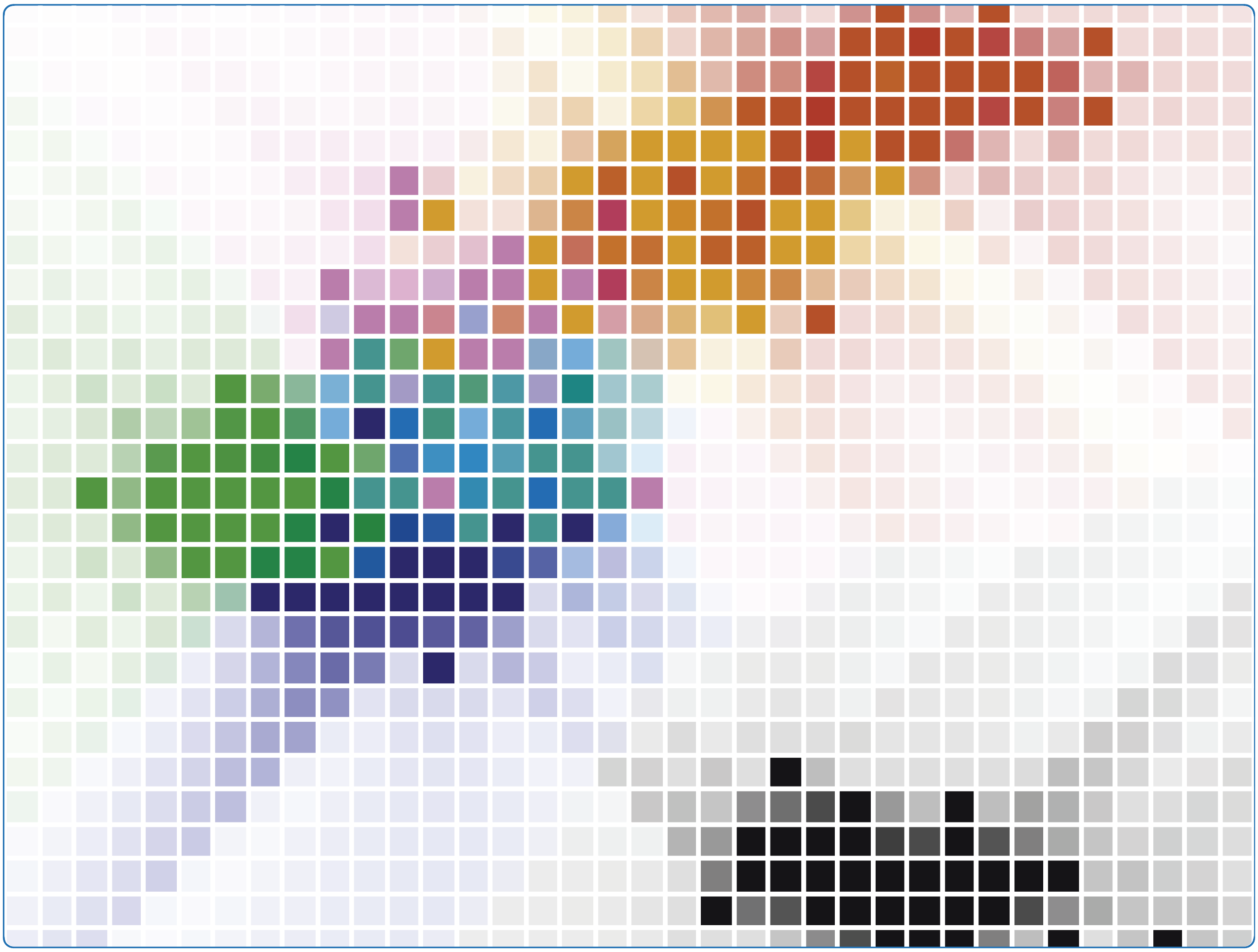
BMC Genetics



Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays

Gross *et al.*

BioMed Central

Gross *et al. BMC Genetics* 2011, **12**:67
http://www.biomedcentral.com/1471-2156/12/67 (28 July 2011)

BMC Genetics

RESEARCH ARTICLE Open Access

Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays

Arnd Gross[1,2], Anke Tönjes[3,4], Peter Kovacs[5], Krishna R Veeramah[6,7,8], Peter Ahnert[1,2], Nab R Roshyara[1,2], Christian Gieger[9], Ina-Maria Rueckert[9], Markus Loeffler[1,2], Mark Stoneking[10], Heinz-Erich Wichmann[9,11,12], John Novembre[6], Michael Stumvoll[3,4] and Markus Scholz[1,2*]

Abstract

Background: The Sorbs are an ethnic minority in Germany with putative genetic isolation, making the population interesting for disease mapping. A sample of N = 977 Sorbs is currently analysed in several genome-wide meta-analyses. Since genetic differences between populations are a major confounding factor in genetic meta-analyses, we compare the Sorbs with the German outbred population of the KORA F3 study (N = 1644) and other publically available European HapMap populations by population genetic means. We also aim to separate effects of over-sampling of families in the Sorbs sample from effects of genetic isolation and compare the power of genetic association studies between the samples.

Results: The degree of relatedness was significantly higher in the Sorbs. Principal components analysis revealed a west to east clustering of KORA individuals born in Germany, KORA individuals born in Poland or Czech Republic, Half-Sorbs (less than four Sorbian grandparents) and Full-Sorbs. The Sorbs cluster is nearest to the cluster of KORA individuals born in Poland. The number of rare SNPs is significantly higher in the Sorbs sample. FST between KORA and Sorbs is an order of magnitude higher than between different regions in Germany. Compared to the other populations, Sorbs show a higher proportion of individuals with runs of homozygosity between 2.5 Mb and 5 Mb. Linkage disequilibrium (LD) at longer range is also slightly increased but this has no effect on the power of association studies.
Oversampling of families in the Sorbs sample causes detectable bias regarding higher FST values and higher LD but the effect is an order of magnitude smaller than the observed differences between KORA and Sorbs.
Relatedness in the Sorbs also influenced the power of uncorrected association analyses.

Conclusions: Sorbs show signs of genetic isolation which cannot be explained by over-sampling of relatives, but the effects are moderate in size. The Slavonic origin of the Sorbs is still genetically detectable.
Regarding LD structure, a clear advantage for genome-wide association studies cannot be deduced. The significant amount of cryptic relatedness in the Sorbs sample results in inflated variances of Beta-estimators which should be considered in genetic association analyses.

Background

The Sorbs living in the Upper Lusatia region of Eastern Saxony are one of the few historic ethnic minorities in Germany. They are of Slavonic origin speaking a west Slavic language (Sorbian), and it is assumed that they have lived in ethnic isolation among the German majority during the past 1100 years [1]. Therefore, this population may be of special interest for genetic studies of complex traits.

The value of isolated populations for the discovery of genetic modifiers of diseases or quantitative traits is discussed controversially [2-6]. On the one hand, reduced genetic and environmental variability of isolated populations could increase genotypic relative risks [7,8]. In combination with the generally higher degree of linkage disequilibrium (LD) in isolated populations, this could

* Correspondence: markus.scholz@imise.uni-leipzig.de
[1] Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany
Full list of author information is available at the end of the article

improve the power of genetic association studies [5,6,9-11]. On the other hand, studies in isolated populations are often limited in size and, therefore, cannot match modern genome-wide association studies and meta-analyses comprising several tens of thousands of individuals.

Nowadays, it is common practice to combine all available genotyped and phenotyped populations in large-scale, whole genome meta-analyses or pooled analyses in order to identify even very small genetic effects as commonly observed for complex traits. Spurious associations caused by the genetic sub-structures of combined populations are the most serious concern of this approach [12-15], implying the need for appropriate adjustment strategies [16,17]. This is especially true if evidence from isolated and outbred populations is combined as this approach necessitates a thorough comparison of populations by population genetic means in order to determine their "degree of isolation" [6]. For this purpose, different methods have been proposed in the literature. For example, length and number of runs of homozygosity (ROHs) are discussed as an appropriate measure of isolation since they measure the degree of parental consanguinity [18]. LD is estimated to be higher in isolated populations because of lower generation numbers resulting in fewer recombination events [5,6]. Due to the smaller size of the founder population, it can also be expected that there is a lower number of polymorphisms in isolated populations [6,19,20]. Other markers of population structure such as F-statistics [21] are related to the measures mentioned above. Furthermore, genetic distances between populations can be determined by principal components analysis (PCA), allowing to quantify how closely populations are related [22]. By this technique genetic information can be mapped to topographic maps [14] allowing the assessment of a new indicator of isolation in the sense that an isolated population could be genetically far away from their geographic location. So far there appears to be no single measure sufficient to characterize the isolation of a population.

Another characteristic feature of isolated populations is the putatively higher degree of cryptic relatedness in randomly drawn samples. This is a serious concern in genetic association analysis and needs to be addressed with appropriate statistical methods [17,23-25]. Relatedness of individuals could also interact with the above mentioned measures of isolation of populations. Thus, when comparing two populations with different degrees of cryptic relatedness, it is not easy to decide whether differences in these measures can be traced back to different degrees of isolation or simply to over-sampling of related subjects.

The degree of isolation of the Sorbs has been studied in the past by the analysis of Y-chromosomal markers [26]. Recently, we compared a subset of about 200 Sorbs with other European isolates using 30,000 SNPs measured by microarrays [1]. In this analysis, the Sorbs expressed only moderate signs of isolation. Here, we analyse a sample of N = 977 Sorbs, which is currently included in several genome-wide association studies e.g. [27,28], and compare the Sorbs with the German outbred population of the KORA study [29]. Using the KORA study (N = 1644) and a larger sample of Sorbs (N = 977) provides more power than previous studies for comparing population genetic patterns between Sorbs and their neighbours. For this purpose, we assess the above mentioned population genetic characteristics: PCA, number of rare SNPs, F-statistics, ROHs, and LD. All analyses are based on genome-wide SNP array data. We also aim to separate effects of cryptic relatedness from effects of genetic isolation.

Furthermore, we analyse how differences between populations can be translated to differences in power of genetic association studies within these samples. We analyse the influence of genetic effect size, LD structure, heritability, and relatedness on power.

Methods

Study Populations

Sorbs

The Sorbs are of Slavonic origin, and lived in ethnic isolation among the Germanic majority during the past 1100 years [1]. Today, the Sorbian-speaking, Catholic minority comprises 15,000 full-blooded Sorbs resident in about 10 villages in rural Upper Lusatia (Oberlausitz), Eastern Saxony. A convenience sample of this population was collected including unrelated subjects as well as families. Details of the study population can be found elsewhere [28,30]. Genotyping and metabolic phenotyping of this sample was approved by the ethics committee of the University of Leipzig and is in accordance with the declaration of Helsinki. All subjects gave written informed consent before taking part in the study. A subset of individuals were genotyped with either Affymetrix Human Mapping 500 K Array Set (N = 483) or Affymetrix Genome-Wide Human SNP Array 6.0 (N = 494). Details on genotyping are described in [28]. A total of 977 subjects were available after quality control.

KORA

The study population was recruited from the KORA/MONICA S3 survey, a population-based sample from the general population living in the region of Augsburg, Southern Germany, which was carried out in 1994/95. In a follow-up examination of S3 in 2004/05 (KORA F3), 3006 subjects participated. Recruitment and study procedures of KORA have been described elsewhere [29,31]. For KORA F3 500 K we selected 1644 subjects of these participants then aged 35 to 79 years. Informed consent has been given, and the study has been approved by the local ethics committee. All KORA participants have a German passport. Genotyping of these individuals was performed with the Affymetrix Gene Chip Human Mapping 500 K Array Set as described in [32].

HapMap

174 CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah) and 88 TSI (Toscans in Italy) samples were taken from a recent HapMap Collection (Public Release 27, NCBI build 36, The International HapMap Project). From the CEU sample, we removed 58 children, five individuals with call rate < 90% and one individual because of cryptic relatedness (NA07045 because of lower call-rate compared to NA12813 [33]). In summary, we analysed 110 CEU and 88 TSI samples.

Data Analysis

Genotype Imputation and Quality Control

Missing genotypes of the KORA and Sorb samples were imputed separately using MACH Imputation Software with standard settings [34].

After Imputation, we checked 471,012 autosomal SNPs in the overlap of the Affymetrix Human Mapping 500 K Array Set and Affymetrix Genome-Wide Human SNP Array 6.0 for quality.

SNPs with a call rate less than 95% in all four study populations combined, prior to imputation, were filtered (34,711 SNPs). Hardy-Weinberg-Equilibrium (HWE) was tested across populations using a stratified test proposed by [35]. 10,712 SNPs with p-values less than 10^{-6} were eliminated. Finally, 14,508 SNPs showing unexpectedly high differences of allelic frequencies between genotyping platforms in the Sorbs sample were eliminated (p-value < 10^{-7}, see [1] for further details).

Since several SNPs violated more than one of our criteria, we discarded a total of 46,536 SNPs and analysed 424,476 remaining SNPs.

For estimation of ROHs (see below) the number of analysed SNPs is reduced to 306,081 by matching SNPs on Affymetrix chips with available SNPs in the HapMap CEU and TSI samples. Due to the high sensitivity of the PCA (see below) we decided to tighten our quality criteria for this kind of analysis. Only SNPs with a call rate of at least 99% were included for PCA, which reduced the number of SNPs to 199,702.

An overview of the data pre-processing workflow can be found in Additional file 1.

Estimation of Relatedness

Pair-wise relatedness between all individuals of KORA and Sorbs was estimated by the method described in [36]. For first degree relatives one would expect a value of r = 0.5, for second degree relatives a value of r = 0.25, and so on. Two individuals were considered as unrelated if the pair-wise relatedness estimate was not greater than 0.2, which approximately corresponds to the exclusion of first and second degree relatives.

For analyses of dependence of measures of population genetic comparison on relatedness, we define two subsamples used for all subsequent analyses: For the first subsample, the complete Sorbs sample (Sorbs$_{977}$, N = 977) was matched with a randomly selected subset of N = 977 unrelated KORA subjects born in Germany (KORA$_{977}$). For the second subsample, a subset of N = 532 unrelated Sorbs (Sorbs$_{532}$) was matched with a subset of N = 532 KORA subjects (KORA$_{532}$) randomly selected from KORA$_{977}$.

Unrelated subjects were selected by an algorithm which implements a step-by-step removal of individuals showing the highest number of relationships to other members of the population until no pair of individuals with relatedness > 0.2 remained.

Principal components analysis

PCA is suitable to map genetic variance to a few dimensions expressing the highest degree of variance [16,22]. It has been shown recently that the application of this technique to genome-wide genetic data is powerful enough to mirror even small geographic distances in Europe [14,37].

Since PCA results are biased in case of unequal population sizes [38], it was necessary to analyse subsamples of our populations. We performed PCA of 350 individuals from 7 subsamples of size N = 50, generated from the most unrelated individuals of our four study populations. The subsamples were defined as follows. Three subsamples were created from N = 1336, N = 140, and N = 80 individuals from KORA, who were born in Germany, in the Czech Republic, and in Poland, respectively. Two subsamples were generated from the Sorbs grouped by their degree of Sorbian ancestry. We identified 786 "Full"-Sorbs who stated that all four grandparents are Sorbs and 160 "Half"-Sorbs where at least one grandparent was not Sorbian. Another two subsamples were built from 110 CEU and 88 TSI samples.

PCA was done with iterative removal of outliers (default 5 iterations) and LD correction in consecutive SNPs (involving two previous SNPs as recommended in the manual of the EIGENSOFT package).

Rare SNPs

Isolated populations are supposed to have reduced genetic variability resulting in a higher number of rare SNPs. By definition, a SNP has a minor allelic frequency (MAF) of at least 1%. To account for variance we calculated the exact 95% confidence interval of the MAF and considered a SNP as rare if the interval was below one percent. This is equivalent to less than 11 observed alleles in Sorbs$_{977}$ or KORA$_{977}$ and less than five observed alleles in Sorbs$_{532}$ or KORA$_{532}$ respectively. The odds to find rare SNPs were compared between KORA and Sorbs using Fisher's exact test.

F-statistics

To characterize the variance of allelic frequencies within and between populations, we calculated F-statistics.

The inbreeding coefficient F_{IS} measures the correlation of alleles within an individual relative to the corresponding population. It is calculated by estimating the deviance of the observed number of heterozygote genotypes from what is expected under HWE. For every SNP, we calculated unbiased estimates as presented in [21], assessed the weighted average and determined the standard error of estimates by jack-knifing over individuals.

Correlation of alleles of individuals in the same population was estimated by the co-ancestry coefficient $F_{ST.}$ Since F_{ST} quantifies the amount of genetic variation between populations, it is used to define genetic distances between populations. We assessed F_{ST} for pairs of populations using a combined estimate of all SNPs [21] and calculated the standard error of estimates again by jack-knifing over individuals.

Runs of homozygosity

Counting ROHs is useful to detect inbreeding [18]. ROHs were determined in all individuals from KORA, Sorbs, CEU, and TSI using the PLINK Package (Version 1.07) with standard settings except for two parameters as noted below. PLINK estimates ROHs by searching for contiguous runs of homozygote genotypes. For this purpose, a window (default length 5000 kb, minimum 50 SNPs) is moved along the genome. To account for possible genotyping errors, at each SNP the homozygosity of the window is assessed allowing one (default) heterozygous genotype and five (default) missing calls. For each SNP the proportion of overlapping homozygous windows is calculated. If this proportion is high enough (default 5%) the SNP is considered to be part of a homozygous segment. Only homozygous segments longer than a given threshold (500 kb, default 1000 kb), consisting of a minimum number of 100 SNPs (default) and comprising a minimum SNP density of one SNP per 50 kb (default) were denoted as ROH. A homozygous segment can be split in two if two SNPs are at least 100 kb apart (default 1000 kb). Details on the algorithm can be found on the PLINK Homepage (see URLs).

Linkage disequilibrium

In the Sorbs and KORA samples, we calculated pairwise LD for all SNPs on Chromosome 22 (5382 markers) using robust estimators [39]. We used the widely accepted measures r [40] and $|D'|$ [41] to quantify LD. Since both measures depend on allelic frequencies, we also used the newly proposed measure $|\eta_1|$, which is independent of allelic frequencies. Hence, it is especially useful when comparing populations [42]. The measure η_1 is a monotone function of the odds ratio λ [43] ranging between -1 and 1. It is defined as

$$\eta_1 = \begin{cases} 2\dfrac{\lambda^2 - \lambda - \lambda \ln \lambda}{(\lambda - 1)^2} - 1 & if\, \lambda \neq 1 \\ 0 & if\, \lambda = 1 \end{cases}$$

Its absolute value is the percentage of SNP pairs under the non-informative uniform distribution with less extreme LD than the one observed (see [42] for details). Measures of LD were averaged using bins of 5 kb length as proposed by Olshen et al. [44]. Resulting means were smoothed by a LOWESS estimator [45].

Comparison of power assuming uncorrelated phenotypes

We analysed how the observed differences in LD structure between KORA and Sorbs can be translated into differences in power of genetic association studies. For this purpose, we assumed a linear regression model $\mathbf{y} = \beta_1 \mathbf{s}_1 + \boldsymbol{\varepsilon}_1$ of a random phenotype y which is influenced by a genotype $\mathbf{s}_1$ of a causative SNP, and $\boldsymbol{\varepsilon}_1$ is the residual Gaussian error of the model.

The SNP is assumed to explain a pre-specified proportion of the total variance of the phenotype which is denoted as R_s^2 in the following. In consequence, we can assume $\beta_1 = 1$ without restriction of generality. Within the distance of ± 2 Mb we now analysed the model $\mathbf{y} = \beta_2 \mathbf{s}_2 + \boldsymbol{\varepsilon}_2$ for a second SNP, which is in maximum LD (measured by r) with the causative SNP. That is, we analysed the best proxy of the causative SNP rather than the causative SNP itself modelling the marker principle of genetic association studies. The estimator $\hat{\beta}_2$ is normally distributed and depends on $\mathbf{s}_1$, $\mathbf{s}_2$, and R_s^2:

$$\hat{\beta}_2 \sim \mathrm{N}\left(\frac{\mathrm{Cov}(s_1, s_2)}{\mathrm{Var}(s_2)}, \frac{\dfrac{\mathrm{Var}(s_1)}{R_s^2} - \dfrac{\mathrm{Cov}(s_1, s_2)^2}{\mathrm{Var}(s_2)}}{\sum\limits_{i=1}^{n} (s_{2i} - \overline{s_2})^2} \right).$$

Where n is the number of individuals, s_{2i} is the genotype of the i-th individual and $\overline{s_2}$ is the average. The formula is derived in Additional file 2. We calculated the power of the regression analysis, i.e. the probability that the observed p-value is smaller than a given significance level (p-value threshold) when testing $\hat{\beta}_2$ against the null hypothesis $\beta_2 = 0$ using the above formula. This was done for all SNPs on Chromosome 22 in KORA$_{977}$, KORA$_{532}$, Sorbs$_{977}$, and Sorbs$_{532}$. Distribution of power was derived using the results of all SNPs of Chromosome 22. Results were compared between the KORA and Sorbs samples of equal size.

Comparison of power assuming correlated phenotypes

In the previous section, we derived formulae for the estimation of power under the assumption of uncorrelated phenotypes. This approach applies for either a negligible relatedness structure of the individuals or a weak correlation of phenotypes of related individuals. Applying a GRAMMAR approach [17], deviations from this situation can be corrected resulting again in the situation considered in the previous section.

However, to our knowledge, it is still not common practice in genome-wide association studies to use this approach to correct for relatedness. Therefore, we aim to study the situation in which the phenotypes are correlated but in which the corresponding individuals were analysed as independent even though they are not.

Following Amin *et al.* [17], we simulated phenotypes $\mathbf{y}$ on the basis of the mixed model $\mathbf{y} = \beta_1\mathbf{s}_1 + \mathbf{g} + \boldsymbol{\varepsilon}_1$, comprising a fixed effect of genotypes $\mathbf{s}_1$, a random effect representing the residual polygenic effects $\mathbf{g} \sim \mathrm{N}_n(\mathbf{0}, \sigma_g^2\mathbf{G})$ and non-genetic residuals $\boldsymbol{\varepsilon}_1 \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Here, $\mathbf{G}$ represents the pair-wise relatedness matrix. The model results in non-trivial covariance of phenotypes of different individuals. For each SNP we drew 1000 samples from the model and analysed the linear model $\mathbf{y} = \beta_2\mathbf{s}_2 + \boldsymbol{\varepsilon}_2$ for a second SNP which is in maximum LD to the first SNP in complete analogy to the procedure developed for uncorrelated phenotypes (see previous section). Different degrees of heritability $R_h^2 = R_s^2 + R_g^2$ were simulated, where R_s^2 is the explained variance by genotypes $\mathbf{s}_1$ and R_g^2 is the explained variance by polygenetic effects $\mathbf{g}$. Providing values for R_h^2 and R_s^2 results in the variance components $\sigma_g^2 = \mathrm{Var}(s_1)(\frac{R_h^2}{R_s^2} - 1)$ and $\sigma^2 = \mathrm{Var}(s_1)\frac{1 - R_h^2}{R_s^2}$, which follow after some calculations.

Statistical Software and Web-Resources

HapMap data were downloaded from [46]. Estimation of Eigenvectors for comparison of all subsamples was done with the EIGENSOFT package (Version 3.0, [47]). ROHs were determined by the PLINK Package (Version 1.07, [48]) [49].

All other calculations were performed using the Statistical Software package R (Version 2.8.0, [50]) [51].

Results

For population genetic comparison of the Sorbian minority in Germany with the German KORA population, several measures of genetic isolation were applied to genome-wide SNP array data.

Relatedness

We analysed the relatedness of all 476,776 pairs of individuals in the Sorbs and all 1,350,546 pairs in the KORA samples. Results are shown in Figure 1. Frequencies of relationships differ remarkably between the two samples. Emphasized by the different scales of the histograms, it can be clearly recognized that the numbers of

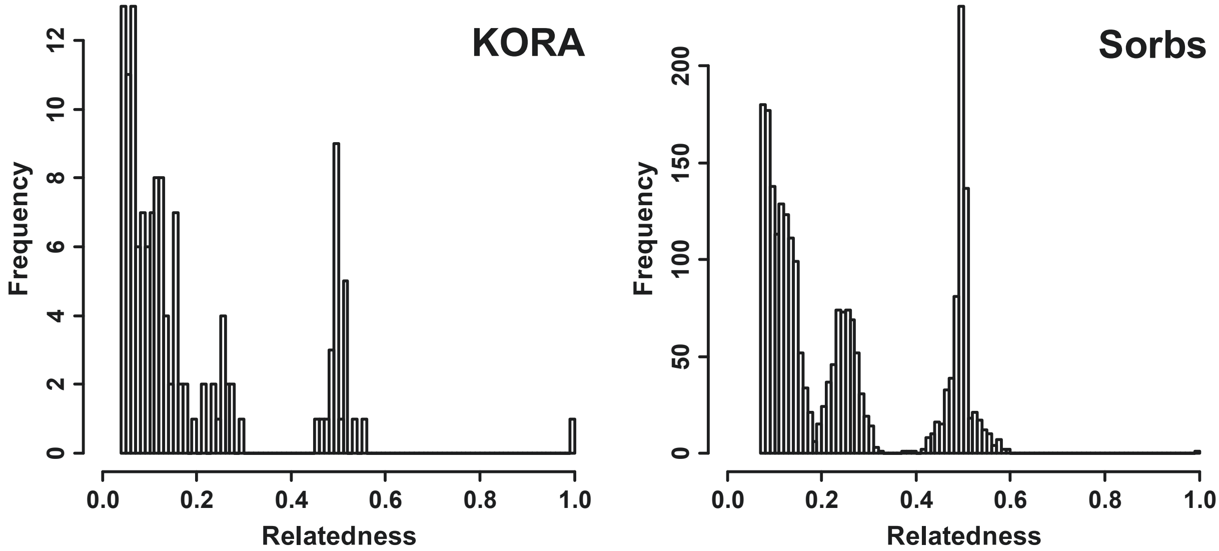


Figure 1 Distribution of degrees of relatedness in KORA and Sorbs. Distribution of degrees of relatedness in the KORA and Sorbs samples. For readability, the distribution of the 0.01% highest relatedness estimates of the KORA samples and the highest 0.5% estimates of the Sorbs samples are shown.

first and second degree relationships are higher in the Sorbs compared to KORA. Numbers of pairs with estimates over a given threshold are shown in Table 1 for both populations. We also provide odds-ratios for the encounter of a related pair.

To achieve samples without pairs of individuals with relatedness-estimates greater than 0.2, it was necessary to exclude 445 Sorbs and 33 KORA individuals, resulting in subsamples of 532 Sorbs and 1,611 KORA individuals.

Principal components analysis

Results of PCA after removal of outliers and LD correction are shown in Figure 2. The figure comprises all 150 individuals from KORA, 97 Sorbs, 49 HapMap CEU and 48 HapMap TSI after outlier removal.

A plot of the genetic variance represented by the first two principal components impressively reflects the geographic origin of these populations. TSI samples are relatively far away from the other clusters giving an orientation of a north to south axis. The KORA population is very close to the CEU HapMap population. In contrast, the Sorbian population clusters significantly eastwardly. There is a clear trend of west to east clustering of KORA individuals born in Germany, KORA individuals born in Poland or Czech Republic, Half-Sorbs, and finally, Full-Sorbs. The Sorbs clusters are nearest to the cluster of KORA individuals born in Poland.

Rare SNPs

When analysing 424,476 quality SNPs in 977 Sorbs (Sorbs_{977}) and the random Sample of 977 individuals from KORA (KORA_{977}), we counted 51,204 rare SNPs in Sorbs_{977} and 49,721 rare SNPs in KORA_{977} (p-value 6.7×10^{-7}). In the subset of 532 unrelated Sorbs (Sorbs_{532}) and the random sample of 532 unrelated individuals from KORA (KORA_{532}), we counted again more rare SNPs in the Sorbs_{532} than in KORA_{532}, i.e. 49,257 and 47,913 (p-value 4.7×10^{-6}), respectively.

F-Statistics

Estimating F_{IS} in the samples KORA_{977} and KORA_{532} resulted in slightly positive values with the smaller value in KORA_{977}. In contrast, in the samples Sorbs_{977} and Sorbs_{532}, we find slightly negative values with smaller value in the sample Sorbs_{977}.

F_{ST} estimates are somewhat higher between KORA_{977} and Sorbs_{977} than between KORA_{532} and Sorbs_{532}. F_{ST} estimates are higher than corresponding F_{IS} estimates, indicating a clear genetic distance between the two populations. All statistics can be found in Table 2.

Runs of Homozygosity

ROHs were determined for the populations KORA, Sorbs_{977}, Sorbs_{532}, CEU, and TSI. Percentages of individuals in these populations containing at least one ROH in a specified length interval were calculated (Figure 3). Compared to the other populations, Sorbs show a higher proportion of individuals with ROHs between 2.5 Mb and 5 Mb.

In a second step, mean total length of ROHs with a given minimum length was estimated averaged over the individuals of each population (Figure 4). Again, Sorbs differ from the other populations and are characterized by higher mean total length of ROHs. However, the effect is less pronounced if only long ROHs are considered. The mean total length of ROHs is shorter for Sorbs_{532} than for Sorbs_{977} but the difference is small.

Linkage Disequilibrium

Three measures of LD were calculated for KORA_{977}, KORA_{532}, Sorbs_{977}, and Sorbs_{532}. Results of η_1 are shown in Figure 5. Other measures such as r and D' behave similarly (data not shown). LD in the KORA sample is markedly lower at long ranges compared to Sorbs. This result is robust against dropping related individuals in the Sorb sample.

As expected for KORA_{977} and KORA_{532} a small sample size bias can be observed. In contrast the estimators for Sorbs_{977} and Sorbs_{532} are virtually identical.

Comparison of power assuming uncorrelated phenotypes

The power to detect causal SNPs was calculated for KORA_{977}, KORA_{532}, Sorbs_{977}, and Sorbs_{532}. Results for SNP effects with explained variances of 2% or 5% can be found in Figure 6. Since the results are virtually identical for KORA and Sorbs, we present the quartiles of the power distribution in Table 3 for p-value thresholds of 1×10^{-5} and 1×10^{-7}.

Table 1 Distribution of pair-wise relatedness estimates

Lower Bound	Number of pairs in KORA	Number of pairs in Sorbs	Odds ratio (KORA = reference category) [95% CI]
0.1	79	1889	68 [54;86]
0.2	38	1186	88 [64;126]
0.4	24	666	79 [52;123]
0.6	1	1	3 [0;222]

Number of pair-wise relatedness estimates above a given boundary for a total of 476776 and 1350546 calculated pair-wise estimates in Sorbs and KORA, respectively. We also present the odds-ratio for an encounter of relatives and corresponding 95% confidence interval.

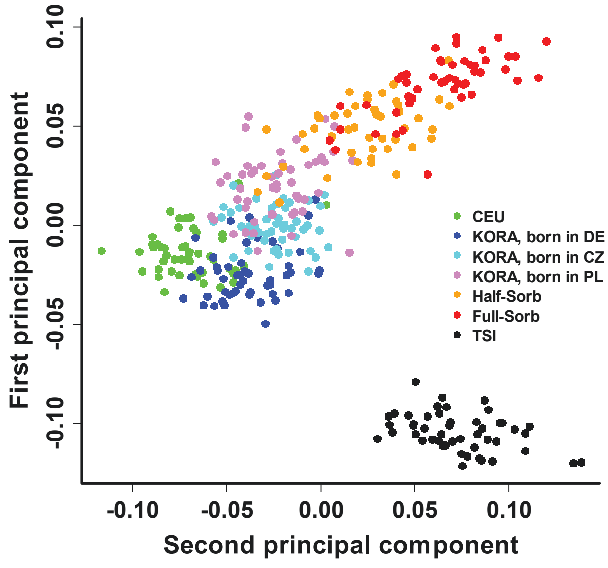


Figure 2 Principal components analysis of study populations. First two principal components of individuals from KORA born in Czech Republic (N = 50), Germany (N = 50), Poland (N = 50) and Full-Sorbs (N = 49), Half-Sorbs (N = 48), CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah, N = 49) and TSI (Toscans in Italy, N = 48).

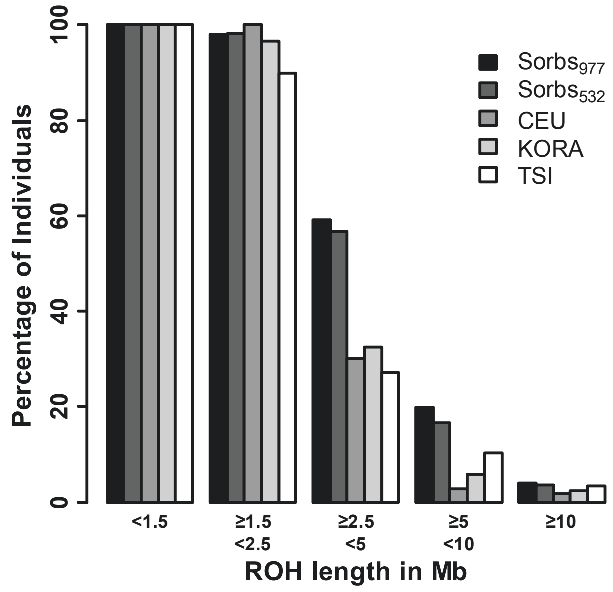


Figure 3 Proportion of individuals with certain ROH length. Proportion of individuals from KORA (N = 1644), Sorbs$_{977}$, Sorbs$_{532}$, CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah, N = 110) and TSI (Toscans in Italy, N = 88) with at least one ROH in the given length interval.

Comparison of power assuming correlated phenotypes

In Table 4 we present the power estimates assuming a heritability of 100% resulting in the greatest differences compared to Table 3. However, except for Sorbs$_{977}$, there are only very small differences between Tables 3 and 4 and even for Sorbs$_{977}$ the differences appear to be not substantial. For an explained variance of 2%, the power in Sorbs$_{977}$ increases, but it decreases for an explained variance of 5%. This is due to dependence on the significance threshold. Independent of the explained variance of the SNPs, the power under maximum heritability (100%) is greater than under minimal heritability (R_s^2) for small p-value thresholds. But for large p-value thresholds, the opposite is true (see Additional file 3).

The explanation for this behaviour is the inflation of the variance of the β-estimator caused by high levels of relatedness in the Sorbs$_{977}$ sample (see Additional file 4).

Results for other degrees of heritability are presented in Additional file 5. As expected, in the case of minimal heritability the results of our simulations under the mixed model and the results obtained with our analytical formula used in the previous section are coincident.

Table 2 Inbreeding and co-ancestry coefficients

Population	F-statistic	Estimate	SE
KORA$_{977}$	F_{IS}	0.0012	2.7×10^{-4}
Sorbs$_{977}$	F_{IS}	-0.0006	2.7×10^{-4}
KORA$_{532}$	F_{IS}	0.0014	3.5×10^{-4}
Sorbs$_{532}$	F_{IS}	-0.0002	3.6×10^{-4}
KORA$_{977}$, Sorbs$_{977}$	F_{ST}	0.0034	5.4×10^{-5}
KORA$_{532}$, Sorbs$_{532}$	F_{ST}	0.0029	6.7×10^{-5}

Estimates and standard errors (SE) of inbreeding coefficients F_{IS} and co-ancestry coefficients F_{ST} for KORA and Sorbs.

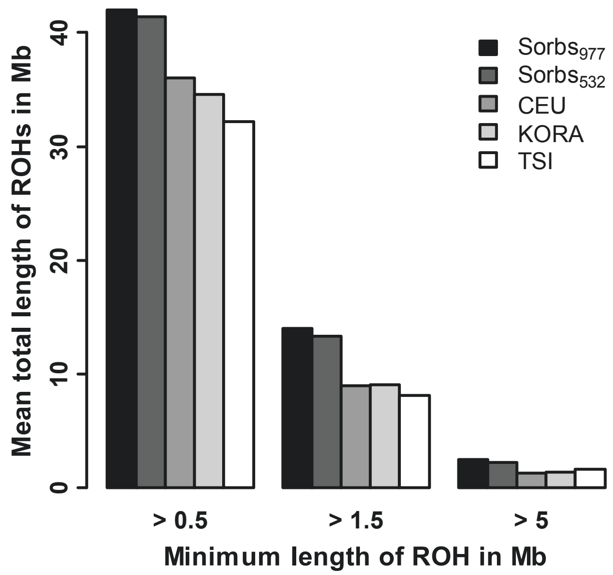


Figure 4 Average total length of ROHs. Average total length of ROHs for KORA (N = 1644), Sorbs$_{977}$, Sorbs$_{532}$, CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah, N = 110) and TSI (Toscans in Italy, N = 88) in dependence on minimal length of a single run.

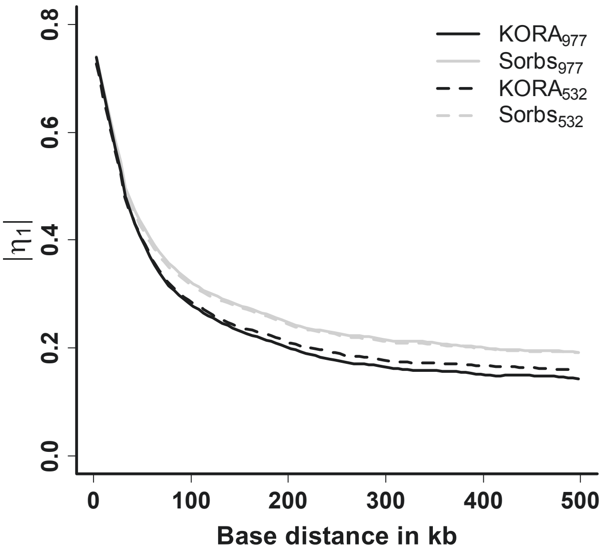


Figure 5 LD structure in KORA and Sorbs. LD structure in the KORA$_{977}$, KORA$_{532}$, Sorbs$_{977}$ and Sorbs$_{532}$ samples. η_1 was estimated for all SNP pairs of chromosome 22. Results are averaged over distance using bins of 5 kb length and smoothed by a LOWESS estimator.

Discussion

The Sorbs, resident in Lusatia, Germany, are an ethnic minority of Slavonic origin. Using genome-wide SNP array techniques, we aimed to compare this putatively isolated population with a German mixed population (KORA study) by various population genetic means. The Sorbs were compared recently with other European populations or isolates on the basis of a limited set of genetic markers and a limited set of unrelated individuals [1,52]. In the present analysis, we studied the Sorbs from the perspective of ongoing genome-wide association studies. That is, we compared the population with a German mixed population on the basis of complete sets of genotyped individuals, and a large number of genotyped SNPs. We also aimed to separate the effect of isolation from potential effects caused by over-sampling of relatives in the Sorbs. Finally, we studied the implications of observed differences between KORA and Sorbs for the analysis, and especially, the power of genome-wide association studies.

Genotype data from a sample of 977 Sorbs were available from genotyping with 500 k and 1000 k Affymetrix SNP chips. While SNP markers come with certain drawbacks (ascertainment bias, need for careful QC), they have proven useful for detecting subtle population structures.

For comparison with a German mixed population, we used the KORA F3 sample (N = 1644) and corresponding genotypes from 500 k Affymetrix SNP chips. Observed differences between regions of Germany are typically an order of magnitude lower than differences observed between Sorbs and KORA [53]. Publicly available European-American HapMap samples were also included in the analysis.

A major goal of our study was to distinguish effects of genetic isolation from simple over-sampling of families in the Sorbs. Since most of the population genetic measures used to compare populations assume

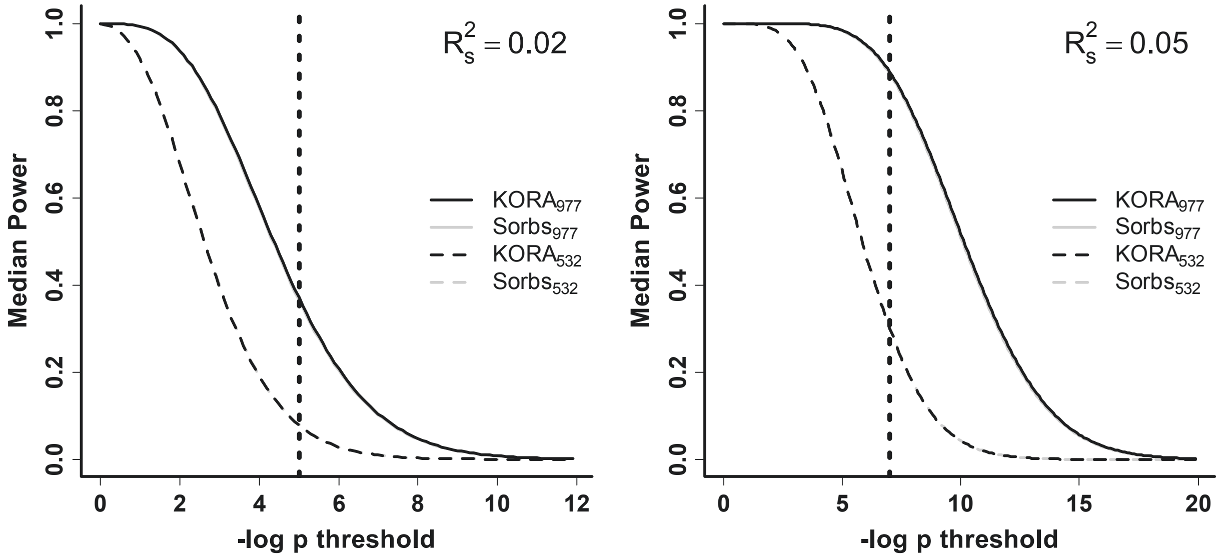


Figure 6 Median power distribution in KORA and Sorbs. Median power to detect SNP effects explaining 2% (left) or 5% (right) of variance, respectively. Power is plotted versus the p-value threshold. The grey lines are virtually covered by the black lines. The dotted line corresponds to p-value thresholds of 1×10^{-5} and 1×10^{-7} respectively.

Table 3 Quartiles of power distribution assuming uncorrelated phenotypes

Explained variance	p-value threshold	Population	1st Quartile	Median	3rd Quartile
2%	1×10^{-5}	KORA_{977}	6.78	37.02	49.19
2%	1×10^{-5}	Sorbs_{977}	6.31	36.51	49.34
2%	1×10^{-5}	KORA_{532}	1.15	7.85	11.52
2%	1×10^{-5}	Sorbs_{532}	1.13	7.88	11.65
5%	1×10^{-7}	KORA_{977}	25.01	88.8	95.81
5%	1×10^{-7}	Sorbs_{977}	23.14	88.37	95.87
5%	1×10^{-7}	KORA_{532}	2.73	30.07	43.41
5%	1×10^{-7}	Sorbs_{532}	2.66	30.17	43.85

Quartiles of the power distribution in percent for an explained variance of 2% with a p-value threshold of 1×10^{-5} and of 5% with a p-value threshold of 1×10^{-7}, respectively.

independence of individuals, over-sampling of families in certain samples may introduce a source of bias which is difficult to control. Indeed, we discovered a large number of closely related individuals within the Sorbs sample. Therefore, we repeated all analyses for a sub-group of Sorbs for which all relationships with relatedness estimates greater than 0.2 were removed. This does not completely resolve the problem of increased relatedness within the Sorbs sample but provides a trend for potential biases introduced by over-sampling of families. Indeed, such biases could be detected in our data but it is not substantial at least for the population genetic measures studied.

Since relatedness cannot be completely removed from the samples, a cut-off of 0.2 for the relatedness estimate seems to be feasible to study the effect of relatedness and to keep the sample size at an acceptable level. We also studied a cut-off of 0.1 reducing the sample size to N = 414. Results can be found in Additional file 6. Although tending slightly towards zero, results are essentially the same as those obtained for the cut-off of 0.2.

For some analyses such as determination of rare SNPs and LD it is known that sample size can introduce bias [39,44,54]. Therefore, for most comparisons we used randomly drawn subsamples of KORA which are of the same size as the Sorbs samples.

PCA is a proven means to detect even very small genetic differences between populations with high power. For European populations, it was demonstrated that the first two appropriately scaled principal components can map individuals to their geographic origin on the European continent with high precision, when all four grandparents are from the same location [14]. Our PCA results showed clear distances between KORA, Sorbs, and individuals from Tuscany. Using individuals from KORA and Tuscany to roughly orient the PCA graph on a map of Europe, Sorbs are positioned towards the East. KORA individuals are very close to the CEU HapMap population, while the distance to Tuscan/TSI individuals is much larger.

We conclude that the Slavonic origin of the Sorbs is still clearly genetically detectable. The analysis revealed that there is a west to east sequence of the clusters of KORA individuals born in Germany, KORA individuals born in Poland or Czech Republic, Half-Sorbs, and finally, Full-Sorbs. Although birthplace is not a stringent indicator of ethnicity, it is a commonly used surrogate in genetic epidemiologic studies if more detailed information cannot be ascertained. On the other hand, most of the KORA individuals born in Poland or Czech Republic are descendents from German minorities of these countries. Hence, on the basis of our data we cannot conclude that the Sorbs are genetically more distant

Table 4 Quartiles of power distribution assuming correlated phenotypes

Explained variance	p-value threshold	Population	1st Quartile	Median	3rd Quartile
2%	1×10^{-5}	KORA_{977}	6.7	37.1	48.4
2%	1×10^{-5}	Sorbs_{977}	10.08	38.95	48.9
2%	1×10^{-5}	KORA_{532}	1.2	7.8	11.6
2%	1×10^{-5}	Sorbs_{532}	1.3	8.2	11.9
5%	1×10^{-7}	KORA_{977}	24.78	88.3	95.12
5%	1×10^{-7}	Sorbs_{977}	27.3	83.6	91.8
5%	1×10^{-7}	KORA_{532}	2.73	29.9	42.9
5%	1×10^{-7}	Sorbs_{532}	2.9	30.4	43.5

Quartiles of the power distribution in percent for an explained variance of 2% with a p-value threshold of 1×10^{-5} and of 5% with a p-value threshold of 1×10^{-7}, respectively. A heritability of 100% is assumed.

from Germany than a random sample from Poland or Czech Republic. Half-Sorbs can be assumed to be closer to the German population than Full-Sorbs due to mating with German neighbours. This is clearly reflected by the localization of Half-Sorbs between KORA individuals and Full-Sorbs. There is a trend that the Sorbs are closer to the KORA individuals born in Poland than to the KORA individuals born in Czech Republic which is in agreement with a recently stated hypothesis that the Sorbs are genetically closer to Polish than to Czech [1].

Since it has been suggested that genetic diversity is lower in isolated populations [6], we analysed the number of rare SNPs. Indeed, we found a higher number of rare SNPs in the Sorbs sample compared to the KORA sample. Although significant, the difference is small in size.

The F_{ST} statistics between KORA and Sorbs were an order of magnitude higher than usually observed between different regions of Germany [53]. Thus, variance between KORA and Sorbs is much higher than expected for different regions in Germany. Surprisingly, the F_{IS} statistic was positive for KORA but negative for Sorbs. Such a phenomenon has also been observed for other isolated populations, suggesting that there may be signs of recent isolation breaking in the Sorbs [44]. Another indicator of isolation breaking is the relatively high number of Half-Sorbs (N = 160) in the present sample, i.e. subjects who claim to have less than four Sorbian grandparents. It should be remarked that the F_{IS} statistic is a population based measure rather than an individual based measure of inbreeding studied in [1].

ROH analysis was proposed to detect signs of isolation by estimation of inbreeding [18]. Despite the simplicity of this concept, calculation of ROH depends on many variable parameter settings such as SNP density or allowed numbers of missings or heterozygous markers, which heavily influence the results. Parameter settings are extensively discussed in McQuillan et al [18]. For our analysis, we used the default settings of PLINK except for two parameters: The threshold for homozygous segments was 500 kb (PLINK default is 1000 kb) and the splitting of homozygous segments can occur if two neighbouring SNPs are 100 kb apart (PLINK default is 1000 kb). Hence, we used the same settings as in McQuillan et al. except for the minimum number of contiguous homozygous SNPs constituting a ROH, for which we kept the PLINK default (N = 100). The results of ROH analysis also depend on allelic frequencies of populations and SNP-selections used by different genotyping technologies. Since McQuillan et al. [18] used a different genotyping platform (Illumina Infinium HumanHap300v2), the latter modification was necessary to obtain similar results.

We found that Sorbs have enriched ROHs of intermediate length (between 2.5 Mb and 5 Mb) compared to KORA, CEU, and TSI. This effect is much less pronounced for longer ROHs. Accordingly, the coverage of the genome by ROHs is higher in the Sorbian population. Following the argumentation of McQuillan et al., we conclude that there is a lack of recent parental relatedness in the Sorbs (no differences for long range ROHs) but that there are signs of ancient parental relatedness or the existence of autozygous segments of older pedigree structures (differences for ROHs of intermediate range). The lack of direct parental relatedness is in accordance with our estimates of F_{IS}.

Furthermore, we compared the LD structure of chromosome 22 between the KORA and the Sorbs population. We used the newly proposed LD measure η_1 for the comparison of KORA and Sorbs. In contrast to the more popular measures *r* and *D'*, the measure η_1 is independent of allelic frequencies [42]. In our opinion, this property is desirable when comparing LD structure between populations of potentially differing allelic frequencies. However, the results obtained by the three measures are very similar (data not shown).

An expected small upward bias caused by smaller sample size in KORA_{532} compared to KORA_{977} could be clearly detected. In contrast, the results for Sorbs_{977} and Sorbs_{532} are virtually identical. We conclude that the expected upward bias of the reduced Sorbs_{532} sample is nullified by the elimination of relationships. This interpretation is supported by the fact that a random sample of N = 532 individuals from Sorbs_{977} resulted in the same sample size bias as observed for KORA (data not shown). That is, LD is upwardly biased by the relatedness structure in the Sorbs. Nevertheless, even if relationships are eliminated to a reasonable degree (first and second degree relationships), Sorbs show generally higher LD at longer distances than is observed in KORA. It has been already shown in the literature that LD excess at longer ranges is a characteristic of isolated populations [5,9-11]. However, the effect is moderate in size which is also in agreement with several other populations considered as isolated [44,55-57].

Since LD structure directly influences the coverage of a SNP technology, and with it, the power of genome-wide association studies, we performed power analyses in the Sorbs and KORA samples. For this purpose, we defined a fixed genetic effect of an arbitrary SNP at chromosome 22. Explained variance was used as a measure of effect in order to adjust for differences in allelic frequencies. For this SNP, we analysed the best proxy SNP available on chromosome 22 in order to mimic a situation in which an unobserved causative variant is detected via a marker in LD. We derived an analytical formula for our model for the case of negligible heritability for which individuals can be considered as independent. This formula also applies to situations where correction for relatedness effects has been performed, for instance with a GRAMMAR approach [17]. Power was calculated for all SNPs on chromosome 22 and the resulting distribution was compared between the Sorbs and KORA samples with and without relatives. No differences regarding power were detected. We conclude that there is no gain in power due to higher LD in the Sorbs.

Since relatedness structure is often neglected in genetic association studies, we also analysed the influence of present relatedness structure on the power of an uncorrected analysis. This analysis is done via simulations of a linear mixed model comprising a fixed effect of a SNP and random polygenetic and non-genetic effects. We showed that the variance of the β-estimator is inflated under relatedness and high heritability. This results in a gain in power for higher p-value thresholds and a loss of power for lower p-value thresholds in the Sorbs$_{977}$, irrespective of the size of the genetic effect considered. The explanation is that normal distributions with different variances are overlapping.

We conclude that relatedness in the Sorbs$_{977}$ sample influences the power of uncorrected genetic association studies. Influence of relatedness on power is highest under maximum heritability of the phenotype. However, directions of power differences depend on the size of the genetic effect in combination with the significance threshold chosen.

In our simulations we did not observe a scenario resulting in a clear power benefit in the Sorbs$_{977}$ sample. However, this does not rule out that there might be a higher power in the Sorbs due to increased effect sizes caused, e.g., by higher environmental homogeneity or lower number of causative variants [7,8].

Conclusions

We could show that there are signs of genetic isolation within the Sorbs which cannot be explained by oversampling of relatives. The effects are moderate in size. The Slavonic origin of the Sorbs is still genetically detectable. Although there is higher LD in the Sorbs, the difference to KORA is small. Power analysis showed that a clear advantage of the Sorbs for genome-wide association studies with respect to coverage cannot be expected.

The significant amount of cryptic relatedness in the Sorbs sample results in inflated variances of β-estimators which should be considered in genetic association analyses.

Additional material

Additional file 1: Workflow of data pre-processing. The workflow of data pre-processing is presented. We start with the autosomal SNP data of four different populations (KORA, Sorbs, HapMap CEU, HapMap TSI). Numbers of remaining markers at each step of pre-processing are presented in bold.

Additional file 2: Derivation of the formula for $\hat{\beta}_2$.

Additional file 3: Comparisons of power for Sorbs$_{977}$ for minimal and maximal heritability of phenotypes. Simulation results of the power for minimal ($\sigma_g^2 = \mathrm{Var}(s_1)(\frac{R_h^2}{R_s^2} - 1)$) and maximal (100%) heritability. For the minimal heritability, we present the results of our analytical formula. The values presented in Tables 3 and 4 are displayed in bold.

Additional file 4: Variance inflation under relatedness. Comparison of the theoretical variance of the β_1-estimator assuming uncorrelated phenotypes (analytical formula $\mathrm{var}\ (\beta_1) = \frac{1}{N-1}(\frac{1}{R_s^2} - 1)$) with the averaged variances over all SNPs of chromosome 22 under a heritability of 100% assuming correlated phenotypes. The standard error of this estimate and the inflation factor are also provided. Sorbs$_{977}$ are presented in bold due to high inflation of variances of β_1-estimates.

Additional file 5: Simulation results for power under assumption of correlated phenotypes. Heritability was modified between R_s^2 and 100%. Explained variances of the SNP are 2% or 5% with corresponding p-value thresholds of 10^{-5} and 10^{-7}, respectively. All simulations were performed for KORA$_{977}$, Sorbs$_{977}$, KORA$_{532}$, and Sorbs$_{532}$. Power distribution is derived using the results of all SNPs of Chromosome 22.

Additional file 6: Additional inbreeding and co-ancestry coefficients. Estimates and standard errors (SE) of inbreeding coefficients F_{IS} and co-ancestry coefficients F_{ST} for KORA and Sorbs and different levels of relatedness: without filtering for relatedness (KORA$_{977}$, Sorbs$_{977}$), filtering for relatedness > 0.2 (KORA$_{532}$, Sorbs$_{532}$), filtering for relatedness > 0.1 (KORA$_{414}$, Sorbs$_{414}$). Indices refer to resulting numbers of cases.

Acknowledgements

We thank Knut Krohn and Beate Enigk for conducting microarray experiments of the Sorbs sample at the IZKF Leipzig at the Faculty of Medicine of the University of Leipzig (Projekt Z03).

We gratefully acknowledge the contributions of P. Lichtner, G. Eckstein, Guido Fischer, T. Strom and all other members of the Helmholtz Centre Munich genotyping staff in generating the SNP dataset as well as the contribution of all members of field staffs who were involved in the planning and conduct of the MONICA/KORA Augsburg studies. The KORA group consists of H.E. Wichmann (speaker), A. Peters, C. Meisinger, T. Illig, R. Holle, J. John and their co-workers who are responsible for the design and conduct of the KORA studies.

We thank Maelle Salmon for helping with data quality control. We thank Karsten Krug and Lars Thielecke for their technical assistance.

Finally, we express our appreciation to all participants of the Sorb and the KORA study for donating their blood and time.

Funding

The KORA research platform (KORA: Cooperative Research in the Region of Augsburg) and the MONICA Augsburg studies (Monitoring trends and determinants on cardiovascular diseases) were initiated and financed by the Helmholtz Zentrum München-National Research Center for Environmental Health, which is funded by the German Federal Ministry of Education, Science, Research and Technology and by the State of Bavaria. Part of this work was financed by the German National Genome Research Network (NGFN). Our research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ. AT, PK and MStu received financial support from the German Research Council (KFO-152), IZKF (B27), and the German Diabetes Association. MSto is funded by the Max Planck

Society. AG and PA are funded by the German Federal Ministry for Education and Research (01KN0702). AG, PA, NRR, and MSch were funded by the Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes, and Environment (LIFE Center, Universität Leipzig). LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF), the European Social Fund (ESF), and by means of the Free State of Saxony within the framework of its excellence initiative.

Author details

[1]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany. [2]LIFE Center (Leipzig Interdisciplinary Research Cluster of Genetic Factors, Phenotypes and Environment), University of Leipzig, Philipp-Rosenthal Strasse 27, 04103 Leipzig, Germany. [3]Department of Medicine, University of Leipzig, Liebigstrasse 18, 04103 Leipzig, Germany. [4]IFB Adiposity Diseases, University of Leipzig, Stephanstrasse 9c, 04103 Leipzig, Germany. [5]Interdisciplinary Center for Clinical Research, University of Leipzig, Liebigstrasse 21, 04103 Leipzig, Germany. [6]Dept Eco & Evo Biol, Interdepartmental Program in Bioinformatics, University of California, 621 Charles E. Young Dr South, Box 951606, Los Angeles, Los Angeles, CA 90095-1606 USA. [7]Center for Society and Genetics. University of California, 1323 Rolfe Hall, Box 957221, Los Angeles, Los Angeles, CA 90095-7221, USA. [8]Dept of History, University of California, 6265 Bunche Hall, Box 951473, Los Angeles, Los Angeles, CA 90095-1473, USA. [9]Helmholtz Centre Munich, German Research Center for Environmental Health, Institute of Epidemiology, Ingolstaedter Landstraße 1, 85764 Neuherberg, Germany. [10]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. [11]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians-University, Marchioninistraße 15, 81377 Munich, Germany. [12]Klinikum Grosshadern, Ludwig Maximilians University, Marchioninistraße 15, 81377 Munich, Germany.

Authors' contributions

Design of the Study: MSch. Design of the Sorbs study and data collection: AT, PK, MStu. Design of the KORA data collection: CG, IR, HW. Data analysis: AG, NRR, MSch. Writing: AG, MSch. Contribution to writing and discussion: KRV, PA, ML, MSto, AT, PK, MStu, JN.
All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 14 March 2011 Accepted: 28 July 2011
Published: 28 July 2011

References

1. Veeramah KR, Tonjes A, Kovacs P, Gross A, Wegmann D, Geary P, Gasperikova D, Klimes I, Scholz M, Novembre J, *et al*: **Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity.** *European Journal of Human Genetics* 2011.
2. Abbott A: **Manhattan versus Reykjavik.** *Nature* 2000, **406(6794)**:340-342.
3. Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA: **The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes.** *Nat Genet* 2000, **25(3)**:320-323.
4. Taillon-Miller P, Bauer-Sardina I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok PY: **Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28.** *Nat Genet* 2000, **25(3)**:324-328.
5. Shifman S, Darvasi A: **The value of isolated populations.** *Nat Genet* 2001, **28(4)**:309-310.
6. Kristiansson K, Naukkarinen J, Peltonen L: **Isolated populations and complex disease gene identification.** *Genome Biol* 2008, **9(8)**:109.
7. Sheffield VC, Stone EM, Carmi R: **Use of isolated inbred human populations for identification of disease genes.** *Trends Genet* 1998, **14(10)**:391-396.
8. Arcos-Burgos M, Muenke M: **Genetics of population isolates.** *Clin Genet* 2002, **61(4)**:233-247.
9. Tenesa A, Wright AF, Knott SA, Carothers AD, Hayward C, Angius A, Persico I, Maestrale G, Hastie ND, Pirastu M, *et al*: **Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations.** *Hum Mol Genet* 2004, **13(1)**:25-33.
10. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, Ospina J, Ruiz-Linares A, Macedo A, Palha JA, *et al*: **Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies.** *Nat Genet* 2006, **38(5)**:556-560.
11. Angius A, Hyland FC, Persico I, Pirastu N, Woodage T, Pirastu M, De la Vega FM: **Patterns of linkage disequilibrium between SNPs in a Sardinian population isolate and the selection of markers for association studies.** *Hum Hered* 2008, **65(1)**:9-22.
12. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: **Genetic structure of human populations.** *Science* 2002, **298(5602)**:2381-2385.
13. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, *et al*: **Genotype, haplotype and copy-number variation in worldwide human populations.** *Nature* 2008, **451(7181)**:998-1003.
14. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, *et al*: **Genes mirror geography within Europe.** *Nature* 2008, **456(7218)**:98-101.
15. Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M: **Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs.** *PLoS One* 2009, **4(11)**:e7888.
16. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38(8)**:904-909.
17. Amin N, van Duijn CM, Aulchenko YS: **A genomic background based method for association analysis in related individuals.** *PLoS One* 2007, **2(12)**:e1274.
18. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, *et al*: **Runs of homozygosity in European populations.** *Am J Hum Genet* 2008, **83(3)**:359-372.
19. Peltonen L, Jalanko A, Varilo T: **Molecular genetics of the Finnish disease heritage.** *Hum Mol Genet* 1999, **8(10)**:1913-1923.
20. Peltonen L: **Positional cloning of disease genes: advantages of genetic isolates.** *Hum Hered* 2000, **50(1)**:66-75.
21. Weir BS: **Genetic Data Analysis II.** Sunderland, MA: Sinauer Associates, Inc; 1996.
22. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2(12)**:e190.
23. Choi Y, Wijsman EM, Weir BS: **Case-control association testing in the presence of unknown relationships.** *Genet Epidemiol* 2009, **33(8)**:668-678.
24. Zhang F, Deng HW: **Correcting for cryptic relatedness in population-based association studies of continuous traits.** *Hum Hered* 2010, **69(1)**:28-33.
25. Thornton T, McPeek MS: **ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure.** *Am J Hum Genet* 2010, **86(2)**:172-184.
26. Krawczak M, Lu TT, Willuweit S, Roewer L: **Genetic diversity in the German population.** *Handbook of Human Molecular Evolution* John Wiley & Sons; 2008.
27. Kottgen A, Pattaro C, Boger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV, *et al*: **New loci associated with kidney function and chronic kidney disease.** *Nat Genet* 2010.
28. Tonjes A, Koriath M, Schleinitz D, Dietrich K, Bottcher Y, Rayner NW, Almgren P, Enigk B, Richter O, Rohm S, *et al*: **Genetic variation in GPR133 is associated with height: genome wide association study in the self-contained population of Sorbs.** *Hum Mol Genet* 2009, **18(23)**:4662-4668.
29. Wichmann HE, Gieger C, Illig T: **KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes.** *Gesundheitswesen* 2005, **67(Suppl 1)**:S26-30.
30. Tonjes A, Zeggini E, Kovacs P, Bottcher Y, Schleinitz D, Dietrich K, Morris AP, Enigk B, Rayner NW, Koriath M, *et al*: **Association of FTO variants with BMI and fat mass in the self-contained population of Sorbs in Germany.** *Eur J Hum Genet* 2010, **18(1)**:104-110.
31. Holle R, Happich M, Lowel H, Wichmann HE: **KORA–a research platform for population based health research.** *Gesundheitswesen* 2005, **67(Suppl 1)**:S19-25.

32. Doring A, Gieger C, Mehta D, Gohlke H, Prokisch H, Coassin S, Fischer G, Henke K, Klopp N, Kronenberg F, *et al*: **SLC2A9 influences uric acid concentrations with pronounced sex-specific effects.** *Nat Genet* 2008, **40(4)**:430-436.
33. Pemberton TJ, Wang C, Li JZ, Rosenberg NA: **Inference of unexpected genetic relatedness among individuals in HapMap Phase III.** *Am J Hum Genet* 2010, **87(4)**:457-464.
34. Li Y, Willer C, Sanna S, Abecasis G: **Genotype imputation.** *Annu Rev Genomics Hum Genet* 2009, **10**:387-406.
35. Troendle JF, Yu KF: **A note on testing the Hardy-Weinberg law across strata.** *Ann Hum Genet* 1994, **58(Pt 4)**:397-402.
36. Wang J: **An estimator for pairwise relatedness using molecular markers.** *Genetics* 2002, **160(3)**:1203-1215.
37. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, *et al*: **Correlation between genetic and geographic structure in Europe.** *Curr Biol* 2008, **18(16)**:1241-1248.
38. McVean G: **A genealogical interpretation of principal components analysis.** *PLoS Genet* 2009, **5(10)**:e1000686.
39. Scholz M, Hasenclever D: **Comparison of Estimators for Measures of Linkage Disequilibrium.** *The International Journal of Biostatistics* 2010, **6(1)**.
40. Hill WG, Robertson A: **Linkage Disequilibrium in Finite Populations.** *Theoretical and Applied Genetics* 1968, **38**:226-231.
41. Lewontin RC: **The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.** *Genetics* 1964, **49(1)**:49-67.
42. **A Canonical Measure of Allelic Association.** [http://arxiv.org/PS_cache/arxiv/pdf/0903/0903.3886v1.pdf].
43. Edwards AWF: **The Measure of Association in a 2 × 2 Table.** *Journal of the Royal Statistical Society, Series A* 1963, **126**:108-114.
44. Olshen AB, Gold B, Lohmueller KE, Struewing JP, Satagopan J, Stefanov SA, Eskin E, Kirchhoff T, Lautenberger JA, Klein RJ, *et al*: **Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping.** *BMC Genet* 2008, **9**:14.
45. Cleveland WS: **Robust locally weighted regression and smoothing scatterplots.** *Journal of the American Statistical Association* 1979, **74**:829-836.
46. **International HapMap Project.** [http://hapmap.ncbi.nlm.nih.gov/].
47. **EIGENSOFT Package.** [http://genepath.med.harvard.edu/~reich/Software.htm].
48. **PLINK Package.** [http://pngu.mgh.harvard.edu/purcell/plink/].
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81(3)**:559-575.
50. **The R Project.** [http://www.r-project.org/].
51. **R: A Language and Environment for Statistical Computing.** [http://www.R-project.org].
52. Rodig H, Grum M, Grimmecke HD: **Population study and evaluation of 20 Y-chromosome STR loci in Germans.** *Int J Legal Med* 2007, **121(1)**:24-27.
53. Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, Entz P, Suk EK, Toliat MR, Klopp N, Caliebe A, *et al*: **SNP-based analysis of genetic substructure in the German population.** *Hum Hered* 2006, **62(1)**:20-29.
54. Chen Y, Lin CHL, Sabatti C: **Volume Measures for Linkage Disequilibrium.** *BMC Genetics* 2006, **7(54)**.
55. Kruglyak L: **Genetic isolates: separate but equal?** *Proc Natl Acad Sci USA* 1999, **96(4)**:1170-1172.
56. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A: **Linkage disequilibrium patterns of the human genome across populations.** *Hum Mol Genet* 2003, **12(7)**:771-776.
57. Bosch E, Laayouni H, Morcillo-Suarez C, Casals F, Moreno-Estrada A, Ferrer-Admetlla A, Gardner M, Rosa A, Navarro A, Comas D, *et al*: **Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: most population isolates do not show increased LD.** *BMC Genomics* 2009, **10**:338.

doi:10.1186/1471-2156-12-67
Cite this article as: Gross *et al.*: Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genetics* 2011 **12**:67.