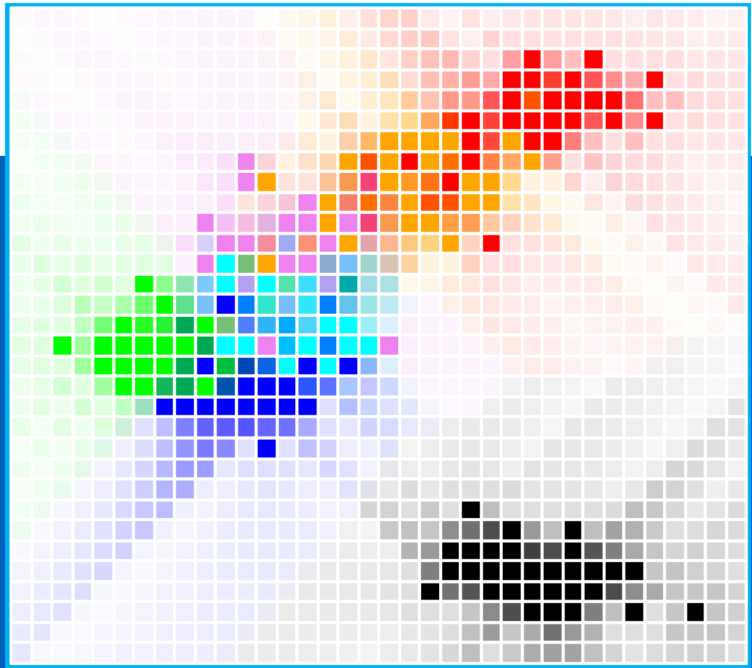
UNIVERSITÄT LEIPZIG

Population-genetic comparison of a German isolated population with a German mixed population on the basis of genome-wide SNP markers



Groß A, Scholz M

26.09.2011

imise.
Institut für Medizinische Informatik, Statistik und Epidemiologie

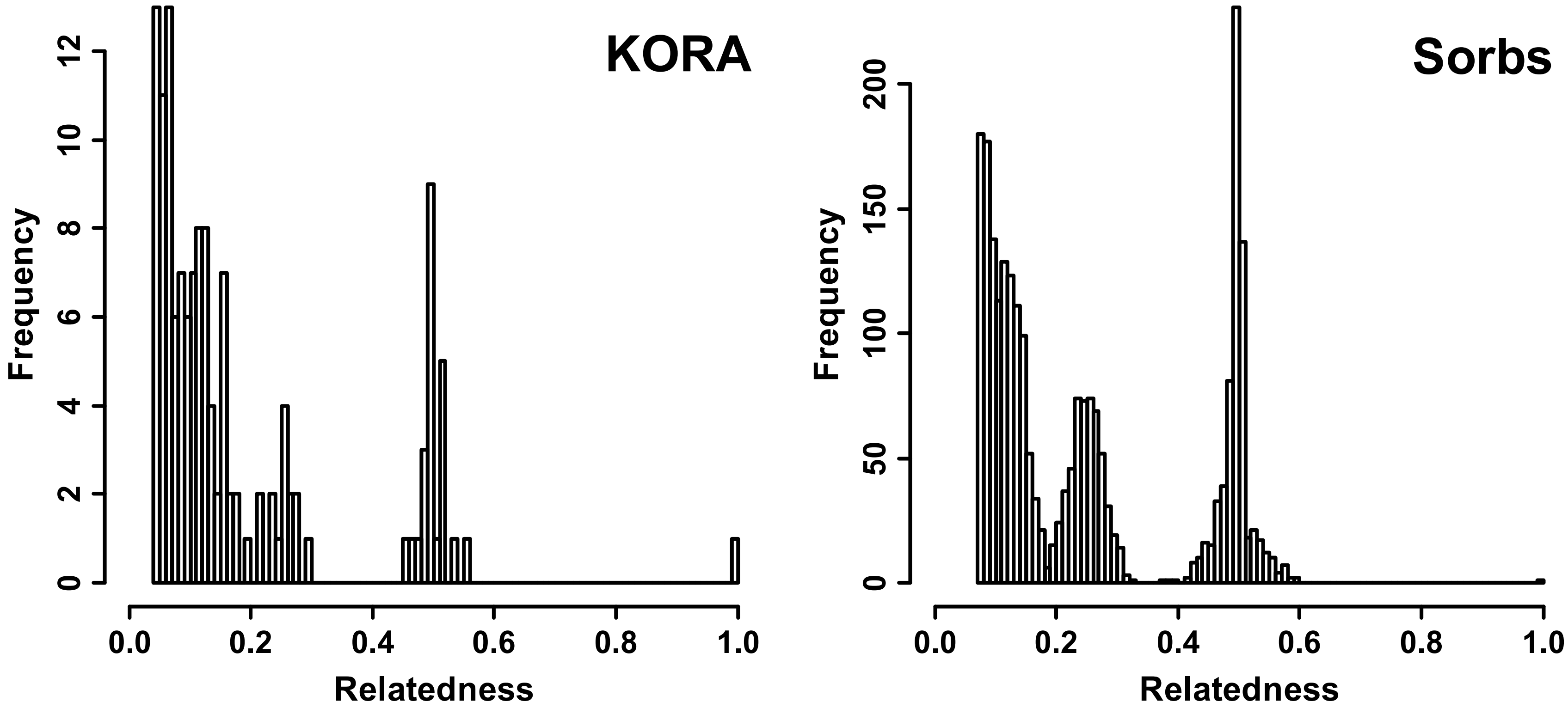
Background

- Isolated populations are assumed to be interesting for disease mapping
- Sample of about 1000 Sorbs is currently analyzed 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 a German outbred population
- Separate effects of oversampling of families in the Sorbs from effects of genetic isolation

Methods

- SNP Genotype Data from Affymetrix 500k and 1000k Arrays
- Study Populations
  - N=977 Sorbs
  - N=1644 KORA Individuals
  - N=198 Hapmap Individuals, N=110 CEU (CEPH from Utah), N=88 TSI (Toscans in Italy)
- Data Analysis
  - Pair-wise relatedness
  - Principal components analysis
  - Rare SNPs
  - F-statistics
  - Runs of homozygosity
  - Linkage disequilibrium
  - Power analysis

Pair-wise relatedness



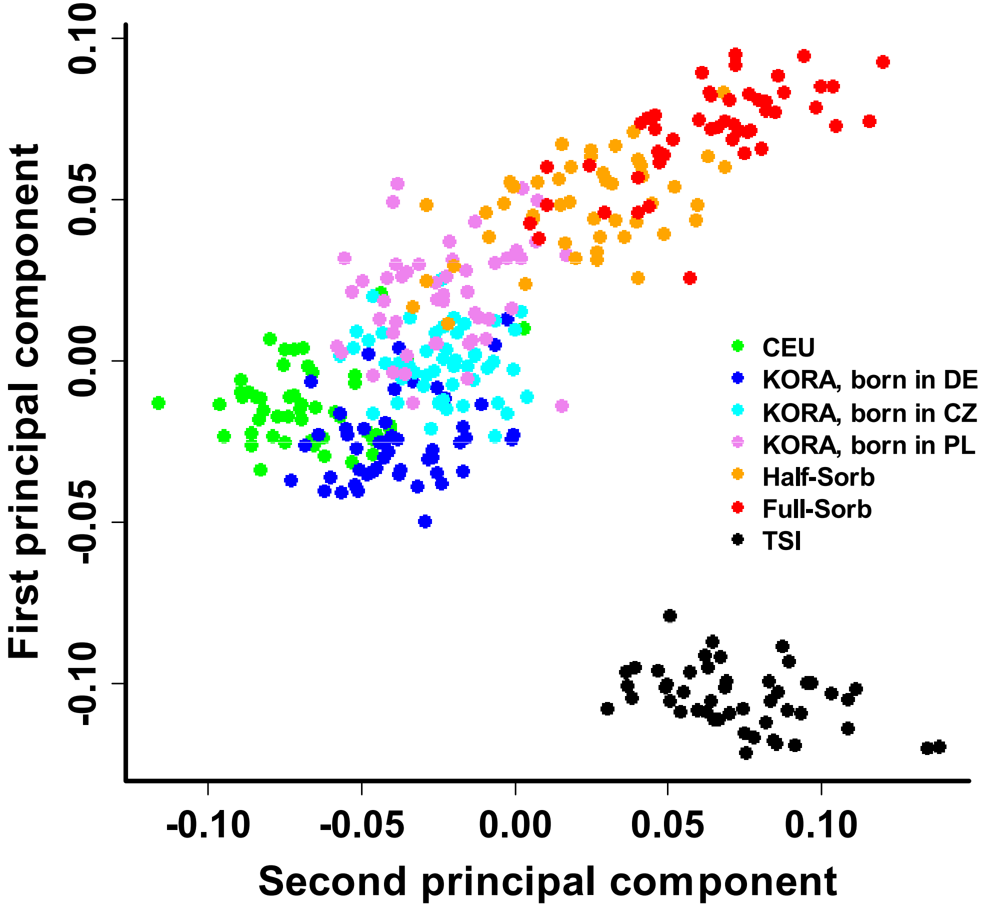
Distribution of degrees of relatedness in the KORA and Sorbs samples.

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.

Subsamples

- For analyses of dependence of measures of population genetic comparison on relatedness, we define two subsamples used for subsequent analyses
  - Complete Sorbs sample ($Sorbs_{977}$, N=977) was matched with a randomly selected subset of N=977 unrelated KORA subjects born in Germany ($KORA_{977}$)
  - N=532 unrelated Sorbs ($Sorbs_{532}$) was matched with a subset of N=532 KORA subjects ($KORA_{532}$) randomly selected from $KORA_{977}$
- 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

Principal components analysis



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 from Utah, N=49) and TSI (Toscans in Italy, N=48).

Rare SNPs

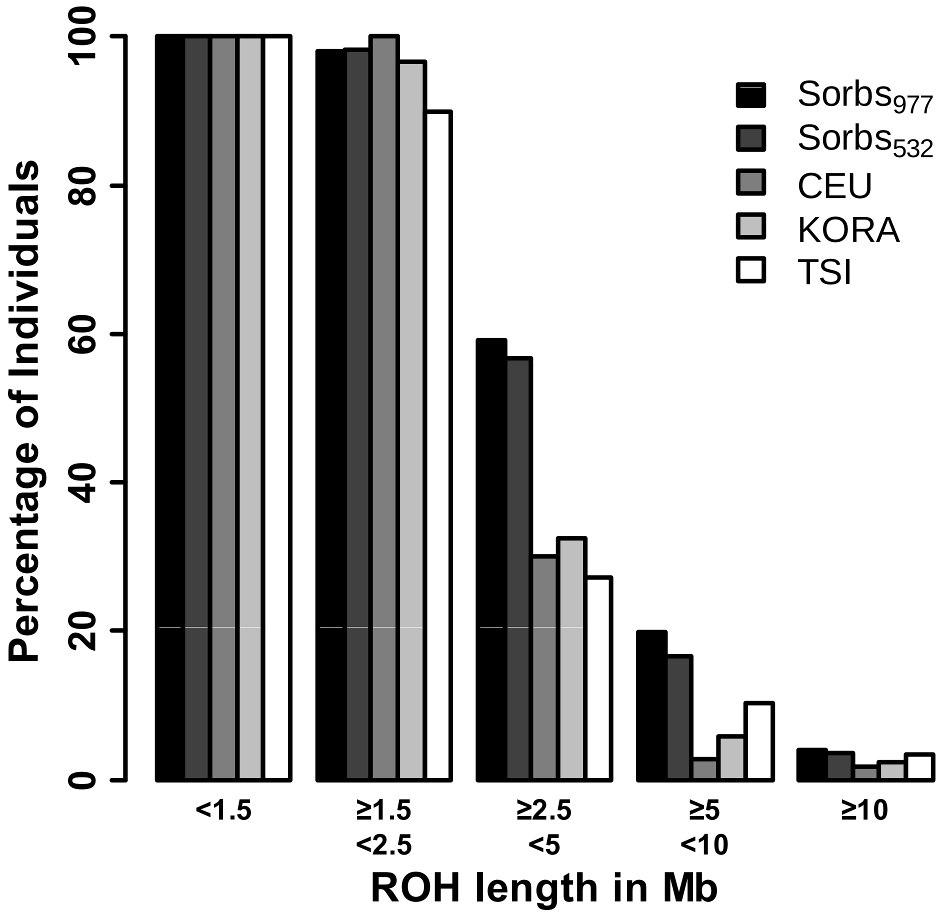
- Consider a SNP as rare if the 95% confidence interval of the minor allele frequency is below 1%
- When analysing 424,476 quality SNPs, we counted
  - 51,204 rare SNPs in Sorbs_{977} and 49,721 rare SNPs in KORA_{977} (p-value 6.7×10^{-7})
  - 49,257 rare SNPs in Sorbs_{532} and 47,913 rare SNPs in KORA_{532} (p-value 4.7×10^{-6})

F-statistics

Population	F-statistic	Estimate	SE
$KORA_{977}$	F_{IS}	0.0012	$2.7x10^{-4}$
$Sorbs_{977}$	F_{IS}	-0.0006	$2.7x10^{-4}$
$KORA_{532}$	F_{IS}	0.0014	$3.5x10^{-4}$
$Sorbs_{532}$	F_{IS}	-0.0002	$3.6x10^{-4}$
$KORA_{977}$,$Sorbs_{977}$	F_{ST}	0.0034	$5.4x10^{-5}$
$KORA_{532}$,$Sorbs_{532}$	F_{ST}	0.0029	$6.7x10^{-5}$

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

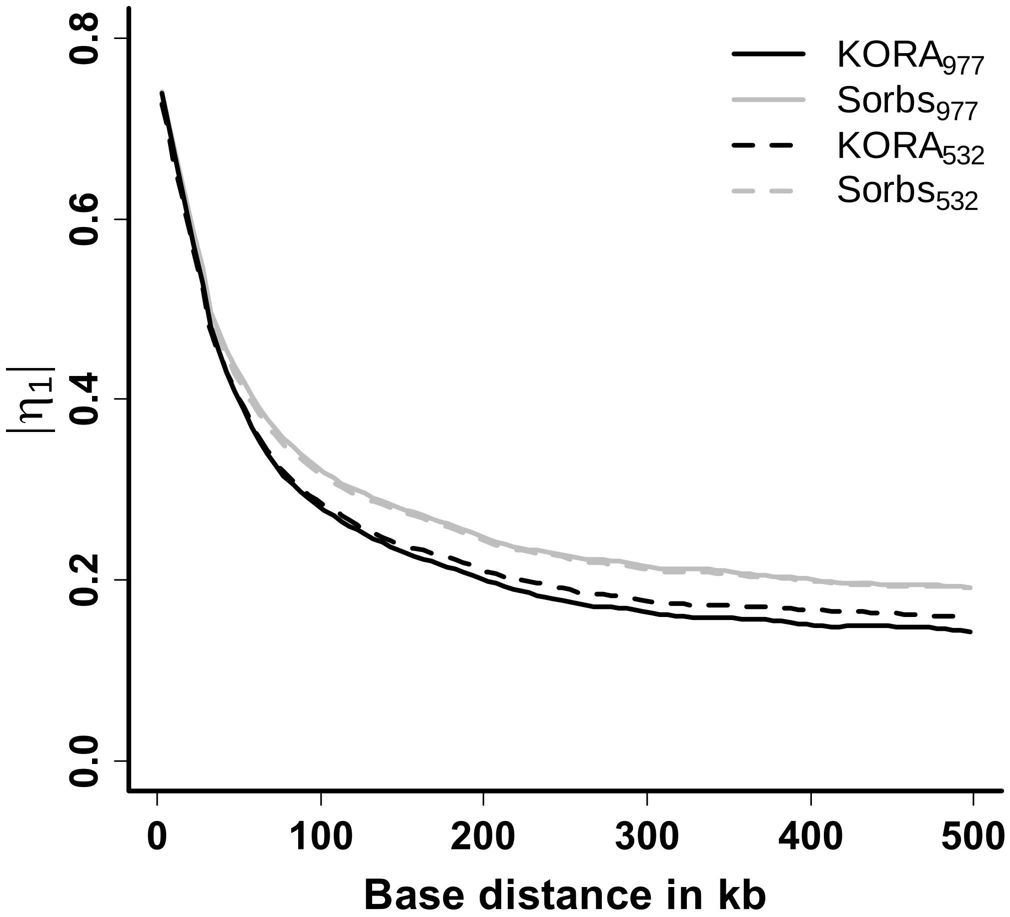
Runs of homozygosity



Proportion of individuals from KORA (N=1644), $Sorbs_{977}$, $Sorbs_{532}$, CEU (CEPH from Utah, N=110) and TSI (Toscans in Italy, N=88) with at least one ROH in the given length interval.

UNIVERSITÄT LEIPZIG

Linkage Disequilibrium



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.

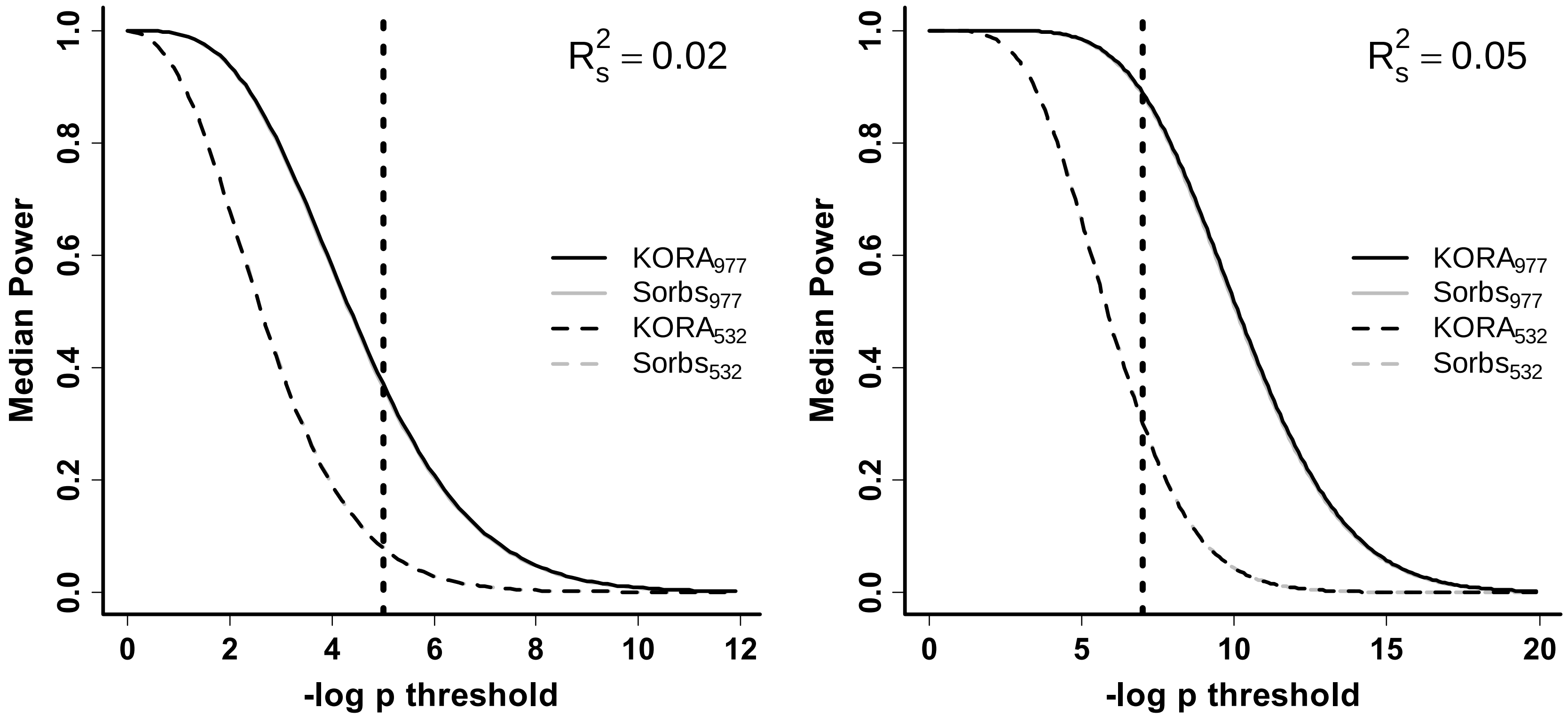
$$\eta_1 = \begin{cases} 2\dfrac{\lambda^2 - \lambda - \lambda \ln \lambda}{(\lambda - 1)^2} - 1 & \text{if } \lambda \neq 1 \\ 0 & \text{if } \lambda = 1 \end{cases}$$

The measure η_1 is a monotone function of the odds ratio ranging between -1 and 1.

Power analysis I

- Assume a linear regression model $\mathbf{y}=\beta_1\mathbf{s}_1+\boldsymbol{\varepsilon}_1$ of a random phenotype y influenced by a genotype $\mathbf{s}_1$ of a causative SNP, and residual error ε_1
- The SNP is assumed to explain a pre-specified proportion of the total variance R_s^2 of the phenotype and assume $\beta_1=1$
- Within the distance of ± 2 Mb analyse 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
- Then $\hat{\beta}_2 \sim \mathrm{N}\left(\frac{\mathrm{Cov}(s_1,s_2)}{\mathrm{Var}(s_2)}, \frac{\frac{\mathrm{Var}(s_1)}{R_s^2}-\frac{\mathrm{Cov}(s_1,s_2)^2}{\mathrm{Var}(s_2)}}{\sum_{i=1}^{n}(s_{2i}-\overline{s_2})^2}\right)$ and depends on $\mathbf{s}_1$, $\mathbf{s}_2$ and R_s^2
- Calculated the power of the regression analysis for all SNPs on Chromosome 22

Power analysis II



Median power to detect SNP effects explaining 2% (left) or 5% (right) of variance.

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.

Conclusions

- The 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

UNIVERSITÄT LEIPZIG

Reference

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

Gross A, Tönjes A, Kovacs P, Veeramah KR, Ahnert P, Roshyara NR, Gieger C, Rueckert IM, Loeffler M, Stoneking M, Wichmann HE, Novembre J, Stumvoll M, Scholz M.

BMC Genet. 2011 Jul 28;12:67.

PMID: 21798003