C. Lamina[1]
M. Steffens[2]
J. Mueller[3]
E. Lohmussaar[3, 4]
T. Meitinger[3]
H. E. Wichmann[1]

Genetic Diversity in German and European Populations: Looking for Substructures and Genetic Patterns

Genetische Diversität in deutschen und europäischen Bevölkerungen: Suche nach Substrukturen und genetischen Mustern

Zusammenfassung

Klassische Fall-Kontroll-Studien sind leistungsfähige und kosteneffiziente Studiendesigns, um Assoziationen zwischen genetischen Markern und Phänotypen genetisch komplexer Krankheiten aufzuspüren. Trotzdem konnte nur ein geringer Anteil von signifikanten Assoziationsergebnissen von anderen Studien repliziert werden. Ein Grund hierfür könnten unerkannte Bevölkerungsstrukturen sein. Das Ziel der Deutschen „Genomic-Control"-Studie ist es, eine mögliche Populationsstratifikation zwischen einer süddeutschen Population, repräsentiert durch den KORA-Survey S4 (1999/2001), und zwei norddeutschen Populationen (SHIP, Greifswald, und POPGEN, Schleswig-Holstein) aufzudecken. Aus KORA S4 wurden bereits Probanden speziell als Kontrollpopulation für genetische Fall-Kontroll-Studien ausgewählt. Daher ist die Kenntnis über die genetische Populationsdifferenzierung innerhalb Deutschlands und in welchem Ausmaß diese Fall-Kontroll-Studien beeinflusst, wichtig für bisherige Projekte und weitere Planungen. In einer anderen Studie, der europäischen LD-Studie, wurde das Muster des Kopplungsungleichgewichtes (LD) im Humangenom zwischen acht verschiedenen europäischen Populationen verglichen. Im Allgemeinen bleiben LD-Muster zwischen den Populationen erhalten. Allerdings wurden in bestimmten chromosomalen Regionen variierende LD-Strukturen beobachtet, was Auswirkungen auf die Feinkartierung von Genen in verschiedenen Populationen haben kann.

Abstract

A classical case-control study is a powerful and cost-efficient approach to detect association of genetic markers with complex disease phenotypes. However, only a small fraction of significant association results has been replicated by other studies. Undetected genetic substructures in the population may be one of the reasons for spurious or biased results. The German "Genomic Control" study aims at detecting genetic differentiation between one Southern German population, represented by the KORA study (KORA Survey S4 (1999/2001)), and two Northern German populations (SHIP, Greifswald, and POPGEN, Schleswig-Holstein). Relevant population-substructures will be assessed, as well as their influence on case-control studies. Since KORA samples are used as controls for different German genetic association studies, the knowledge gained through this Genomic Control project will influence the planning of further genetic association studies. A second project, the European LD study, deals with the detection and comparison of linkage disequilibrium (LD) patterns in the human genome for eight distinct European populations. In general, a conservation of LD patterns across European samples can be observed for most gene regions. However, there are chromosomal regions with variable LD structure which may have implications on the fine-mapping of genes in different populations.

affiliation
[1] GSF National Research Center for Environment and Health, Institute of Epidemiology, Neuherberg, Germany
[2] Rheinische Friedrich-Wilhelms-University, Institute of Medical Biometry, Informatics and Epidemiology, Bonn, Germany
[3] GSF National Research Center for Environment and Health, Institute of Human Genetics, Neuherberg, Germany
[4] University of Tartu, Institute of Molecular and Cell Biology, Estonian Biocentre, Estonia

correspondence
Claudia Lamina · Institute of Epidemiology, GSF – National Research Center for Environment and Health · Ingolstädter Landstraße 1 · 85764 Neuherberg · Germany · E-mail: claudia.lamina@gsf.de

bibliography
Gesundheitswesen 2005; 67 Sonderheft 1: S127 – S131 © Georg Thieme Verlag KG Stuttgart · New York
DOI 10.1055/s-2005-858254
ISSN 0949-7013

Schlüsselwörter
Populationsstratifikation · „Genomic Control" · „Structured Association" · Kopplungsungleichgewicht · „tagSNPs"

Key words
Population stratification · genomic control · structured association · linkage disequilibrium · tagSNPs

Introduction

The detection of the genetic basis of complex human diseases is one of the next big goals in human genetic research. The genetic fraction for most monogenic disorders has already been identified. But the detection of the underlying causes for complex diseases remains a challenge. For many common diseases, like heart attack, stroke or cancer, many different genes as well as environmental factors play a role in the development of these diseases. Case-control-studies are efficient tools to assess association between variants in candidate genes and a disease. A group of affected patients (= cases) is compared to a group of unaffected controls. If the frequency of a specific allele or genotype in the cases is higher or lower than expected by chance, an association is detected. One strength of genetic case-control-studies lies in the fact, that specimens, which have been collected in one survey, can be used as controls to study a variety of gene-disease associations. For example, blood samples from the KORA Survey S4 (1999/2001) are used as controls for different case groups that were collected in different studies and even different regions of Germany. In recent years, though, candidate gene association studies have produced inconsistent results. One explanation for the lack of replication can be found in the possible effect of population stratification [1, 2]. Population stratification occurs, when the population of interest is not homogeneous with respect to allele frequencies, but consists of subgroups, that have different allele frequencies for a gene on the chromosomal region of interest (Fig. **1**). If these subgroups also have different frequencies of a true risk factor, then subgroup membership is a confounder. Not accounting for this confounding factor in the matching procedure of cases and controls, may lead to spurious findings. One popular example for the effect of population stratification is the following [3]: An association can be found between the A1 allele at the HLA-A locus and the ability of eating with chopsticks. But this association would emerge as a statistical artefact if the origins of the study participants were considered. Asian and Caucasian populations differ in frequencies of the A1 allele and also differ in the ability of eating with chopsticks. Thus, there is no association with the allele, but with geographical origin and cultural affiliation. Such strong differentiations are mostly detected, but a weak population substructure may be cryptic and needs some adjustment on a genetic basis.

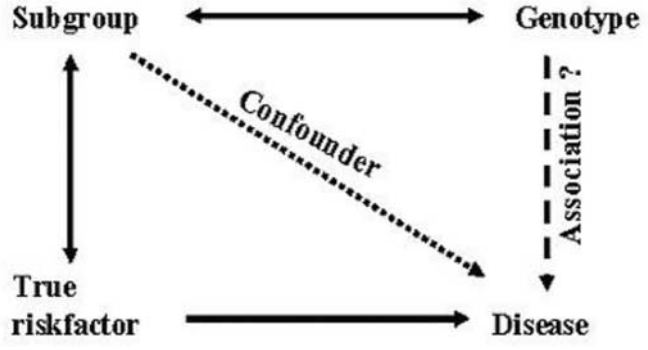

Fig. **1** Schematic model for population stratification: If the study population is divided into subgroups, which differ in frequencies of genotypes and risk factors of disease (full arrows), the subgroup acts as confounding factor (dotted arrow). The association of interest (dashed line) between genotype and disease is biased. (see [16]).

In recent years, the methods proposed for "Genomic Control" (GC) [4, 5] and "Structured Association" (SA) [6] have been developed to identify such cryptic population stratifications and to correct for it in association analyses. For example, the GC-methods have been tested on three US-case-control studies and one from Poland [7]. In the US-Afro-American population, there was a slight indication of stratification. However, it could be eliminated after exclusion of persons, whose parents and grand-parents were not born in the USA. This finding may indicate, that a thorough selection of cases and controls may be sufficient to exclude spurious findings due to population stratification. For Germany, there is little empirical evidence on population substructure and admixture, yet. Although, according to Cavalli-Sforza et al. [8] there is an a priori expectation of some genetic differentiation along a north-south gradient within Germany. If case and control groups are collected in different regions, this fact may become a confounding factor. Then, genetic differences found between case and control groups could be due to population stratification. Data from the KORA survey S4 and two northern German studies (SHIP, Greifswald, and POPGEN, Schleswig-Holstein) are used to assess the extent of population stratification in Germany (German Genomic Control Study).

KORA samples are also used in the European LD-study [9]. The aim of this study is to compare the genetic patterns in several distinct European populations and to evaluate the transferability of genetic information among populations in relevance to association studies.

German Genomic Control Study

In the German Genomic Control Study, the allele and genotype frequencies of three regional German populations are compared (see Fig. **2**): KORA S4 survey from Augsburg, SHIP ("Study of health in Pomerania") from Greifswald (Pomerania), and POPGEN (Population Genetic Cohort), conducted in Kiel and sorrounding counties (Schleswig-Holstein). Sub-samples of about 720 persons have been drawn randomly in each cohort and 210 Single-Nucleotide-Polymorphisms (SNPs) have been genotyped. SNPs are variations on single base pairs in the genomic sequence and can be found approximately at every 300th base pair. 140 SNPs out of the 210 selected SNPs were chosen out of the non-functional region of the genome, the so-called genomic desert. These loci are presumed to be not under selection and thus, they are only subject to neutral processes of drift and migration in demographic history. The other 70 SNPs were chosen out of in-

Fig. **2** Study populations of the Germany study on Genomic Controls (GC) (highlighted): KORA, SHIP and POPGEN (each n = 720) and of the European Study on Linkage Disequilibrium (LD): KORA (n = 170), SHIP (n = 100), POPGEN (n = 160), samples from Estonia (n = 170), South Tyrol (n = 170), Ladinia (n = 160), Brisighella (n = 98) and Calabria (n = 100); the circle represents the emigration region of the HapMap CEPH trios (n = 90), that were used as a reference population. The KORA population is a random sample of the KORA survey S4 (1999/2001).

tragenic regions. All SNPs are uniformly dispersed throughout the genome and have moderate allele frequencies. After the selection process, these SNP loci have been used for the Genomic Control (GC) and Structured Association (SA) methods. Both methods are based on the idea, that population substructure does not only affect the candidate genes, but can be seen across the whole genome. Alleles that are not associated with the disease or candidate genes can be used to estimate the effect of population stratification in the sample and to identify the underlying subgroups. In the method of Genomic Control, for example, a factor is estimated, which is used for the appropriate correction of results derived from case-control association analyses. The genotyping for this study and the quality management is still in progress. First results will be made publicly available in 2005.

European LD-Study

The LD pattern is critical for association studies, in which disease causing variants are identified by allelic association with adjacent markers.

Consider several SNPs in a defined region on one chromosome. It seems obvious, that the probability, that neighboring SNP alleles are inherited together is higher than for SNPs that are further apart. This pairwise association between alleles at two different loci is referred as linkage disequilibrium (LD). LD measures, such as D' or r^2, range from 0 (= no disequilibrium, no association between SNPs) to 1 (= complete LD, association between SNPs). Although LD patterns are rather unpredictable and cannot be inferred simply by distance between the SNPs, there are certain structures of LD in the human genome. It has been shown, that the human genome can be partitioned into blocks of high LD, also called haplotype blocks [10, 11]. These blocks are separated by regions, in which recombination has occurred more often. Within these recombination hot spots, LD is rather low. Understanding the LD patterns and thus the distribution of haplotype blocks in the human genome is crucial for further association studies [12]. For example, consider a genomic region with 31 SNPs (Fig. **3**). This region might be structured by the pattern of high and low LD into 3 haplotype blocks. In the first haplotype block, 7 SNPs could be found, that are in complete LD. That means that they are correlated and information of one SNP is sufficient to infer the information on the others. The second block shows 11 SNPs, that can be reduced to 3 SNPs without loosing much information and so on. Those highly informative SNPs are called tagSNPs or tagging SNPs. Thus, genotyping ressources and costs can be substantially reduced by genotyping only the tagSNPs. In recent years, several groups have been working on the installation of a haplotype map of the human genome. Popu-

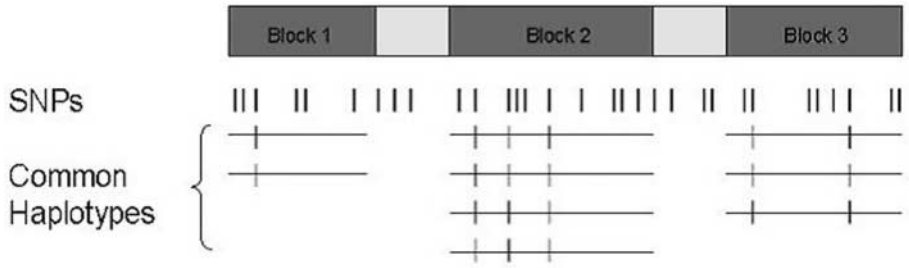

Fig. **3** Partitioning of DNA sequence into 3 haplotype blocks, defined by high Linkage Disequilibrium (LD) between pairs of SNPs, separated by recombination hot spots. In each block, the information of all occurring SNPs can be reduced to several tagSNP. Combination of tagSNP characteristics (black and grey lines) define common haplotypes.

lations from the HapMap Project [13] are suggested as reference populations for the selection of tagging SNPs in association studies. The European patterns within this project are represented by 30 trios from a US Utah population of Northern and Western European ancestry (CEPH samples, [14]). However, the general applicability of the HapMap data has to be confirmed by samples from several local populations.

The aim of the European LD study is to compare the LD patterns in several distinct European populations. Four genomic regions (in total 749 kb) containing candidate genes for different complex traits have been analyzed. Individuals were genotyped for markers that are evenly distributed with an average spacing of ~ 2 – 4 kb in eight population-based samples from ongoing epidemiological studies across Europe plus the CEPH trios of the HapMap project.

In general, a conservation of the LD patterns across European samples could be observed. Nevertheless, shifts in the position of boundaries of high LD regions could be demonstrated between populations, when assessed by a novel procedure based on bootstrapping. Transferability of LD information among populations was also tested. In two of the analyzed gene regions, tagSNPs selected in the HapMap CEPH trios performed surprisingly well in all local European samples. However, significant variation in the two other gene regions predicts a restricted applicability of CEPH derived tagging markers. Simulations based on the observed data sets show the extent by which further gain in tagSNP efficiency and transferability can be achieved by increased SNP density.

Conclusion and Outlook

In recent years, there has been a debate over the impact of population stratification on genetic association studies [15, 16]. It has been shown, that the difference between distinct populations constitute only 3 – 5 % of the overall genetic variation among individuals [17]. Thus, the within-population differences among individuals accounts for the vast majority (93 – 95 %) of genetic variation. Nevertheless, stratification cannot be eliminated in real case-control studies. Even small effects of population stratification can lead to false positive results, especially in studies with large sample sizes, where the power to detect small effects is increased. To avoid this problem, family-based study designs have been proposed. Using non-affected family-members as controls rules out the possibility of population stratification. However, these designs have stronger sampling requirements, are therefore more expensive and lack power to detect small effects. Since there are methods to detect and control for population stratification, there is no need to abandon case-control studies.

KORA samples have been used to answer crucial questions regarding population genetics: Is there genetic differentiation between Northern German and Southern German populations? How big is the influence of a possible population stratification on case-control studies? Do European populations differ with respect to LD patterns, haplotype blocks and tagging SNPs?

The last question has already been answered: For most gene regions, LD patterns are conserved, although LD structures vary in certain chromosomal regions. These findings have to be considered in the fine-mapping of genes in different populations in the future.

Altogether, the answers to these questions close the gap in one's knowledge on population structure in Germany and Europe. Samples from KORA S4 have already been used as universal controls for case-control studies. Since further use is promoted, information on population substructure and genetic patterns in the population will play an important role in future study planning.

Acknowledgement

This work has been supported by the GSF and grants from BMBF – Federal Ministry of Education and Research (NGFN).

The article refers specifically to the following contributions of this special issue of Das Gesundheitswesen: [18 – 21].

References

[1] Freedman ML, Reich D, Penney KL et al. Assessing the impact of population stratification on genetic association studies. Nat Genet 2004; 36: 388 – 393
[2] Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet 2003; 361: 598 – 604
[3] Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994; 265: 2037 – 2048
[4] Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999; 55: 997 – 1004
[5] Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 2001; 60: 155 – 166
[6] Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor Popul Biol 2001; 60: 227 – 237
[7] Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. Am J Hum Genet 2002; 71: 304 – 311

[8] Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes New Jersey USA: Princeton Univ. Press, 1994
[9] Mueller JC, Lohmussaar E, Magi R et al. Linkage Disequilibrium Patterns and tagSNP Transferability among European Populations. Am J Hum Genet 2005; 76 (3): 387 – 398
[10] Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 2003; 4: 587 – 597
[11] Gabriel SB, Schaffner SF, Nguyen H et al. The structure of haplotype blocks in the human genome. Science 2002; 296: 2225 – 2229
[12] Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. Trends Genet 2003; 19: 135 – 140
[13] The International HapMap consortium. The international HapMap project. Nature 2003; 426: 789 – 796
[14] Dausset J, Cann H, Cohen D et al. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics 1990; 6: 575 – 577
[15] 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
[16] 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
[17] Rosenberg NA, Pritchard JK, Weber JL et al. Genetic structure of human populations. Science 2002; 298: 2381 – 2385
[18] Löwel H, Döring A, Schneider A et al. The MONICA Augsburg surveys – basis for prospective cohort studies. Gesundheitswesen 2005; 67 S1: S13 – S18
[19] Holle R, Happich M, Löwel H et al. KORA – A research platform for population based health research. Gesundheitswesen 2005; 67 S1: S19 – S25
[20] Wichmann HE, Gieger C, Illig T et alKORA-gen – Resource for population genetics, controls and a broad spectrum of disease phenotypes. 2005; 67 S1: S26 – S30
[21] Löwel H, Meisinger C, Heier M et al. The population-based Acute Myocardial Infarction (AMI) Registry of the MONICA/KORA study region of Augsburg. Gesundheitswesen 2005; 67 S1: S31 – S37