411[1a]	NS[a]	NS	NS	NS	NS
	222/212[1b]	NS[a]	NS	NS	NS	NS
	49/49[1c]	NS[a]	NS	NS	NS	NS
Chinese[41]	173/156	NS	NS	NS	NS	NS
Chinese[42]	211/127	ND	0.011 (G)	NS	ND	ND
Fins[43]	1603 total[b]	ND	NS	ND	NS	NS
Japanese[44]	81/81	NS	NS	ND	ND	ND
Japanese[45]	177/172	ND	NS	NS	NS	NS
Mexican[46]	134/114	0.017 (C)	NS	NS	NS	NS
Oji-Cree[47]	121/468	ND	NS	ND	ND	ND
Polish[48]	229/148	ND	NS	NS	NS	0.038 (121/121)
Samoans[49]	172/96	ND	NS	NS	NS	NS
Scandinavians[50]	1159 total[c]	NS	NS	NS	NS	NS

ND = not determined; NS = not significant. In case of significant results, the *P*-value, and the associated allele/haplogenotype are shown in round brackets.

[a]In a combined analysis of all UK. studies (1. Case–control study[1a]: trios probands/controls from birth cohort.[36] 2. Case–control study[1b]: diabetic probands, adult controls.[37] 3. Discordant-sib study[1c]: diabetic subjects, unaffected sibs.[38]), and in combination with a Mexican-American study,[11] the C allele at SNP-44 was associated with T2DM ($P = 0.015$, $P = 0.004$).

[b]Two samples of 526 (FUSION 1) and 255 (FUSION 2) index cases, 185 and 414 unaffected spouses and offspring of FUSION 1 index cases or their affected siblings, and 223 elderly normal glucose-tolerant control subjects were tested.

[c]In all, 409 type 2 diabetic patients, 200 glucose-tolerant control subjects, 322 young healthy subjects, 206 glucose-tolerant offspring of diabetic patients, 457 glucose-tolerant 70-year-old men (the participants underwent a 120-min euglycemic-hyperinsulinemic clamp study at 70 years of age for measurements of insulin sensitivity) were tested.

It is likely that those conflicting results were due to a strong genetic and phenotypic heterogeneity of T2DM. The existence of subtle subphenotypes with a different genetic background can be assumed. Selection pressure under different environmental conditions might result in promoting the 'survival' of different mutations in various genes, all leading to comparable but slightly different subphenotypes. As shown by Baier *et al*,[13] 'the diabetes genotype' may have been beneficial for survival during the evolution of man, therefore implicating a variety of such subphenotypes.

In our association study involving previously mentioned SNPs at the calpain-10 locus we therefore wanted to systematically test for genetic stratification of our sample and, if found, perform separate association tests for each subgroup. In order to minimize heterogeneity further we applied a 2-step procedure. (1) We chose end-stage diabetic nephropathy requiring hemodialysis as a very specific diabetes subphenotype. The specificity relies on the fact that only 25–40% of type 2 diabetic patients develop this form of nephropathy after 25 years of diabetes duration. The clientele is usually characterized by severe insulin resistance and frequent micro- and macrovascular complications. Life expectancy after starting hemodialysis is only 2 years for 50% of the patients, based on fatal vascular complications.[14–16] (2) We extended this classical and widely used approach by a genetic diversity test in our cases and controls (German ancestry) before we carried out association analyses. By selecting a specific subphenotype and, in addition, narrowing down the phenotype by a genetic diversity test we were hoping to enrich the sample with those phenotype(s) in subgroups that share nearly the same genetic basis.

We chose a multivariate approach, the genetic vector space method (modified Genomic Control method) that relies on the concept of 'biological ethnicity' (see Materials and methods). We identified significant genetic diversity in an apparently homogeneous German population. When testing calpain-10 SNPs (UCSNP-19, -43, -63) for association with T2D, our results changed completely when taking this genetic diversity into account.

Materials and methods

Subjects

We recruited our case sample of 612 type 2 diabetic patients with end-stage diabetic nephropathy on hemodialysis throughout Germany from dialyses centers within the study frame of the German 4D (Die Deutsche Diabetes-Dialysestudie) trial, with the headquarter at the Division of Nephrology, Department of Medicine, University of Würzburg, Würzburg, Germany.[17] The study was approved by the local ethics committee. All participants gave their written consent. Only patients with end-stage diabetic nephropathy and German ancestry (questionnaire) were included in the trial. Baseline parameters are summarized in Table 2.

Table 2 Clinical and laboratory characteristics of the patients

Parameter	*Value*
Age (years)	65.8±8.1
N	612
Sex (males/females)	328/284
BMI (kg/m^2)	27.5±4.8
Duration of diabetes (years)	17.5±8.5
Total cholesterol (mM)	5.69±1.49
Triglyceride (mM)	3.10±2.56
Coronary artery disease (%)[a]	21.6
Stroke (%)	16.8
HbA_{1c} (%)	6.6±1.3

Data are given as mean±SD.
[a]Combination of myocardial infarction, bypass surgery, PTCA.

We found no history of diabetes, stroke, or myocardial infarction in our control sample of 214 healthy controls (112 males, 102 females). Subjects came from the area around Würzburg, Bavaria, Germany. We found hypertension in two (0.93%), hyperlipoproteinemia in two (0.93%), and smoking in 42 subjects (19.63%). Our questionnaire revealed that 8.41% of parents, 0.47% of siblings, and 0% of children of controls were diabetic. Since the average age of the control group was 33.05±9.32, we had to assume that approximately 5% of the controls were still at risk for developing T2DM (according to general population prevalence of T2DM in Germany).

Genotyping

We genotyped 20 microsatellite markers in DNA samples from all subjects. The selection of those markers was based on the preferences for the genetic vector space method by Stassen *et al.*[19] that will be described later. We used this marker set to detect unknown population stratification through the concept of biological ethnicity (Table 3). Polymerase chain reaction (PCR) was performed under standard reaction conditions. We redesigned primer sequences for UCSNP-43 and -63 (UCSNP-43 forward 5′-HEX-GACCCTCACCATGAGTCATAATTG-3′, UCSNP-43 reverse 5′-TCACCAAGTACAAGGCTTAGCCTCACCTTCGTA-3′, UCSNP-63 forward 5′-FAM-CTCCTGATCAACACCTAGCCAAGG-3′, UCSNP-63 reverse 5′-AAGGGGGGCCAGCGCCTGACGGGGGTGGCG-3′). We performed SNP testing as a modified RFLP method (DRMP-PCR) as described in Berger *et al.*[18]

Genetic diversity test by multivariate feature vectors (population substructure analysis)

The main principles of this method are described in detail by Stassen *et al.*[19] All algorithms were implemented in the *Master.GEN* program package.[20] Here, we present only the basic aspects of the method.

Table 3 Twenty uncorrelated polymorphic markers were combined to a multidimensional feature vector in order to assess genetic diversity and to model biological ethnicity in terms of interindividual genetic similarities

No.	*Marker*	*Chromosome*	*Marker type*	*Decode (cM)*
1	D2S1360	2	Tetra	40.4
2	D4S2632	4	Tetra	54.2
3	D6S1006	6	Tri	30.2
4	D6S1050	6	Tetra	46.0
5	D6S1036	6	Tetra	86.9
6	D6S1082	6	Tetra	99.8
7	D6S474	6	Tetra	116.7
8	D6S1009	6	Tetra	138.8
9	D6S2436	6	Tetra	161.3
10	D6S305	6	Di	173.5
11	D7S1804	7	Tetra	135.9
12	D7S2195	7	Tetra	152.1
13	D11S1999	11	Tetra	17.0
14	D11S1977	11	Tetra	44.0
15	D11S2002	11	Tetra	87.2
16	D11S4464	11	Tetra	130.4
17	D13S788	13	Tetra	54.8
18	D15S822	15	Tetra	13.6
19	D17S928	17	Di	135.7
20	D20S470	20	Tetra	44.3

Unknown population admixture can substantially reduce the power of studies that aim to link phenotype to genotype. Since allele distributions of microsatellites generally display subtle to marked differences between populations, a multivariate configuration of sufficiently polymorphic microsatellites enables quantification of the genetic heterogeneity of genetically diverse sample sets. Such multivariate configurations can be regarded as multivariate 'feature vectors' which span a genetic vector space. Subjects are characterized in a vector space as distinct 'points' such that genetically similar subjects form compact clouds ('clusters'), while genetically dissimilar subjects are located in more distant regions of the vector space. 'Natural' clusters can then be used to define genetically homogenous subgroups, thus leading to the concept of 'biological ethnicity'. This concept reduces the problem of genomic control for genetic association studies, where unknown population admixture can produce false-positive as well as false-negative signals.

In our study, we relied on a slightly modified set of 20 di-, tri- and tetranucleotide polymorphisms (Table 3), which had previously been applied successfully in studies investigating differences in genetic diversity between various US-American populations,[21–23] European populations,[24,25] and population isolates.[26,27] Those markers were unlinked with each other, albeit not randomly distributed over the genome. The method was initially developed for microsatellite markers, for which many reference genotypes and allele frequencies were available. Once the genetic vector space was constructed, cluster analysis was carried out under the following optimization criteria: (1) cluster detection started exclusively with the cases and searched for the largest homogenous group among the cases, thereby excluding the controls; (2) the controls were subsequently treated as independent replication samples, thus supplementing the clusters derived from the patients. As a direct consequence, our cluster analysis method has a slight preference for actually classified cases over controls.

Evaluation of linkage disequilibrium between UCSNP-43, -19, -63

We estimated haplotype frequencies at three loci separately for the total case and control groups using SNPHAP v1.0. In addition, we carried out haplotype estimation separately for each subgroup within cases and controls for the stratified association analyses. We estimated the LD for each SNP pair separately for cases and controls. Further, we determined Lewontin's D′ and the squared correlation coefficient Δ^2 as measures of linkage disequilibrium (LD), and the *P*-value according to the χ^2 test.[28] We used the statistical package R for testing HWE and for the evaluation of LD.

Analyses of association for UCSNP-43, -19, -63

Allele frequencies for each SNP in controls and cases were computed by allele counting. As in Horikawa *et al*,[11] allele 1 denotes the G allele for UCSNP-43, the C allele at UCSNP-63, and consists of two copies of the 32-bp repetitive sequence at UCSNP-19. We used a significance level of 5% for the initial tests of association. In addition, we evaluated whether an association remained significant after Bonferroni correction for multiple comparisons. Two test statistics were used to test for association between UCSNP-43, -19, -63 and T2DM: (1) the general χ^2 test with two degrees of freedom (df) comparing genotypes of cases and controls, and (2) a trend test with one df comparing genotypes in a multiplicative allelic relative risk model. For the stratified analysis a modified Cochran–Mantel–Haenszel test statistic[29] was applied, which sums up the relative genotype frequency differences for all subgroups without requiring additional df. This method allows the estimation of a common OR across different subgroups. For small *P*-values, ORs and 95%CI were calculated. Exact CI for the ORs are shown for small subgroups. For all association analyses, the statistical package SAS was used.

Haplotype-based analysis of association using SAS

We calculated the estimated number of haplotypes per group for fully genotyped individuals by multiplying the estimated relative frequencies with twice the number of fully genotyped individuals. The rare haplotypes 122, 211, 212, 222 were pooled. We then carried out a global χ^2 test with 4 df comparing the haplotype distribution between cases and controls. For the stratified analysis we applied a modified Cochran–Mantel–Haenszel test statistic as global test statistic, like in the association analyses for individual

SNPs. The *P*-value corresponding to each haplotype shows its importance with respect to the global test statistic. It was calculated by decomposing the global χ^2 test statistic into the contributions of the individual haplotypes. For small *P*-values, we calculated the ORs comparing one haplotype *versus* all other haplotypes. We calculated the exact CI for the ORs for small subgroups. Frequencies for haplogenotypes, containing haplotypes with a significant association to T2DM were calculated as well. Only individuals whose haplogenotype could be determined with a probability of more than 90% were included in this analysis.

Results

Population substructure analyses

The tested sample consisted of 826 subjects (612 cases, 214 controls). Population substructure analyses revealed four subgroups (one large, three small groups) in our sample. In a stratified analysis, we had to exclude 87 controls of subgroup 4 because they did not match any patient group. The sample sizes of the subgroups 2 and 3 were very small so that the results of those groups had to be viewed with caution. The 547 cases and 101 controls of group 1 formed a genetically homogenous population (Table 4). We detected no significant deviation from HWE for SNPs in either the entire sample or any of the identified subgroups of cases and controls. All three markers were in LD as expected.

Analyses of association for UCSNP-43, -19, -63

In the combined sample, we did not find evidence for association between T2DM with end-stage diabetic nephropathy and UCSNP-43/-19. In the nonstratified analysis, UCSNP-63 did not show an association with diabetes either. However, when we analyzed the data stratified by the three subgroups, we found a significant association between UCSNP-63 and T2DM ($P=0.031$) after disregarding the 87 unmatched controls of subgroup 4. The observed association in the entire stratified sample resulted from subgroup 1 ($P=0.002$). In subgroup 1, the rare allele 2 was more frequent in controls than in cases. When we used a Bonferroni correction for the χ^2 test appropriate for testing three independent polymorphisms, that is, a significance level of $\alpha=0.05/3=0.017$, the association tested across all subgroups in the stratified sample failed to reach significance. Since the polymorphisms are in strong LD with each other, this correction probably leads to a very conservative threshold. In subgroup 1, however, the association remained significant even after additional correction for the analysis in three independent subgroups with a Bonferroni-corrected significance level of $\alpha=0.05/9=0.006$ (Table 5).

Table 4 Numbers of cases and controls in the subgroups (number of fully genotyped subjects)

Probands	*Subgroup 1*	*Subgroup 2*	*Subgroup 3*	*Subgroup 4*	*Total*
Cases	547 (531)	39 (39)	26 (26)	0 (0)	612 (596)
Controls	101 (94)	13 (12)	13 (13)	87 (85)	214 (204)
Total	648 (625)	52 (51)	39 (39)	87 (85)	826 (800)

Haplotype-based analysis of association

Four of the eight possible haplotypes were observed at common frequencies in the total data set (111, 112, 121, 221; Table 6). The global χ^2 test in the total nonstratified sample and across subgroups did not reveal any haplotype-based evidence for association. We found a significant global *P*-value across haplotypes ($P=0.035$) only in subgroup 1. As mentioned before, we could not observe an overall association across groups in the stratified analysis since the two other subgroups showed different haplotype distribution patterns between cases and controls. The association in subgroup 1 did not remain significant when adjusting for multiple testing in three independent subgroups, since the *P*-value was larger than the Bonferroni-corrected significance level of $\alpha=0.05/3=0.017$. Further analysis showed that the highest contribution to the global *P*-value in subgroup 1 resulted from haplotype 112 ($P=0.006$), which was the main cause of the observed association for UCSNP-63 described in the previous paragraph. Haplotype 112 was the only haplotype with a frequency higher than 1% in the population having allele 2 at UCSNP-63. In our analysis, haplotype 112 was associated with a lower risk of T2DM because it occurred at higher frequency in the control group. We also compared the frequencies of the different haplogenotypes containing haplotype 112. The analysis showed that the observed association between haplotype 112 and diabetes was mainly caused by individuals carrying the haplotype combination 112/121 (Table 7), which was more often observed in controls *versus* the cases.

Discussion

The formidable problems of detecting association in complex diseases such as T2DM lie in the significant reduction of power that is associated with the etiological complexity (clinical, genetic, ethnic heterogeneity; polygenic character; gene–environmental interactions).[30,31] Since there is no reliable test available to differentiate distinct subforms of diabetes it is not surprising that association studies present with conflicting results. A part of the explanation of this situation might be related to potentially hidden population substructure in addition to very heterogeneous phenotypes. Several authors have shown that sampling strategies need to take account of population substructure, among them the Icelandic genetic isolate.[9]

There is no single study on the role of the described CAPN10 polymorphism in the development of T2DM that

Table 5 Allele frequencies and tests for association between UCSNP -43, -19, -63 and T2DM

	Complete sample		*Subgroup 1*	*Subgroup 2*	*Subgroup 3*
	Not stratified	*Stratified*			
No. of subjects	*n = 612/214*	*n = 612/127*	*n = 547/101*	*n = 39/13*	*n = 26/13*
UCSNP-43					
Cases/controls (allele 1%)	70.7/69.0	70.7/71.0[a]	71.3/71.9	64.1/53.9	67.3/80.8
P value χ^2 (2df)	0.139	0.200	0.245	0.623	0.110
P value trend test (1df)	0.500	0.834	0.861	0.393	0.172
UCSNP-19					
Cases/controls (allele 1%)	38.7/37.4	38.7/37.7[a]	39.2/38.7	32.1/29.2	38.5/38.5
P value χ^2 (2df)	0.209	0.630	0.545	0.578	0.601
P value trend test (1df)	0.650	0.843	0.896	0.768	1.000
UCSNP-63					
Cases/controls (allele 1%)	93.8/91.5	**93.8/90.4**[a]	**94.0/88.4**	93.6/100.0	90.4/96.2
P value χ^2 (2df)	0.158	**0.031**	**0.002**	0.175	0.347
P value trend test (1df)	0.107	**0.047**	**0.005**	0.175	0.347
OR (22 or 12 *versus* 11)	0.67	**0.56**	**0.42**	—[b]	2.86
95%CI for OR	(0.43, 1.04)	**(0.34, 0.93)**	**(0.25, 0.72)**	—[b]	(0.27, 146)

[a]The overall allele frequencies for the stratified sample not including subgroup 4 are only given for comparison and are not used for the test statistic, which is based only on subpopulation genotype frequencies.
[b]Since all subgroup 2 controls were 11 homozygotes, the OR and 95%CI for subgroup 2 could not be estimated.
Significant results are shown in bold.

Table 6 Haplotype frequencies and tests for association between the haplotypes and T2DM

	Complete sample		*Subgroup 1*	*Subgroup 2*	*Subgroup 3*
	Not stratified	*Stratified*			
No. of subjects[a]	*n = 596/204*	*n = 596/119*	*n = 531/94*	*n = 39/12*	*n = 26/13*
Haplotype 111					
Cases/controls (%)	31.7/27.9	31.7/26.9[b]	32.4/27.0	25.6/26.9	24.8/26.1
P-value	NS	—[c]	NS	NS	NS
Haplotype 112					
Cases/controls (%)	6.2/8.2	6.2/9.6[b]	**6.0/11.7**	6.4/0.0	9.6/3.8
P-value	0.180	—[c]	**0.006**	0.125	0.386
OR	0.74	0.62	**0.48**	—[d]	2.66
95%CI	(0.48, 1.13)	(0.38, 1.01)	**(0.29, 0.81)**	—[d]	(0.27,130.7)
Haplotype 121					
Cases/controls (%)	32.7/32.7	32.7/34.3[b]	32.8/32.9	32.1/26.9	32.9/50.8
P-value	NS	—[c]	NS	NS	NS
Haplotype 221					
Cases/controls (%)	28.6/30.1	28.6/28.3[b]	28.0/28.4	35.9/46.2	28.7/10.7
P-value	NS	—[c]	NS	NS	NS
Remaining haplotypes					
Cases/controls (%)	0.8/1.1	0.8/0.9[b]	0.7/0.0	0.0/0.0	4.0/8.5
P-value	NS	—[c]	NS	NS	NS
P-value (global χ^2 test)	0.425	0.302	**0.035**	0.697	0.260

[a]For comparison, the relevant number of subjects is given, although the analysis is actually based on the number of haplotypes which is twice the number of subjects.
[b]Note that overall haplotype frequencies for the stratified sample not including subgroup 4 are only given for comparison and are not used for estimation of the OR, which is based only on subpopulation haplotype frequencies.
[c]The stratified global χ^2 test could not be decomposed into the contributions of the individual haplotypes because the covariance matrix of the test statistic is calculated as the sum over the subpopulations and its inverse could not be decomposed as before.
[d]In subgroup 2 the OR and corresponding 95%CI could not be estimated since there were no controls with haplotype 112.
Significant results are shown in bold.

Table 7 Haplogenotype frequencies for haplogenotypes containing haplotype 112

	Complete sample		*Subgroup 1*	*Subgroup 2*	*Subgroup 3*
	Not stratified	*Stratified*			
No. of subjects	*n = 597/207*	*n = 589/119*	*N = 531/95*	*n = 39/13*	*n = 19/11*
Haplogenotype: 112/111					
Cases/controls (%)	3.4/1.9	3.4/2.5	3.4/3.1	5.1/0.0	0.0/0.0
Haplogenotype: 112/112					
Cases/controls (%)	0.7/0.5	0.7/0.0	0.8/0.0	0.0/0.0	0.0/0.0
Haplogenotype: 112/121					
Cases/controls (%)	4.0/8.7	4.0/10.9	**3.8/12.6**	2.6/0.0	15.8/9.1
OR	0.44	0.38	**0.27**	—	1.88
95%CI for OR	(0.23, 0.83)	(0.19, 0.77)	**(0.13, 0.57)**	—	(0.13,108.1)
Haplogenotype: 112/221					
Cases/controls (%)	3.5/4.8	3.5/5.9	3.2/7.4	5.1/0.0	10.5/0.0

Significant results are shown in bold.

performed a systematic test for genetic diversity in the tested samples from different populations. As in previous studies, our sampling strategy for diabetic nephropathy also relied exclusively on a questionnaire for a 'valid' ancestry of study participants. Although also we had apparently recruited relatively homogeneous German groups of cases and controls, we nevertheless detected three distinct subgroups in our cases and controls, with obviously differing genetic ancestry, by a genetic vector space method with 20 microsatellite markers. There were even 87 controls that had to be removed since they did not match any case from our sample. When we analyzed the entire sample disregarding population substructure, we did not detect association between end-stage diabetic nephropathy requiring hemodialysis and the three individual calpain-10 polymorphisms, including possible haplotypes and haplogenotypes. When we grouped all individuals by their population substructure we found a significant association of the common allele 1 at UCSNP-63 with diabetes ($P = 0.005$) in the largest subgroup 1 (547 cases, 101 controls). Even after a very conservative correction for multiple testing (Bonferroni), the calculated *P*-value remained significant (corrected $\alpha = 0.006$). We were aware of the potential loss of statistical power by subclassification. However, the 'effective' statistical power may have been increased by the use of homogeneous subpopulations. Further, the cross comparison between groups enabled the distinction between 'population-independent' (the same signal shows up in all subpopulations) and 'population-related' (the signal shows up in a compact subpopulation while failing to be detected in the population as a whole) vulnerability factors. This was a generalization of standard genomic control methods that follow a probability-oriented approach in order to test for population stratification.

The direction of the association came somewhat unexpectedly: The rare allele 2 was more frequent in controls than in cases and decreased the risk for the development of T2DM with end-stage diabetic nephropathy. This difference could not be observed in groups 2 and 3. The stratified test statistic also supported the association ($P = 0.031$) but did not remain significant after Bonferroni correction. This might be explained by the low power of the Cochran–Mantel–Haenszel test statistic for detecting an association if the effect is heterogeneous across the subgroups. We found further that the haplogenotype 112/121 was more often observed in controls *versus* cases. Our findings indicated a protective function of haplogenotypes 112/121 against the development of diabetes with end-stage diabetic nephropathy.

Other studies support the functional impact of our results: Shima *et al*[32] found a lower body mass index (BMI) and a lower HbA_{1c} being associated with the haplogenotype 112/121 ($P = 0.016$, $P = 0.008$). The same findings were replicated by Ehrmann*et al*[33] in African-American subjects with a specific T2DM phenotype, the polycystic ovary syndrome (PCOS). In terms of common polygenic T2DM, it makes sense that individuals at risk for the disease demonstrate a higher BMI compared to non-diabetic controls. However, these studies, like our own, were in contradiction to the results of Horikawa *et al* who found the haplotype combination 112/121 as increasing the risk for diabetes in a Mexican-American population. Another British study supports the findings of Horikawa *et al*. Subjects with the 112/121 haplotype combination ($n = 29$) had increased fasting ($P = 0.004$) and 2-h plasma glucose levels ($P = 0.003$) compared with the remaining group of subjects having all other haplogenotypes. The 112/121 haplotype combination was also associated with a marked decrease in the insulin secretory response, adjusted for the level of insulin resistance ($P = 0.002$).[34] Conflicting results in different populations and even within the same population may indicate a different genetic background for

the trait, thus explaining contradictory findings. On the other hand, there is a chance that different subphenotypes with another genetic background were studied. This may have been the case in a meta-analysis reported by Weedon *et al*[12] involving four different Japanese populations. The results ranged from evidence for and against association. There is clearly a problem: A stronger focus on genetic background stratification is required, as supported by our findings.

We found association in a subsample but not in the undivided sample. A possible interpretation is (1) that the primary genetic mechanism under investigation (CAPN10) may be ethnicity-specific rather than ethnicity-independent in terms of 'biological ethnicity', and (2) that this 'biological ethnicity' can be quantified through a set of polymorphic microsatellites as demonstrated, for example, by Di Rienzo *et al*[35] for African, Egyptian and Sardinian populations using only 10 microsatellites.

It could be argued that our study results were just a matter of coincidence. However, even a very conservative Bonferroni correction of the significance level for multiple testing did not change the significance. There was also no deviation from HWE in all subgroups supporting our procedure. Further, it could be argued that we had a younger control group when compared with the cases. However, taking into account that association was still detected even though 5–10% of the control subjects will develop diabetes sometime in the future, one would expect an even greater effect in a well-matched sample regarding age and sex.

One might argue that the use of microsatellite markers for a test such as we performed could limit the applicability of genetic vector space methods. In fact, in the age of SNPs it would be ideal to have hundreds of thousands of SNPs with which we could establish a much finer population (or individual) differentiation – yet at the cost of additional complexity as a relatively large number of SNPs is necessary just to get the same information content inherent in one single microsatellite.

Since there is little chance of distinguishing subtle phenotypic differences, we propose tests for genetic homogeneity in the study sample along with the use of advanced phenotyping strategies. In this way, we might be able to enhance the chances for identifying both genetic and nongenetic factors contributing to the disease. The identification of population substructures – or in other words, the identification of genetically similar clusters of individuals – should sharpen up the results of association studies.

Acknowledgements

THL was supported by Grants from the Deutsche Forschungsgemeinschaft (LiDFG768/3–1/3–3/4–1/4–2). MB and MZ were supported by the Interdisziplinäres Zentrum für Klinische Forschung (IZKF-E9). HHS was supported in part by a Grant of the Swiss National Science Foundation (SNF 32–61578.00) and Philip Morris USA Inc. HB was supported by the German National Genome Research Network (NGFN; GE-S09T01, HK-S09T01). KK was supported also by NGFN-GE-S09T01 and by the interdisciplinary PhD program 'applied statistics and empirical methods'. KH was supported by a Grant from the DFG (SFB 577).

References

1 Wacholder S, Rothman N, Caporaso N: Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev* 2002; **11**: 513–520.
2 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. *Am J Hum Genet* 2000; **67**: 170–181.
3 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.
4 Ardlie KG, Lunetta KL, Seielstad M: Testing for population subdivision and association in four case–control studies. *Am J Hum Genet* 2002; **71**: 304–311.
5 Knowler WC, Williams RC, Pettitt DJ, Steinberg AG: Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 1988; **43**: 520–526.
6 Kittles RA, Chen W, Panguluri RK *et al*: CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification? *Hum Genet* 2002; **110**: 553–560.
7 Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003; **33**: 177–182.
8 Freedman ML, Reich D, Penney KL *et al*: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36**: 388–393.
9 Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K: An Icelandic example of the impact of population structure on association studies. *Nat Genet* 2005; **37**: 90–95.
10 Hanis CL, Boerwinkle E, Chakraborty R *et al*: A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nat Genet* 1996; **13**: 161–166.
11 Horikawa Y, Oda N, Cox NJ *et al*: Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 2000; **26**: 163–175.
12 Weedon MN, Schwarz PE, Horikawa Y *et al*: Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *Am J Hum Genet* 2003; **73**: 1208–1212.
13 Baier LJ, Permana PA, Yang X *et al*: A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels and insulin resistance. *J Clin Invest* 2000; **106**: R69–73.
14 Chantrel F, Enache I, Bouiller M *et al*: Abysmal prognosis of patients with type 2 diabetes entering dialysis. *Nephrol Dial Transplant* 1999; **14**: 129–136.
15 Ritz E: Nephropathy in type 2 diabetes. *J Intern Med* 1999; **245**: 111–126.
16 Lindner TH, Monks D, Wanner C, Berger M: Genetic aspects of diabetic nephropathy. *Kidney Int Suppl* 2003; **84**: S186–S191.
17 Wanner C, Krane V, Ruf G, Marz W, Ritz E: Rationale and design of a trial improving outcome of type 2 diabetics on hemodialysis. Die Deutsche Diabetes Dialyse Studie Investigators. *Kidney Int Suppl* 1999; **71**: S222–226.
18 Berger M, Zschemisch S, Hocher B *et al*: Alternative approach for rapid and reliable single-nucleotide polymorphism typing with

double restriction mutagenesis primer PCR. *Clin Chem* 2004; **50**: 2376–2378.
19 Stassen HH, Hoffman K, Scharfetter C: Similarity by state/descent and genetic vector spaces: analysis of a longitudinal family study. *BMC Genet* 2003; **4** (Suppl 1): S59.
20 Stassen HH, Meier M: Master.GEN: A program package for the multivariate analysis of genetic diversity. *Mol Psychiatry* 1999; **4** (Suppl 1): S62.
21 Stassen HH, Begleiter H, Porjesz B, Rice J, Scharfetter C, Reich T: Structural decomposition of genetic diversity in families with alcoholism. *Genet Epidemiol* 1999; **17** (Suppl 1): S325–S330.
22 Stassen HH, Begleiter H, Beirut L *et al*: Oligogenic approaches to the predisposition of alcohol dependence. A genome-wide search on 255 families. *Neurol Psychiatr Brain Res* 2004.
23 Stassen HH, Hoffmann K, Scharfetter C: Analysis of a longitudinal family study; In: Almasy L, Amos CI, Bailey-Wilson JE, Cantor RM, Jaquish CE, Martinez M, Neuman RJ, Olson JM, Palmer LJ, Rich SS, Spence MA, MacCluer JW (eds.): Genetic Analysis Workshop 13: Analysis of longitudinal family data for complex diseases and related risk factors: *BMC Genet* 2003, vol. 4: pp S59, 51–56.
24 Stassen HH, Bridler R, Hell D, Weisbrod M, Scharfetter C: Ethnicity-independent genetic basis of functional psychoses. A genotype-to-phenotype approach. *Neuropsychiatr Genet* 2004; **124**: 101–112.
25 Stassen HH, Begleiter H, Hergersberg M *et al*: A multivariate feature vector approach to quantifying genetic diversity. *Mol Psychiatry* 1999; **4** (Suppl. 1): S62.
26 Stassen HH, Scharfetter C: Oligogenic approaches to the predisposition of asthma in ethnically diverse populations. *Genet Epidemiol* 2001; **21** (Suppl 1): S284–S289.
27 Hoffmann K, Stassen HH, Reis A: Genkartierung in Isolatpopulationen. *Medizinische Genetik* 2000; **12**: 428–437.
28 Devlin B, Risch N: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29**: 311–322.
29 Agresti A: *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons Inc., 1996, p section 7.3.5.
30 Suarez BK, Hampe CL, Van Eerdewegh P: Problems of replicating linkage claims in psychiatry; in: Gershon ES, Cloninger CR (eds).: *Genetic Approaches to Mental Disorders*. American Psychiatric Press, 1994.
31 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
32 Shima Y, Nakanishi K, Odawara M, Kobayashi T, Ohta H: Association of the SNP-19 genotype 22 in the calpain-10 gene with elevated body mass index and hemoglobin A1c levels in Japanese. *Clin Chim Acta* 2003; **336**: 89–96.
33 Ehrmann DA, Schwarz PE, Hara M *et al*: Relationship of calpain-10 genotype to phenotypic features of polycystic ovary syndrome. *J Clin Endocrinol Metab* 2002; **87**: 1669–1673.
34 Lynn S, Evans JC, White C *et al*: Variation in the calpain-10 gene affects blood glucose levels in the British population. *Diabetes* 2002; **51**: 247–250.
35 Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB: Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci USA* 1994; **91**: 3166–3170.
36 Frayling TM, McCarthy MI, Walker M *et al*: No evidence for linkage at candidate type 2 diabetes susceptibility loci on chromosomes 12 and 20 in United Kingdom Caucasians. *J Clin Endocrinol Metab* 2000; **85**: 853–857.
37 Frayling TM, Walker M, McCarthy MI *et al*: Parent-offspring trios: a resource to facilitate the identification of type 2 diabetes genes. *Diabetes* 1999; **48**: 2475–2479.
38 Evans JC, Frayling TM, Cassell PG *et al*: Studies of association between the gene for calpain-10 and type 2 diabetes mellitus in the United Kingdom. *Am J Hum Genet* 2001; **69**: 544–552.
39 Garant MJ, Kao WH, Brancati F *et al*: SNP43 of CAPN10 and the risk of type 2 diabetes in African-Americans: the Atherosclerosis Risk in Communities Study. *Diabetes* 2002; **51**: 231–237.
40 Orho-Melander M, Klannemark M, Svensson MK, Ridderstrale M, Lindgren CM, Groop L: Variants in the calpain-10 gene predispose to insulin resistance and elevated free fatty acid levels. *Diabetes* 2002; **51**: 2658–2664.
41 Sun HX, Zhang KX, Du WN *et al*: Single nucleotide polymorphisms in CAPN10 gene of Chinese people and its correlation with type 2 diabetes mellitus in Han people of northern China. *Biomed Environ Sci* 2002; **15**: 75–82.
42 Chen LX, Ji LN, Han XY, Zhu F: Study on Calpain10 gene polymorphism in Chinese type 2 diabetes families. *Zhonghua Yi Xue Za Zhi* 2003; **83**: 1856–1859.
43 Fingerlin TE, Erdos MR, Watanabe RM *et al*: Variation in three single nucleotide polymorphisms in the calpain-10 gene not associated with type 2 diabetes in a large Finnish cohort. *Diabetes* 2002; **51**: 1644–1648.
44 Daimon M, Oizumi T, Saitoh T *et al*: Calpain 10 gene polymorphisms are related, not to type 2 diabetes, but to increased serum cholesterol in Japanese. *Diabetes Res Clin Pract* 2002; **56**: 147–152.
45 Horikawa Y, Oda N, Yu L *et al*: Genetic variations in calpain-10 gene are not a major factor in the occurrence of type 2 diabetes in Japanese. *J Clin Endocrinol Metab* 2003; **88**: 244–247.
46 del Bosque-Plata L, Aguilar-Salinas CA, Tusie-Luna MT *et al*: Association of the calpain-10 gene with type 2 diabetes mellitus in a Mexican population. *Mol Genet Metab* 2004; **81**: 122–126.
47 Hegele RA, Harris SB, Zinman B, Hanley AJ, Cao H: Absence of association of type 2 diabetes with CAPN10 and PC-1 polymorphisms in Oji-Cree. *Diabetes Care* 2001; **24**: 1498–1499.
48 Malecki MT, Moczulski DK, Klupa T *et al*: Homozygous combination of calpain 10 gene haplotypes is associated with type 2 diabetes mellitus in a Polish population. *Eur J Endocrinol* 2002; **146**: 695–699.
49 Tsai HJ, Sun G, Weeks DE *et al*: Type 2 diabetes and three calpain-10 gene polymorphisms in Samoans: no evidence of association. *Am J Hum Genet* 2001; **69**: 1236–1244.
50 Rasmussen SK, Urhammer SA, Berglund L *et al*: Variants within the calpain-10 gene on chromosome 2q37 (NIDDM1) and relationships to type 2 diabetes, insulin resistance, and impaired acute insulin secretion among Scandinavian Caucasians. *Diabetes* 2002; **51**: 3561–3567.