Original Paper

Human Heredity

Hum Hered 2006;62:20–29
DOI: 10.1159/000095850

Received: April 28, 2006
Accepted after revision: July 25, 2006
Published online: September 21, 2006

SNP-Based Analysis of Genetic Substructure in the German Population

Michael Steffens[a] Claudia Lamina[b] Thomas Illig[b] Thomas Bettecken[d] Rainer Vogler[e] Patricia Entz[f] Eun-Kyung Suk[f] Mohammad Reza Toliat[f] Norman Klopp[b] Amke Caliebe[g] Inke R. König[h] Karola Köhler[i] Jan Lüdemann[j] Amalia Diaz Lacava[a] Rolf Fimmers[a] Peter Lichtner[d] Andreas Ziegler[h] Andreas Wolf[g] Michael Krawczak[g] Peter Nürnberg[f] Jochen Hampe[e] Stefan Schreiber[e] Thomas Meitinger[c, d] H.-Erich Wichmann[b] Kathryn Roeder[k] Thomas F. Wienker[a] Max P. Baur[a]

[a]Institute of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-University, Bonn, [b]Institute of Epidemiology, GSF National Research Center, Munich-Neuherberg, [c]Institute of Human Genetics, Technical University Munich, Munich, [d]Institute of Human Genetics, GSF National Research Center, Munich-Neuherberg, [e]Department of General Internal Medicine, Christian-Albrechts-University, Kiel, [f]Gene Mapping Center at the Max-Delbrück-Center, Berlin, [g]Institute of Medical Statistics and Informatics, Christian-Albrechts-University, Kiel, [h]Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, [i]Department of Genetic Epidemiology, University of Göttingen, Göttingen, [j]Institute of Clinical Chemistry and Laboratory Medicine, University of Greifswald, Greifswald, Germany; [k]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pa., USA

Key Words

German population · Genetic substructure · F-statistics · Genomic control

Abstract

Objective: To evaluate the relevance and necessity to account for the effects of population substructure on association studies under a case-control design in central Europe, we analysed three samples drawn from different geographic areas of Germany. Two of the three samples, POPGEN (n = 720) and SHIP (n = 709), are from north and north-east Germany, respectively, and one sample, KORA (n = 730), is from southern Germany. ***Methods:*** Population genetic differentiation was measured by classical F-statistics for different marker sets, either consisting of genome-wide selected coding SNPs located in functional genes, or consisting of selectively neutral SNPs from 'genomic deserts'. Quantitative estimates of the degree of stratification were performed comparing the genomic control approach [Devlin B, Roeder K: Biometrics 1999;55:997–1004], structured association [Pritchard JK, Stephens M, Donnelly P: Genetics 2000;155:945–959] and sophisticated methods like random forests [Breiman L: Machine Learning 2001;45:5–32]. ***Results:*** F-statistics showed that there exists a low genetic differentiation between the samples along a north-south gradient within Germany (F_{ST}(KORA/POPGEN): $1.7 \cdot 10^{-4}$; F_{ST}(KORA/SHIP): $5.4 \cdot 10^{-4}$; F_{ST}(POPGEN/SHIP): $-1.3 \cdot 10^{-5}$). ***Conclusion:*** Although the F_{ST}-values are very small, indicating a minor degree of population structure, and are too low to be detectable from methods without using prior information of subpopulation membership, such as STRUCTURE [Pritchard JK, Stephens M, Donnelly P: Genetics 2000;155:945–959], they may be a possible source for confounding due to population stratification.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/06/0621–0020$23.50/0

Accessible online at:
www.karger.com/hhe

Dr. Michael Steffens
Institute of Medical Biometry, Informatics and Epidemiology
Rheinische Friedrich-Wilhelms-University, Sigmund-Freud-Strasse 25
DE–53105 Bonn (Germany), Tel. +49 228 287 15243, Fax +49 228 287 15854
E-Mail steffens@imbie.meb.uni-bonn.de

Introduction

Undetected or disregarded population structure may appear to mask, change, reverse or mimic genetic effects of genes underlying complex traits [1], and may lower statistical power in linkage analysis [2]. In particular, genetic epidemiological case-control studies are susceptible to confounding effects of differential allele frequencies at disease and marker loci associated with local populations of different demographic history or ethnic background. This problem is well known since the early days of genetics [6], but, nonetheless, has been largely ignored in genetic epidemiological research. The revived interest in the case-control study design using genetic markers has stirred a recent debate about the relevance of confounding by population substructure and admixture [7–10] with lack of empirical evidence, especially when authors chose models with unrealistically differential prevalences leading to large effects [11, 12].

Generally, one can distinguish between two concepts to handle the problem of unknown population stratification in epidemiological studies, genomic control and structured association. In a series of papers [3, 13, 14], Devlin, Roeder and co-workers elaborated the concept of genomic control. The method uses a collection of supplementary non-candidate loci to estimate any inflation, λ, in the distribution of the association test statistics between unlinked genetic variants of cases and controls generated e.g. by population structure, and then corrects the association test statistics at the candidate loci by the inflation factor λ.

The second concept, structured association, was developed by Pritchard et al. [4]. It is based on a latent class model and stratifies the analysis according to the optimal number of subpopulation classes. Subsequent tests for association are performed within each subpopulation by treating the number of subpopulations and the subpopulation membership or admixture proportions for each individual as known quantities. In doing so, the null hypothesis is that the allele frequencies depend only on the subpopulation and not on the phenotype.

Even though both concepts have been published more than five years ago, there has been no human based, large-scale association study conducted in central Europe applying these concepts, particularly none with especially selected neutral markers of the human genome. This is remarkable in light of the findings of Helgason et al. [15], who found that the Icelandic population, previously thought to be a prime example of a homogeneous population, turned out to be stratified with the potential for a notable impact on association studies. One reason may be the unproven conjecture, that population differentiation is too small within the European population to cause substantial effect in association studies. However, Campbell et al. [16] recently demonstrated that population stratification can, indeed, lead to false positive associations in a sample of European Americans confounded by their European ancestry. To date, most studies inferring population structure using genomic controls or structured association analysis have been limited to human samples of predefined ethnic or geographic origin, or animal and plants involving strains or local races [17–24].

The rapid progress of the high-throughput genotyping technology with its preference for the case-control study design makes it more and more important to assess in advance the magnitude of genetic differences between samples, which are scheduled to be joined in future medical genetic research projects. Therefore, this study addresses the possible impact of population genetic effects on case-control association studies within the three main existent population cohorts in Germany.

Materials and Methods

Study Subjects

Population samples were recruited from three on-going cross-sectional epidemiological surveys of regional German populations: (1) KORA (Co-operative Health Research in the Region of Augsburg) from southern Germany [25, 26]; (2) POPGEN (Population Genetic Cohort) from Schleswig-Holstein, northern Germany [27], and (3) SHIP (Survey of Health in Pommerania) from east and northeast Germany [28, 29]. No degree of kinship was expected among the individuals, either between or within any population sample, but individuals were not explicitly tested for relationship a priori.

Sub-samples of more than 700 people for each sample (KORA: 730, POPGEN: 720, SHIP: 709) were genotyped at the same 212 SNP marker loci (fig. 1). The sub-samples were matched for age (KORA, POPGEN, SHIP: 54 $\pm$ 13 years) and proportions of males (KORA, POPGEN, SHIP: 50%) and showed no specific disease phenotype. Informed written consent was given by each person of the three population samples prior to enrolment in the study.

Genotyping, Marker Selection and Quality Control

All persons were genotyped at the same 212 SNP marker loci, which were subdivided into three marker sets of approximately 70 markers each. The marker sets, GCBerlin, GCKiel and GCMunich, were named according to the location of the genotyping centers where the genotyping for each marker set took place (fig. 1). Each center used a different genotyping technology, namely Berlin (Pyrosequencing™ and Taqman®), Kiel (Taqman®) and Munich (MALDI-TOF™; matrix assisted laser desorption/ionization).

The marker loci in this study were of two types: (i) One set of coding SNPs (GCKiel), which are located in exons of functional genes, causing an amino acid exchange in the resulting protein,

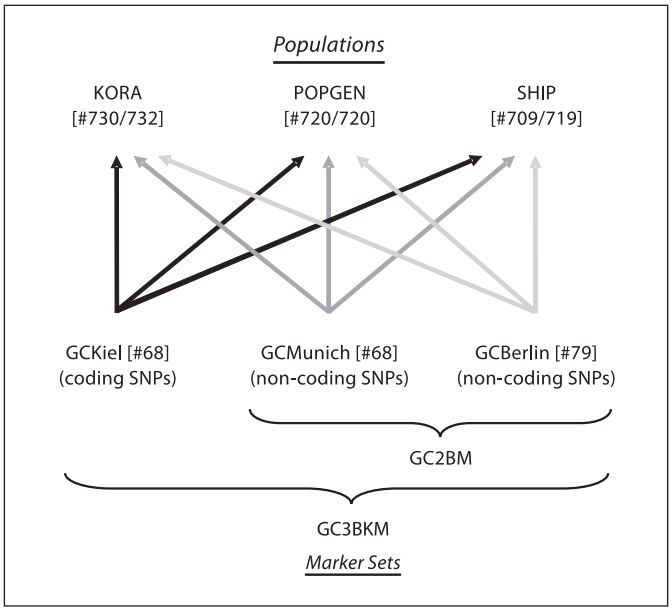

Fig. 1. Study design. Each population was genotyped by each of the three marker sets. The numbers in brackets under the populations show the sample size after and before data revision. The brackets to the right of the marker sets show the total number of markers per set. The sets GCMunich and GCBerlin include three overlapping markers.

Table 1. Characteristics of marker sets

GCMunich, GCBerlin: • *Intergenic* SNPs from genomic deserts, e.g. null loci, randomly chosen, uniformly spread across the genome • validated by minor allele frequency between 10 and 50% in Caucasian • 500 Kbp intermarker distance • >100 Kbp apart from any known gene region • >1 Mbp apart from centromers and telomers • 4 SNPs on each chromosome for each marker set • 2 SNPs on each chromosome arm for each set • 1 SNP for each interval of heterozygosity from 10 to 50% in steps of 10% on each chromosome • no markers selected from the Y-chromosome
GCKiel: • *Intragenic* exonic SNPs, protein coding genes, associated with an amino acid exchange or an effective promotor alteration

and thus are potentially subject to selective forces. (ii) Two sets of neutral SNPs (GCMunich, GCBerlin), which are located far from known genes, and which are uniformly spaced throughout the genome in putative 'genomic deserts'. These loci are assumed to fulfil the requirement of selective neutrality in the absence of specific information, and are expected to depict the neutral processes of drift and migration in demographic history. These GC SNPs were selected according to a set of rules envisaged by Devlin and Roeder (1999) and which were formalized by the Bonn group (table 1). The joined marker sets (GCBerlin + GCMunich) and (GCBerlin + GCKiel + GCMunich) will be referred to as 'GC2BM' and 'GC3BKM', respectively.

Table 2. Genotyping call rate broken down by marker set and population

Population	Genotyping center		
	Kiel	Munich	Berlin
SHIP	0.966	0.979	0.968
KORA	0.962	0.974	0.984
POPGEN	0.989	0.982	0.990

The data were subjected to an extensive quality process which comprised:

Standardized reporting of variants. Genotyping on different platforms resulted in non-uniform allele and genotyping designations, which were translated into dbSNP standards throughout.

Determination of the genotype call rate. The averaged call rate over all population samples was 96.1%. Call rates per marker set and sample are summarized in table 2.

Determination of DNA sample typing rate. A typing rate below 50% was regarded as indicative of impaired DNA quality, which may be dependent on genotyping methodology. In total 40 persons were excluded from the marker sets (GCBerlin: 11, GCKiel: 27, GCMunich: 2).

Unlikely multilocus genotypes were regarded as indicative of quality problems. One person with 80% of homozygous genotypes was removed from the data set.

Unintentional duplicates. 3 pairs of persons were found with a perfect genotype matching. For each pair, the person with the lower typing rate was removed.

Cryptic relatedness. By using 'Graphical Representation of Relationship Errors' (GRR) [30] we found 3 pairs of persons displaying outlying high values of IBS >1,73. These pairs were resolved in the same way as described in the previous point.

Hardy-Weinberg equilibrium (HWE) in each sample. The distribution of p values of the HWE test for each marker was tested for non-uniformity. No significant deviation was observed, and in particular no excess of low p values.

Pairwise linkage disequilibrium (LD) between markers in each sample (D' and r^2). LD values were randomly dispersed and did not show any inhomogeneity or irregularity.

Methods

Data analysis comprises classical population genetic statistics based on prior information about the population structure (F-statistics, the lambda inflation factor, genetic distances), as well as some more sophisticated methods (structured association analysis, random forests, prediction rates), which estimate the number of sub-populations in the sample from multi-locus marker data and account for the uncertainty related to unknown population

structure. In doing so, these methods rely on assumptions concerning Hardy-Weinberg equilibrium and linkage equilibrium within populations.

The *fixation indices* were introduced by S. Wright in 1921 [31, 32] and consist of three parameters, F_{ST} F_{IT}, and F_{IS}, which describe the allelic correlations in a hierarchically substructured population [33]. The programs FSTAT (version 2.9.3.2) [34] and PowerMarker (version 3.23) [35] were used to calculate the F-statistics. X-linked markers were excluded from this analysis to avoid bias due to the effects of natural selection and lower recombination rates on the X-chromosome.

The *inflation factor* λ was calculated by the median of the Armitage's trend test statistics divided by the expected 50%-quantile (0.455) of the association test χ-square distribution (d.f. = 1) under the null hypothesis of no association between the SNPs typed in cases and controls [3]. A bootstrapping procedure was used to calculate the confidence interval for the λ factors [36]. To provide a better insight into the nature of the data we report the λ factors without being rounded up to unity according to the definition of Devlin and Roeder [3] ($\lambda \equiv \max(1, \text{median}(\lambda)/0.455)$).

To infer the relative *individual admixture level* within the three population samples (KORA, POPGEN and SHIP) and among each other, the unsupervised clustering algorithm implemented in the program STRUCTURE (version 2.1) [4] was used. The analysis was conducted for several models with different populations numbers ranging from K = 1 to 5. For each model we run the program several times, using a burn-in period of 10,000 replicates and 100,000 iterations, resulting in stable model parameters and consistent results. First, each population was analysed on its own to detect any possible within population sub-structuring. Second, we pooled populations pair-wise together to see whether STRUCTURE might be able to separate and recover the original ones, and, third, we analysed all three populations simultaneously.

We derived the *prediction rate* as a measure of genetic structure in the style of a Naive Bayes classifier. The prediction rate indicates to what extent it is possible to identify the predefined populations from a given number of marker loci. It is defined as the expected posterior probability of a randomly drawn new individual from one of these samples being correctly classified to its predefined population. The prediction rate was estimated by leave-one-out cross-validation as follows: The posterior probability of each individual being classified to its predefined population is calculated given its multi-locus genotype data and the population genotype frequencies estimated by leaving that individual out. To determine the posterior probability, Bayes' formula is applied based on the assumption that all subpopulations are a priori equally likely. For each subpopulation the likelihood of the individual's genotype data assuming that the individual originates from that population has to be determined. This likelihood is calculated by multiplying the probabilities of the individual's genotypes over the loci, either based on genotype or the allele frequencies in the respective population. Finally, the prediction rate is estimated by averaging the posterior probabilities over all individuals. Confidence intervals for the prediction rate were calculated via bootstrapping over all loci.

Sub-population membership was also predicted by *random forests* according to Breiman [5]. In brief, random forests is an ensemble method comprising a pre-specified number of classification trees [36]. To allow for a better estimation of the classification error frequency, the total sample was randomly split into a training data (approximately two thirds) and a test data set (approximately one third). Random forests were developed in the training data with a missing value imputation based on estimated proximities from the random forests. Specifically, 500 classification trees were grown on bootstrap samples of all individuals in the training data, and a random selection of approximately square root of the available SNPs were used for each node within every tree. The resulting random forests were applied to predict individuals in the test data set where missing values were imputed by median values. Every individual in the test data set was finally classified according to the majority vote across all trees in the forest. Random forests were used for prediction of pairwise populations KORA vs. POPGEN, KORA vs. SHIP, and POPGEN vs. SHIP separately from all, only coding SNPs and from only non-coding SNPs. Frequencies of the classification error were determined in training and test data together with 95% confidence binomial intervals according to Clopper-Pearson [37].

Data Analysis

Analysis of population structure was performed using the five marker sets (GCBerlin, GCKiel, GCMunich, GC2BM, GC3BKM) and derived for all possible combinations of populations, joining populations if statistics are intended only for pairwise comparisons. We simulated two different scenarios when analysing the data set. In general, one might expect to see the biggest possible population differential in prevalence when comparing populations from large geographic distances. Thus, analysing population samples as they were collected in this study, from north (POPGEN), north-east (SHIP) and south (KORA) Germany, can be regarded as a 'worst case' scenario, where all the samples in the first group were drawn from one population and all the samples in the second group were from another. In contrast to this scenario, we simulated random population samples (GCRandom1, GCRandom2) to serve as representatives for a cross-section of the German population. The random population samples were created by bootstrapping using the pool of individuals of the original population (KORA, POPGEN and SHIP) as re-sampling units.

Results

F-Statistics

F_{ST} is the estimator for the coancestry coefficient, the correlation of alleles of different individuals in the same population, which quantifies the amount of genetic variation due to differences among populations or geographical regions. In the majority of cases, F_{ST} estimates were positive and quite low in our study (table 3a). F_{ST} estimates ranged from –0.00017 to 0.00071 with a maximal standard deviation of 0.00033. The highest F_{ST} estimates were identified for the 'KORA versus SHIP' comparison regardless of which distinct marker set (GCBerlin, GCKiel, GCMunich), or combined marker sets (GC2BM, GC3BKM), was used. F_{ST} values were essentially zero if a random sample was compared to any other sample.

Table 3. F-statistics

	GCKiel	GCMunich	GCBerlin	GC2BM	GC3BKM
a F_{ST}					
KORA, POPGEN	0.00003 ± 0.00014	0.00008 ± 0.00014	0.00039 ± 0.00017	0.00025 ± 0.00011	0.00017 ± 0.00009
KORA, SHIP	0.00032 ± 0.00018	0.00071 ± 0.00033	0.00066 ± 0.00022	0.00068 ± 0.00020	0.00054 ± 0.00014
POPGEN, SHIP	–0.00017 ± 0.00009	0.00009 ± 0.00012	0.00008 ± 0.00015	0.00008 ± 0.00010	–0.00001 ± 0.00007
KORA, POPGEN, SHIP	0.00006 ± 0.00010	0.00030 ± 0.00016	0.00039 ± 0.00013	0.00035 ± 0.00010	0.00024 ± 0.00008
b F_{IT}					
KORA, POPGEN	0.0061 ± 0.0033	0.0054 ± 0.0041	0.0083 ± 0.0039	0.0069 ± 0.0028	0.0066 ± 0.0021
KORA, SHIP	0.0049 ± 0.0028	0.0092 ± 0.0044	0.0108 ± 0.0036	0.0100 ± 0.0028	0.0080 ± 0.0021
POPGEN, SHIP	0.0005 ± 0.0031	0.0060 ± 0.0041	0.0067 ± 0.0036	0.0064 ± 0.0027	0.0041 ± 0.0020
KORA, POPGEN, SHIP	0.0038 ± 0.0024	0.0072 ± 0.0034	0.0080 ± 0.0033	0.0076 ± 0.0024	0.0062 ± 0.0017
c F_{IS}					
KORA, POPGEN	0.0061 ± 0.0033	0.0053 ± 0.0041	0.0079 ± 0.0039	0.0067 ± 0.0028	0.0065 ± 0.0022
KORA, SHIP	0.0046 ± 0.0028	0.0084 ± 0.0044	0.0102 ± 0.0036	0.0093 ± 0.0028	0.0075 ± 0.0021
POPGEN, SHIP	0.0006 ± 0.0031	0.0060 ± 0.0040	0.0066 ± 0.0036	0.0063 ± 0.0027	0.0041 ± 0.0020
KORA, POPGEN, SHIP	0.0038 ± 0.0024	0.0069 ± 0.0034	0.0076 ± 0.0033	0.0073 ± 0.0024	0.0059 ± 0.0017
d F_{IS}					
KORA	0.0096 ± 0.0046	0.0083 ± 0.0056	0.0074 ± 0.0047	0.0078 ± 0.0036	0.0084 ± 0.0028
POPGEN	0.0010 ± 0.0047	–0.0010 ± 0.0056	0.0058 ± 0.0047	0.0028 ± 0.0036	0.0022 ± 0.0029
SHIP	–0.0022 ± 0.0043	0.0092 ± 0.0061	0.0108 ± 0.0045	0.0100 ± 0.0037	0.0058 ± 0.0029

Mean estimator of the F-statistics with standard error calculated by jackknifing.
F_{ST} = Correlation of alleles of different individuals in the same population (coancestry); F_{IT} = correlation of alleles within individuals over all populations (inbreeding); F_{IS} = correlation of alleles within individuals within populations.

Table 4. Inflation factor λ

	GCKiel	GCMunich	GCBerlin	GC2BM	GC3BKM
KORA vs. POPGEN	0.948 (0.359, 1.888)	1.541 (0.973, 2.203)	1.707 (0.691, 2.768)	1.360 (0.911, 2.052)	1.138 (0.874, 1.806)
KORA vs. SHIP	1.496 (0.668, 2.386)	1.071 (0.567, 1.916)	1.779 (1.266, 3.180)	1.455 (0.975, 1.935)	1.455 (1.010, 1.867)
POPGEN vs. SHIP	0.609 (0.356, 1.464)	1.210 (0.681, 2.341)	1.246 (0.703, 1.737)	1.282 (0.897, 1.722)	0.977 (0.681, 1.472)
Random1 vs. Random2	1.030 (0.557, 1.666)	1.029 (0.531, 1.715)	1.031 (0.563, 1.652)	1.015 (0.656, 1.461)	1.013 (0.714, 1.372)
KORA vs. Random	0.970 (0.540, 1.537)	0.982 (0.517, 1.606)	1.125 (0.634, 1.742)	1.041 (0.690, 1.476)	1.008 (0.7211, 1.359)
POPGEN vs. Random	0.769 (0.414, 1.237)	0.816 (0.416, 1.351)	0.890 (0.494, 1.403)	0.845 (0.554, 1.200)	0.810 (0.574, 1.094)
SHIP vs. Random	0.920 (0.506, 1.470)	1.037 (0.552, 1.678)	1.046 (0.565, 1.683)	1.030 (0.669, 1.473)	0.980 (0.695, 1.317)

Inflation factor with 95% confidence interval for pairwise comparison of populations.

F_{IT} and F_{IS} describe the correlation of alleles within individuals (e.g. due to inbreeding) over all populations and within one sub-population, respectively. In the human species these values are typically one order of magnitude higher than F_{ST} values, confirming the well-known fact that most variability in human populations is observed within populations and only a minor fraction of genetic variation is due to differences between populations. If evolutionary forces have been acting over sufficient periods, F_{IT} values exceed F_{IS} values. Comparison of KORA from the south with the two northern population samples POPGEN and SHIP coincided in this sense indicating a small, but still measurable genetic differentiation between the southern and northern part of Germany (table 3b and c). For example using the GC2BM marker set, the maximum F_{IT} value reached was 0.0100 ± 0.0028 compared to 0.0093 ± 0.0028 for the F_{IS} value.

Lambda Inflation Factor

Pairwise comparisons of population samples according to the 'worst case' scenario showed λ values ranging from 0.609 to 1.779 (table 4). The marker sets GCKiel and

Table 5. Prediction rate

		GCKiel	GCMunich	GCBerlin	GC2BM	GC3BKM
a	KORA, POPGEN	0.5005 (0.4926, 0.5084)	0.5027 (0.4961, 0.5094)	0.5119 (0.5028, 0.5211)	0.5143 (0.5030, 0.5256)	0.5139 (0.5014, 0.5265)
	KORA, SHIP	0.5101 (0.5007, 0.5196)	0.5203 (0.5055, 0.5350)	0.5174 (0.5076, 0.5273)	0.5353 (0.5197, 0.5510)	0.5421 (0.5261, 0.5580)
	POPGEN, SHIP	0.4952 (0.4898, 0.5007)	0.5019 (0.4958, 0.5081)	0.5076 (0.4979, 0.5173)	0.5090 (0.4981, 0.5199)	0.5046 (0.4930, 0.5162)
	KORA, POPGEN, SHIP	0.3351 (0.3300, 0.3401)	0.3406 (0.3339, 0.3473)	0.3459 (0.3386, 0.3532)	0.3518 (0.3426, 0.3609)	0.3526 (0.3430, 0.3621)
b	KORA, POPGEN	0.5007 (0.4909, 0.5105)	0.5096 (0.5003, 0.5190)	0.5096 (0.4997, 0.5195)	0.5185 (0.5057, 0.5313)	0.5177 (0.5032, 0.5323)
	KORA, SHIP	0.5119 (0.4999, 0.5239)	0.5232 (0.5082, 0.5382)	0.5143 (0.5034, 0.5251)	0.5356 (0.5194, 0.5518)	0.5428 (0.5249, 0.5606)
	POPGEN, SHIP	0.4948 (0.4874, 0.5023)	0.5127 (0.5014, 0.5241)	0.5059 (0.4943, 0.5175)	0.5174 (0.5035, 0.5312)	0.5112 (0.4968, 0.5256)
	KORA, POPGEN, SHIP	0.3357 (0.3294, 0.3421)	0.3484 (0.3400, 0.3567)	0.3433 (0.3353, 0.3514)	0.3564 (0.3460, 0.3667)	0.3561 (0.3452, 0.3670)

Prediction rates with 95% confidence intervals based on allele frequencies (a) and genotype frequencies (b).

GCBerlin yielded the highest factors for the comparison 'KORA versus SHIP' and the lowest values for 'POPGEN versus SHIP'. Likewise, the marker set GCMunich identified higher factors for the comparison 'KORA versus POPGEN' than for 'POPGEN versus SHIP', but showed an outstanding low value for the comparison 'KORA versus SHIP'.

Marker set GCKiel resulted in smaller and GCBerlin in the higher absolute values than the other sets. Despite the apparent differences between the marker sets, the estimated inflation factor of each set was inside the 95%-confidence intervals of the inflation factor of the other sets. Unity was not included in the 95%-confidence interval only for the marker sets GCBerlin and GC3BKM comparing 'KORA versus SHIP'.

Lambda factors for comparisons of randomly created samples were close to unity and showed noticeable smaller confidence intervals than the comparisons between the original population samples. The upper limit of the 95% confidence interval of the inflation factor for the GC3BKM marker set comparing two randomly simulated samples was 1.372.

Structured Association

Regardless of the scenario, STRUCTURE failed to detect any substructure within or between the three populations KORA, POPGEN and SHIP. The model with the highest posterior probability for the data was always the one assuming the number of populations to be unity (K = 1), corresponding to a single population with no outliers. The estimate of the average individual admixture rates within a sample was in all scenarios nearly perfectly equal to the expected probability of random classification.

Prediction Rate

In most pairwise comparisons the prediction rate was close to 50%, with slight advantages for the marker sets including neutral, non-coding loci (table 5). Comparing 'KORA versus SHIP' the prediction rate increased slightly with an increasing number of marker loci up to 54% in the whole data set GC3BKM. For 'POPGEN versus SHIP' the prediction rate was always around 50%. Indeed the estimate was less than 50% using the GCKiel marker set. This corresponds with the negative estimate for F_{ST} for this comparison.

Random Forests

The results of predicting the pairwise population affiliation at a time is depicted in table 6. In the training data the classification error frequencies using all SNPs or merely the non-coding SNPs were always less then 50%, a value which would be expected by random classification. In contrast, upon use of the coding SNPs, even the

Table 6. Random forests

	Training data			Test data		
GC3BKM						
KORA	0.4752 (0.4300, 0.5208)	0.3306 (0.2888, 0.3745)		0.5021 (0.4372, 0.5669)	0.3320 (0.2728, 0.3953)	
POPGEN	0.4472 (0.4023, 0.4928)		0.3602 (0.3174, 0.4049)	0.4153 (0.3517, 0.4810)		0.3856 (0.3232, 0.4509)
SHIP		0.4328 (0.3875, 0.4790)	0.5139 (0.4676, 0.5599)		0.5381 (0.4723, 0.6030)	0.6695 (0.6055, 0.7292)
Total	0.4612 (0.4294, 0.4932)	0.3809 (0.3500, 0.4126)	0.4359 (0.4041, 0.4681)	0.4591 (0.4137, 0.5050)	0.4340 (0.3890, 0.4798)	0.5275 (0.4814, 0.5733)
GC2BM						
KORA	0.5103 (0.4648, 0.5557)	0.4566 (0.4116, 0.5022)		0.5975 (0.5326, 0.6600)	0.4149 (0.3520, 0.4799)	
POPGEN	0.4741 (0.4288, 0.5197)		0.4431 (0.3982, 0.4886)	0.4576 (0.3928, 0.5235)		0.4322 (0.3681, 0.4980)
SHIP		0.5224 (0.4761, 0.5684)	0.5522 (0.5060, 0.5979)		0.5678 (0.5020, 0.6319)	0.6271 (0.5620, 0.6890)
Total	0.4922 (0.4603, 0.5243)	0.4890 (0.4568, 0.5212)	0.4968 (0.4646, 0.5291)	0.5283 (0.48240, 0.5739)	0.4906 (0.4448, 0.5364)	0.5297 (0.4835, 0.5754)
GCKiel						
KORA	0.4339 (0.3892, 0.4794)	0.3182 (0.2769, 0.3617)		0.5809 (0.5159, 0.6439)	0.3568 (0.2964, 0.4209)	
POPGEN	0.4327 (0.3880, 0.4782)		0.3602 (0.3174, 0.4049)	0.4322 (0.3681, 0.4980)		0.4153 (0.3517, 0.4810)
SHIP		0.4009 (0.3562, 0.4468)	0.4691 (0.4232, 0.5154)		0.5508 (0.4850, 0.6154)	0.6059 (0.5405, 0.6687)
Total	0.4333 (0.4018, 0.4652)	0.3589 (0.3284, 0.3902)	0.4139 (0.3824, 0.4459)	0.5073 (0.4615, 0.5531)	0.4528 (0.4075, 0.4987)	0.5106 (0.4645, 0.5566)

Classification error frequencies with 95% confidence intervals for prediction of pairwise populations.

training cases could not be predicted more accurately than by chance. In the test sample, prediction of 'KORA versus POPGEN' or 'POPGEN versus SHIP' is not better than the random classification frequency of 50%. However, the confidence interval of the error frequency in predicting 'KORA versus SHIP' is below 50%, and therefore better than by chance for the marker sets GCKiel and GC3BKM including the coding SNPs.

Discussion

For the first time three large samples drawn from geographically different regions within Germany were analysed to address population genetic issues relevant for future case-control type epidemiological studies.

Summary statistics of population genetic differentiation describing the distributions of heterozygosity and F_{ST} values were consistent with the literature, in that only a minor proportion of the total variance of genetic variation is due to differences between populations, whereas the major part is found within populations [2, 38, 39]. Average F_{ST} values accounted for far less than 5% of the total variance in our sample. Known as Lewontin's fallacy, this does not justify the conclusion that populations are not differentiable by genetic data, because Lewontin's conclusion is based solely on the assumption that information arises only on a locus-by-locus analysis not including correlations amongst the different loci [40]. Instead, we could show that F_{ST} values, despite being small, were significant for the geographically most distant populations KORA and SHIP. However, between KORA and POPGEN the difference was small and not significant, and finally between SHIP and POPGEN, the two populations from the northern part of Germany, there was no clear-cut indication of relevant substructure. These observations are compatible with the results of Cavalli-Sforza et al., who reported a slight degree of population differentiation along a north-south gradient within Germany, using blood group data [41].

F_{ST} values for the marker sets of putatively neutral SNPs (GCBerlin and GCMunich) were in the range typical for European populations [42], but, as expected, much lower than reported for non-white US populations [43]. The estimated F_{ST} = 0.00054 between the most distant populations KORA and SHIP were 2–4 times lower than the F_{ST} value between Germans and other Germanic populations (Dutch, Danish, English, Austrian, Swiss, Belgians), which range from 0.0010 for Germans and Swiss to 0.0022 for Germans and English [41].

Though population genetic differences were quantifiable and detectable by all marker sets, they were too small to serve as a safe basis to predict population membership by current methodology. None of the tested classification or prediction methods was able to produce convincing classification results. The best prediction rate we achieved was only 54% for the pairwise comparison 'KORA versus SHIP', in which both populations showed an F_{ST} value of 0.00054. Low prediction rates in context with low F_{ST} values are not surprising. For instance, Shriver et al. [2] used 8525 autosomal SNPs and achieved a clustering rate of only 57% when clustering Japanese and Chinese individuals with a pairwise F_{ST} value of 0.045. Furthermore, restricting clustering to makers with the highest F_{ST} values (the upper 10%), which are believed to be subject to selection that has been more recent, or resulted from completely neutral evolutionary history, yielded only slightly better prediction rates [2]. We, too, did not achieve significant improvements of the prediction rates when restricting the analyses to markers with the highest F_{ST} values.

STRUCTURE failed to separate between KORA, POPGEN and SHIP. This is concordant with the results of Hao et al. [21], who could not detect any population stratification, when combining two groups of Caucasian samples. Moreover, Köhler and Bickeböller [44] showed in simulations that population stratification generally could be detected for $F_{ST} \geq 0.005$ but not for F_{ST} = 0.0025 using a mixture model as a clustering method. Since the Bayesian model implemented in STRUCTURE is also a probability based clustering method, a similar performance is to be expected. Thus, below a threshold of population genetic differentiation, the methods designed to estimate population structure from data without using prior information of population membership, do not perform well.

In contrast, the genomic control method allows adjusting for confounding even in the presence of subtle population stratification. The same observation were made by Hao et al., who compared two Caucasian populations collected from different geographic regions [21]. In addition, they demonstrated that bias in association tests can still be corrected and adjusted even if population stratification is not statistically significant. This is concordant with our findings that low F_{ST} values can accompany relevant inflation factors. For instance, under the assumption of population stratification due to a genetic differentiation equal in amount as evaluated between KORA and POPGEN with the marker set GC2BM (F_{ST} = 0.00025), the χ^2 statistic is inflated by 1.36. Depending on the size of the test statistic, even this seemingly small level of population differentiation implies an important shift of p values, e.g. $\chi^2 = 9$ (p = 0.0027) divided by 1.36 gives $\chi^2 = 6.62$ (p = 0.0101).

Another source of confounding, which can also result in elevated lambda values is cryptic relatedness, i.e. the kinship among cases or controls unknown to the investigator. Cryptic relatedness can occur, whenever there has been rapid and recent population growth or extensive inbreeding. As demonstrated by Voight and Pritchard [45] for a wide range of possible disease parameters and generations of ancestry (number of meioses separating two relatives), the magnitude of inflation due to cryptic relatedness is in practice quite small, hardly ever exceeding 1.07 in human populations. Thus, even if there were cryptic relatedness in our sample, it could only partly explain the observed levels of the inflation factors.

With larger sample sizes the power to detect confounding effects either due to population stratification or cryptic relatedness increases. Accordingly, under the genomic control approach, even small F_{ST} – values may result in sizeable inflation factors depending on sample size and mixture parameters. This can most obviously be seen by the approximation formula for the inflation factor λ: $\lambda \approx 1 + F_{ST} \cdot N \cdot P$, whereby N is the sample size per case and control group drawn from the mixed population and P is a factor proportional to the fraction of each subpopulation within the case- and control group ($P = (a1 - b1)^2 + (a2 - b2)^2$; a1, a2 = fraction of population 1 and 2 within cases; b1, b2 = fraction of population 1 and 2 within controls), thus P varies from 0 (ideal situation) to 2 (worst case scenario) [3].

It is to be mentioned that in their recent paper Campbell et al. [16] observed that STRUCTURE and the genomic control method did not detect significant stratification, but they still obtained false positive association between lactase resistance and height among European Americans. However, they discarded a larger quantity of markers not passing the quality control or being out of Hardy-Weinberg equilibrium, which might have resulted in a biased, downward estimate of the inflation factor λ.

Conclusion

The observed low levels of population sub-structuring gives a good indication that the German population is a suitable source for association studies in complex diseases as well as pharmaco-genetic studies. Nevertheless, even in populations with little heterogeneity it may be worthwhile to address population substructure by genomic controls or other methods. The quantitative estimate of the degree of stratification, derived from the genomic control approach in this study, suggests that at least in some cases, there is a need to correct for inflation to avoid false spurious association due to unobservable population stratification. This reasoning is in concordance with Freedman et al. [46], who emphasized the relevance of modest amounts of stratification even in well designed studies, and the findings of Campbell et al. [16] and Helgason et al. [15], who showed, that the level of population genetic differentiation found within the European population, has in reality the potential to act as a notable confounding factor on association studies.

Acknowledgements

The study was supported by the German BMBF/GEM and national genotyping platform of the NGFN. The SHIP study was funded by grants from the German Federal Ministry for Education and Research (BMBF, grant no. 01ZZ96030), of the Ministry for Education, Research and Cultural Affairs and the Ministry for Social Affairs of the State of Mecklenburg-West Pomerania.

Supplementary Material

Supplementary information is available at the Human Heredity website.

Table S1: Data Set. Genotypes, minor allele frequency with 95%-confidence interval, p value of Hardy-Weinberg deviation and observed and expected heterozygosity are given for each population (KORA, POPGEN and SHIP) and marker.

References

1 Deng HW: Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 2001;159:1319–1323.

2 Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW: The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics 2004;1:274–286.

3 Devlin B, Roeder K: Genomic control for association studies. Biometrics 1999;55:997–1004.

4 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959.

5 Breiman L: Random Forests. Machine Learning 2001;45:5–32.

6 Wahlund S: Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas 1928;11:65–106.

7 Jorgenson E, Liu X, Witte JS: Case-control analyses: Geneopardy! Genet Epidemiol 2005;29(suppl 1):S86–S90.

8 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.

9 Wacholder S, Rothman N, Caporaso N: Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2000;92:1151–1158.

10 Thomas DC, Witte JS: Point: population stratification: A problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002;11:505–512.

11 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. Nat Genet 2004;36:512–517.

12 Devlin B, Bacanu SA, Roeder K: Genomic Control to the extreme. Nat Genet 2004;36:1129–1130.

13 Devlin B, Roeder K, Bacanu SA: Unbiased methods for population-based association studies. Genet Epidemiol 2001;21:273–284.

14 Bacanu SA, Devlin B, Roeder K: The power of genomic control. Am J Hum Genet 2000;66:1933–1944.

15 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.

16 Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN: Demonstrating stratification in a European American population. Nat Genet 2005;37:868–872.

17 Pritchard JK: Deconstructing maize population structure. Nat Genet 2001;28:203–204.

18 Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G: Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 2002;12:602–612.

19 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–228.

20 Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB: Human population genetic structure and inference of group membership. Am J Hum Genet 2003;72:578–589.

21 Hao K, Li C, Rosenow C, Wong WH: Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip Human Mapping 10K array. Eur J Hum Genet 2004;12:1001–1006.

22 Yang BZ, Zhao H, Kranzler HR, Gelernter J: Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. Genet Epidemiol 2005;28: 302–312.

23 Maraganore DM, de AM, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PV, Frazer KA, Cox DR, Ballinger DG: High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 2005;77:685–693.

24 Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW: Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLoS Genet 2005;1:e70.

25 Holle R, Happich M, Lowel H, Wichmann HE, OKSG: KORA – a research platform for population based health research. Gesundheitswesen 2005;67:19–25.

26 Wichmann HE, Gieger C, Illig T, otKSG: KORA-gen – Resource for population genetics, controls and a broad spectrum of disease phenotypes. KORA-gen – Ressource für Bevölkerungsgenetik, Kontrolle und ein breites Spektrum an Krankheitsphänotypen. Gesundheitswesen 2005;67:26–30.

27 Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S: PopGen: Population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genet 2006;9:55-61.

28 John U, Greiner B, Hensel E, Ludemann J, Piek M, Sauer S, Adam C, Born G, Alte D, Greiser E, Haertel U, Hense HW, Haerting J, Willich S, Kessler C: Study of Health In Pomerania (SHIP): A health examination survey in an east German region: Objectives and design. Soz Präventivmed 2001;46:186–194.

29 Luedemann J, Schminke U, Berger K, Piek M, Willich SN, Doring A, John U, Kessler C: Association between behavior-dependent cardiovascular risk factors and asymptomatic carotid atherosclerosis in a general population. Stroke 2002;33:2929–2935.

30 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: GRR: Graphical representation of relationship errors. Bioinformatics 2001;17: 742–743.

31 Wright S: Coefficients of inbreeding and relationship. Am Nat 1922;56:330–338.

32 Wright S: Systems of mating. I. The biometric relations between parent and offspring. Genetics 1921;6:111–178.

33 Wright S: The genetical structure of populations. Annals of Eugenics 1951;15:323–354.

34 Goudet J: Fstat version 1.2: A computer program to calculate F-statistics. J Hered 1995; 86:485–486.

35 Liu K, Muse SV: PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 2005;21:2128–2129.

36 Weir BS: Genetic Data Analysis II. Sunderland, Sinauer Associates, 1996.

37 Clopper CJ, Pearson E: The use of confidence or fiducial limits illustrated in the case of binomial. Biometrica 1934;26:404–413.

38 Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL: An apportionment of human DNA diversity. Proc Natl Acad Sci USA 1997; 94:4516–4519.

39 Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA: The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 2000;66:979–988.

40 Edwards AW: Human genetic diversity: Lewontin's fallacy. Bioessays 2003;25:798–801.

41 Cavalli-Sforza LL, Menozzi P, Piazza A: The History and Geography of Human Genes. Princeton University Press, 1994.

42 Morton NE: The future of genetic epidemiology. Ann Med 1992;24:557–562.

43 Chakraborty R, Jin L: A unified approach to study hypervariable polymorphisms: Statistical considerations of determining relatedness and population distances. EXS 1993;67: 153–175.

44 Köhler K, Bickeböller H: Case-Control Association Tests Correcting for Population Stratification. Ann Hum Genet 2006;70:98–115.

45 Voight BF, Pritchard JK: Confounding from cryptic relatedness in case-control association studies. PLoS Genet 2005;1:e32.

46 Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn: Assessing the impact of population stratification on genetic association studies. Nat Genet 2004;36:388–393.