I. M. Heid[1]
C. Lamina[1]
F. Bongardt[1]
G. Fischer[1]
N. Klopp[1]
C. Huth[1]
H. Küchenhoff[2]
F. Kronenberg[1, 3]
H. E. Wichmann[1]
T. Illig[1]

Wie gut können Haplotypen in den populationsbasierten KORA-Studien rekonstruiert werden?

How About the Uncertainty in the Haplotypes in the Population-Based KORA Studies?

Zusammenfassung

In den KORA-Surveys werden derzeit verschiedene Kandidatengene, die in Zusammenhang mit Typ 2 Diabetes, Herzinfarkt, Atherosklerose, Adipositas und anderen Erkrankungen stehen, untersucht. Hierbei werden SNPs (Single Nucleotide Polymorphisms, Einzelbasenaustausche) in verschiedenen Genen bei den Probanden der Querschnittstudie genotypisiert. Ferner gewinnen Haplotypen an Bedeutung: Haplotypen sind Kombinationen von Allelen innerhalb von bestimmten Abschnitten eines Chromosomenstrangs. Die Betrachtung solcher Haplotypen in genetischen Assoziationsstudien ist oft effizienter als die Betrachtung der einzelnen SNPs. Ein statistisches Problem ist hierbei die Rekonstruktion der Phaseninformation: Bei der Genotypisierung werden nur die Allele (also die Ausprägungen) eines Individuums an den SNPs bestimmt, jedoch nicht, welche Base auf welchem Chromosomenstrang angesiedelt ist. Verschiedene statistische Haplotyp-Rekonstruktionsverfahren ermöglichen die Identifizierung der wahrscheinlichsten Haplotypen. Dabei ist ein gewisser Prognosefehler unausweichlich. Auch Genotypisierungsfehler können zur Unsicherheit in den Haplotypen beitragen. Dieser Genotypfehler kann von Bedeutung werden, selbst wenn der Genotypfehler je SNP sehr klein ist. Dies liegt daran, dass mehrere SNPs an den Haplotypen beteiligt sind. Ein Ziel dieses Projekts ist die Quantifizierung der Haplotyp-Unsicherheiten bei Genen, die in KORA untersucht wurden. Wir verwenden einerseits Computersimulationen basierend auf den in den KORA-Probanden beobachteten Haplotypen und deren Häufigkeiten. Andererseits vergleichen wir Ergebnisse mit Simulationen basie-

Abstract

In the KORA surveys, numerous candidate genes in the context of type 2 diabetes, myocardial infarction, atherosclerosis or obesity are under investigation. Current focus is on genotyping single nucleotide polymorphism (SNPs). Haplotypes are also of increasing interest: haplotypes are combinations of alleles within a certain section of one chromosome. Analysing haplotypes in genetic association studies is often more efficient than studying the SNPs separately. A statistical problem in this context is the reconstruction of the phase: genotyping the SNPs determines the alleles of an individual at one particular locus of the DNA, but does not reveal which allele is located on which one of the two chromosomes. This information is required when talking about haplotypes. There are statistical approaches to identify the most likely two haplotypes of an individual given the genotypes. However, a certain error in prognosis is unavoidable. There are also errors in the genotypes. These errors are assumed to be small for one SNP but can accumulate over the SNPs involved in one haplotype and thus can induce further uncertainty in the haplotype. It is therefore the aim of our project to quantify the uncertainties in the haplotypes particularly for genes investigated in the KORA surveys. We conduct computer simulations based on the haplotypes and their frequencies observed in the KORA individuals and compare the results with simulations based on mathematical modelling of the evolutionary process ("coalescent models"). The uncertainties in the haplotypes have an impact on the search for association between genes and disease: an association may not be detected as the haplotype uncertainty

Institutsangaben
[1] Institute of Epidemiology, GSF National Research Center for Environment and Health, Neuherberg, Germany
[2] Department of Statistics, Ludwig-Maximilians-Universtität München, Germany
[3] Innsbruck Medical University, Department of Medical Genetics, Molecular and Clinical Pharmacology, Division of Genetic Epidemiology, Innsbruck, Austria

Korrespondenzadresse
Iris M. Heid · GSF – Forschungszentrum für Umwelt und Gesundheit, Institut für Epidemiologie · Ingolstädter Landstraße 1 · 85764 Neuherberg · E-mail: heid@gsf.de

Bibliografie
Gesundheitswesen 2005; 67 Sonderheft 1: S132 – S136 © Georg Thieme Verlag KG Stuttgart · New York
DOI 10.1055/s-2005-858253
ISSN 0949-7013

rend auf mathematischen Modellen zur Evolution („coalecent models“). Diese Unsicherheiten in den Haplotypen können dazu führen, dass vorhandene Assoziationen zwischen Gen und Erkrankung nicht gefunden werden, da die Unsicherheit in den Haplotypen den Unterschied der Haplotyp-Häufigkeiten zwischen Erkrankten und Nichterkrankten verwischt. Das Ausmaß dieses Problems und Lösungsmöglichkeiten aufzuzeigen, ist das zweite Ziel dieses Projekts.

Schlüsselwörter
Haplotyp-Rekonstruktion · Prognosefehler · Assoziationsstudien · populationsbasiert

obscures the haplotype frequency differences between cases and controls. It is a further aim of our project to elucidate the extent of this problem and to develop strategies for reducing it.

Key words
Haplotype reconstruction · prognosis error · association studies · population-based

Einleitung

SNPs, Haplotypen und Assoziationsstudien

Neben der Sequenzierung des menschlichen Genoms brachten Hochdurchsatz-Techniken zur Genotypisierung einen Durchbruch in der genetischen Erforschung komplexer Erkrankungen, weil dadurch systematische Analysen auch von großen epidemiologischen Studien möglich wurden. Eine herausragende Fragestellung ist hierbei die Analyse der Assoziation zwischen Erkrankung und Kandidatengenen. Kandidatengene sind Gene, bei denen ein funktioneller Zusammenhang mit der Erkrankung aufgrund verschiedener Kriterien vermutet wird. Hierbei steht also die Bestätigung eines modifizierten Erkrankungsrisikos durch Varianten des Gens und dessen Modellierung im Vordergrund. Solche Assoziationen können in populationsbasierten Querschnittstudien nachgewiesen werden.

Bei der Genotypisierung setzt sich als Labormethode zunehmend die Untersuchung von SNPs (Einzelbasenaustausche, Single Nucleotide Polymorphisms) durch, da diese die am häufigsten vorkommenden DNA-Varianten sind und damit ein genaues Eingrenzen der assoziierten DNA-Varianten möglich ist. Ein SNP ist ein Basenpaar in der DNA, das in der untersuchten Population abweichende Ausprägungen zeigt. Ein SNP muss nicht in kausalem Zusammenhang mit der Erkrankung stehen. Die Assoziation kann auch aufgrund der räumlichen Nähe mit dem unbekannten Erkrankungslokus auftreten. Der SNP mit der beobachteten Assoziation dient dann als Stellvertreter (Marker) für den Erkrankungslokus. Beispielsweise zeigt ein AT-Polymorphismus die Ausprägungen A und T in der Bevölkerung. Da jedem Menschen von seinen beiden Elternteilen je ein Chromosomenstrang vererbt wird, besitzt der Mensch pro SNP zwei Allele, die im Beispiel entweder A oder T sind. Damit gibt es drei Ausprägungsmöglichkeiten des Genotyps für einen SNP, also A/A (homozygot A), A/T (heterozygot) und T/T (homozygot T). Hierbei ist die Reihenfolge ohne Belang, d. h. A/T und T/A sind als Genotyp identisch, da bei der Genotypisierung nicht unterschieden werden kann, welches Allel auf welchem Chromosomenstrang lokalisiert ist. Eine der Ausprägungen könnte mit einer erhöhten Erkrankungswahrscheinlichkeit einhergehen. Wenn die Erkrankung in zwei Ausprägungen (krank, nicht krank) gegeben ist, erfolgt die Risikomodellierung durch Vergleich der Allelhäufigkeiten zwischen Fällen und Kontrollen.

Es ist schwierig, vorliegende genetische Assoziationen in epidemiologischen Studien zu zeigen. Dies kann darin begründet sein, dass die „Power“ von Assoziationsanalysen einzelner SNPs mit einer Erkrankung nicht ausreichend war. Werden viele SNPs untersucht, entsteht das Problem des multiplen Testens bzw. das Problem von vielen Parametern im Modell.

Um mehr Information in den Assoziationsanalysen zu erhalten, multiples Testen zu vermeiden und die Power zu verbessern, wird mehr und mehr Augenmerk auf Haplotypen gelegt. Die Kombination von Allelen auf einem Chromosomenstrang nennt man Haplotyp.

Das zentrale Problem hierbei ist, dass durch die Genotypisierung im Allgemeinen nur festgestellt werden kann, welche beiden Allele ein Individuum an einer Stelle (Locus) der DNA aufweist; es kann jedoch nicht festgestellt werden, auf welchem Chromosomenstrang sich die jeweiligen Allele befinden. Der Genotyp ist also beobachtbar, aber nicht die sogenannte Phase. Betrachtet man nur einen SNP, ist die Phase nicht von Belang. Betrachtet man jedoch zwei oder mehr SNPs, kann der Haplotyp oft nicht eindeutig bestimmt werden, wie das folgende Beispiel zeigt: Ein Proband sei heterozygot für beide betrachteten SNPs (ein A/T-Polymorphismus und ein C/G-Polymorphismus):

Chromosomenstrang I .... A T T C G [T] C C [C] T A T T T A. ....
Chromosomenstrang II .....A T T C G [A] C C [G] T A T T T A. .....

Die beobachtbaren Genotypen an den beiden SNPs sind also A/T und C/G, wobei die Reihenfolge hier keine Information beinhaltet. Betrachtet man die oben dargestelle DNA-Sequenz, dann ergeben sich aufgrund der Genotypen des Individuums zwei mögliche Haplotyp-Paare:

Chromosomenstrang I ....[A T T C G T C C C T A T T T A.]....
Chromosomenstrang II .....[A T T C G A C C G T A T T T A.] ....

oder

Chromosomenstrang I ....[A T T C G A C C C T A T T T A.]....
Chromosomenstrang II .....[A T T C G T C C G T A T T T A.] .....

Die insgesamt vier möglichen Haplotypen an den beiden Loci können mit AC, AG, TC und TG bezeichnet werden, wobei der Unterstrich indiziert, dass nun die Reihenfolge von Bedeutung ist. Im Allgemeinen gibt es also für k SNPs maximal 2^k mögliche Haplotypen. Ist der Proband homozygot in einem der SNPs (z. B. A/A für den ersten SNP), reduziert sich die Anzahl der Möglichkeiten um

den Faktor 2. Für den Probanden aus dem Beispiel ist das Haplotyp-Paar (AC, TG) oder (AG, TC) möglich. Allgemein kann das Haplotyp-Paar für Probanden, die heterozygot an mehr als einem SNP sind, nicht eindeutig durch die Genotypen bestimmt werden.

Methoden und erste Ergebnisse

Haplotyp-Rekonstruktion

Die exakten Haplotypen im Labor festzustellen, ist sehr zeitaufwändig und kostenintensiv und deshalb nur bei wenigen Probanden praktikabel. Ein statistisches Problem ist also die Haplotyp-Rekonstruktion aus vorgegebenen Genotypen: Die Identifizierung des wahrscheinlichsten Haplotypen eines Individuums, dabei gegeben die Genotypen der SNPs für das Individuum. Die Komplexität der Rekonstruktion (a) steigt mit der Anzahl der Loci, (b) mit dem Anteil an heterozygoten Probanden, (c) ist invers zur Allelhäufigkeit und (d) invers zum Grad der gemeinsamen Vererbung der beteiligten SNPs. Verschiedene statistische Verfahren ermöglichen die Identifizierung der wahrscheinlichsten Haplotypen. Ein Maximum-Likelihood-Verfahren unter Verwendung des E-M-Algorithmus (expectation-maximization) wird in [1] beschrieben. Ferner steht Software (PHASE, [2, 3]) zur Verfügung, die auf einem Bayes-Verfahren beruht. Die Autoren der Originalpublikationen führen teilweise Simulationen durch, um ihre Verfahren zu testen.

Prognosefehler

Ein Prognosefehler von 20% wird in [1] beschrieben. Die Entwickler von PHASE sprechen davon, diesen Prognosefehler mit ihrem Verfahren um bis zu 50% senken zu können [2]. Aber ein direkter Vergleich der Verfahren miteinander und eine genaue Analyse des Prognosefehlers unterbleiben. Sehr interessant ist ein Ansatz, der Ergebnisse von Computersimulationen zur Schätzung des Prognosefehlers durch E-M-Haplotyp-Rekonstruktion mit dem Stichprobenfehler in Abhängigkeit von verschiedenen Parametern (Stichprobengröße, Anzahl der untersuchten Loci, Allelhäufigkeit und Abweichung vom Hardy-Weinberg-Equilibrium) vergleicht [4].

Wir verwenden einerseits Computersimulationen basierend auf den in den KORA-Probanden geschätzten Haplotypen und deren Häufigkeiten. Andererseits vergleichen wir Ergebnisse mit Simulationen basierend auf dem Koalescent-Modell unter Verwendung des Programms „ms“ [5], wobei 10 000 Chromosomenabschnitte mit zufälligen Mutationen simuliert werden.

Vorläufige Ergebnisse unserer Untersuchungen zeigen deutlich kleinere Fehlklassifikationsraten als die, von denen die Originalautoren berichteten, wenn realistische Szenarien simuliert werden, die sich an den in KORA-Probanden untersuchten Genen anlehnen. Es zeigt sich allerdings auch, dass die Fehlerraten von Gen zu Gen sehr unterschiedlich sein können, je nachdem wie hoch die SNPs korreliert und wie groß die Allelfrequenzen sind. Deshalb sind Strategien zur Reduktion des Haplotyp-Rekonstruktionsfehlers für einige Gene eher notwendig als für andere.

Fehlklassifikation des Genotyps

Nicht nur die Prognosefehler, auch die Fehlklassifikation des Genotyps beeinträchtigt die Aussagekraft der Haplotyp-Assoziationsanalyse. Fehlerquellen beinhalten u.a. Probenvertauschung und Genotypisierungsfehler.

Es wurden einige Fehlermodelle für den Genotypsierungsfehler beschrieben [6, 7]. Wir setzen für unsere Untersuchungen folgendes dreistufiges Fehlermodell an:

1. unabhängige Fehlklassifikation: Der Fehler hängt nicht von der zu typisierenden DNA ab, d.h. die Wahrscheinlichkeit, dass bei einem A/C-Polymorphismus das A-Allel falsch bestimmt wird, ist gleich der Wahrscheinlichkeit, dass das C-Allel falsch bestimmt wird;
2. allelabhängige Fehlklassifikation: Der Fehler hängt von dem zu typisierenden Allel ab;
3. genotypabhängige Fehlklassifikation: Der Fehler hängt von dem zu typisierenden Genotyp ab. Dies ist ein sehr allgemeines Modell, das durch sechs unterschiedliche Parameter in einer 3-×-3-Fehlklassifikationsmatrix beschrieben werden kann. Hierbei können verschiedene Spezialfälle, die teilweise in der Literatur beschrieben wurden, abgebildet werden.

Um ein gültiges Fehlermodell aufzustellen und Aussagen über die Fehlklassifikationswahrscheinlichkeiten machen zu können, muss der Messprozess der Genotypisierung analysiert werden. Bei der an der GSF etablierten Genotypisierungsmethode MALDITOF-MS (matrix-assisted laser desorption ionization-time-of-flight mass spectrometer by Sequenom, San Diego, USA) werden bestimmte DNA-Abschnitte eines Probanden vervielfältigt, sodass sich zum Schluss genau die Ausprägungen des einen zu untersuchenden SNP derart vervielfacht haben, dass deren Masse im Massenspektrometer gemessen werden kann. Ein solches Bild wird an zwei auf der X-Achse definierten Stellen auf das Vorhandensein eines Signals ausgewertet, wie in diesem Beispiel ein Proband, der heterozygot für einen A/C-Polymorphismus ist:

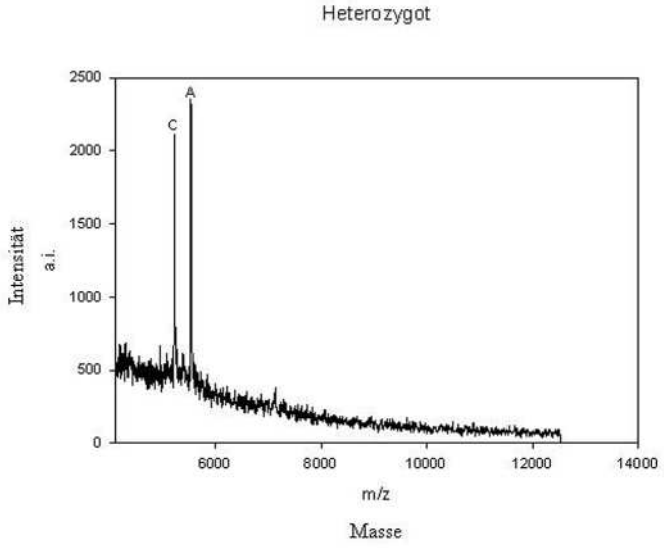

Abb. **1** Ergebnis der Genotypsierung durch MALDITOF-MS für eine Person, die heterozygot für einen A/C-Polymorphismus ist. Es werden zwei Signale detektiert, die jeweils das „C“ und das „A“-Allel an der untersuchten Stelle der DNA nachweisen.

Im Allgemeinen bedeutet ein einziges Signal an der ersten bzw. zweiten Stelle, dass der Proband das Allel C bzw. A doppelt besitzt (also homozygot C bzw. A ist); je ein Signal an den beiden Stellen, wie im Bild oben, zeigt einen heterozygoten Probanden an. Nicht zu vermeiden ist zufälliges Hintergrundrauschen, das zu Fehlern führen kann. Ein Signal wird dadurch identifiziert, dass die Amplitude eine gewisse Schwelle überschreitet. Die Amplitude des einen Signals bei homozygoten Personen ist höher als die Amplitude der zwei Signale bei heterozygoten Personen. Ein vorhandenes Signal kann fälschlicherweise im Hintergrundrauschen untergehen. Es ist im Fehlermodell zu berücksichtigen, (a) dass die Fehlklassifikation von heterozygoten Probanden als falsch homozygot (ein Signal wird übersehen) größer ist als die Fehlklassifikation von homozygoten Probanden als heterozygot (ein zweites Signal wird fälschlicherweise identifiziert) und (b) dass die Wahrscheinlichkeit homozygote Probanden als falsch homozygot (z.B. A/A statt C/C) zu kodieren gleich null ist.

Zur Größe der Fehlklassifikation bei gängigen Genotypisierungsmethoden gibt es in der Literatur keine Angaben. Viele theoretische Simulationen über Genotypisierungsfehler arbeiten mit Fehlklassifikationsraten von 5%. Das erscheint aufgrund der in unserem Labor durchgeführten Qualitätskontrollen als unrealistisch hoch: Bei jeder Genotypisierung werden neben Positiv- und Negativkontrollen 10% der Proben doppelt analysiert. Vorläufige Auswertungen zeigen eine Fehlklassifikationsrate von ca. 0,5%. Die Auswertung von Wiederholungsmessungen ist das Fundament der Charakterisierung und Quantifizierung des Genotypfehlers in diesem Projekt.

Zukünfige Arbeiten

Effekte von Unsicherheiten in den Haplotypen

Eine Fehlklassifikation des Geno- oder Haplotyps, die nicht in der Analyse korrigiert wird, kann zu einer Fehlschätzung eines Erkrankungsrisikos führen [8]. Ob dies eine Unterschätzung oder Überschätzung ergibt, hängt von dem zugrunde liegenden Fehlermodell ab (z.B. [9, 10]). Durch Berücksichtigung der Fehlklassifikation in der Analyse kann die Verzerrung behoben und das Signifikanzniveau wiederhergestellt werden [11, 12]. In dem hier beschriebenen Zusammenhang ist eine Kernfrage, inwieweit der Genotypfehler von dem Genotyp bzw. von dem Allel abhängt.

Der Effekt des Genotypisierungsfehlers auf verschiedene Aspekte bei Familienstudien wurde bereits beschrieben (z.B. [13, 14]). Familien bieten den Vorteil, dass unmögliche Allelkombinationen in den Stammbäumen erkannt werden können. Dafür gibt es bereits Verfahren und Software (PedCheck [15]). Bei populationsbasierten Studien ist eine solche Kontrolle nicht möglich, da im Allgemeinen keine Information über Eltern oder Geschwister vorhanden ist. Insofern sind bei Studien wie den KORA-Querschnittstudien einerseits Genotypfehler nicht als Mendelfehler erkennbar, anderseits ist mit größerem Prognosefehler zu rechnen, da in der Haplotyp-Rekonstruktion bei Teilnehmern in populationsbasierten Studien keine Stammbauminformation einbezogen werden kann.

Die Erfassung und Beschreibung der Unsicherheiten in den Haplotypen und die Entwicklung von Strategien zum Umgang mit den Unsicherheiten in den Haplotypen sind also gerade für die genetischen Fragestellungen in populationsbasierten Studien wie den KORA-Querschnittstudien von besonderer Bedeutung.

Danksagung

Wir möchten uns bei allen Mitarbeitern der KORA-Studiengruppe, des Genotypsierungslabors und der Genetischen Statistik bedanken, durch deren Kooperation dieses Vorhaben ermöglicht wird.

Methodisch-genetische Untersuchungen mit MONICA/KORA wurden gefördert durch die GSF, das BMBF – Bundesministerium für Bildung und Forschung (NGFN, 01GR0464 TP8.3, 01GR0464 TP8.8) und die DFG – Deutsche Forschungsgemeinschaft (SFB 386 TPB10).

Der Artikel nimmt besonderen Bezug auf folgende Beiträge dieser Sonderausgabe von Das Gesundheitswesen: [16 – 21].

Literatur

[1] Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evolution 1995; 12: 921 – 927
[2] Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001; 68: 978 – 989
[3] Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstructionfrom population genotype data. Am J Hum Genet 2003; 73: 1162 – 1169
[4] Fallin D, Cohen A, Essioux L et al. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 2001; 11 (1): 143 – 151
[5] Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 2002; 18 (2): 337 – 338
[6] Akey JM, Zhang K, Xiong M et al. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet 2001; 68: 1447 – 1456
[7] Gordon D et al. A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet 2001; 69: 371 – 380
[8] Bross I. Misclassifikation in 2 × 2 tables. Biometrics 1978; 10: 478
[9] Rubin T, Rosenbaus AB, Cobb S. The use of interview data for the detection of associations in field studies. J Chronic Diseases 1956; 4: 253 – 266
[10] Wacholder S, Dosemeci M, Lubin JH. Blind assignment of exposure does not always prevent differential misclassification. Am J Epidemiol 1991; 1134: 433 – 437
[11] Duffy SW, Rohan TE, Day NE. Misclassification in more than one factor in a case-control study: a combination of Mantel-Haenszel and maximum likelihood approaches. Stat Med 1989; 8: 1529 – 1536
[12] Kaldor J, Clayton D. Latent class analysis in crhonic disease epidemiology. Stat Med 1985; 4: 327 – 335
[13] Ott J. Linkage analysis with misclassifiication at one locus. Clin Genet 1977; 12: 119 – 124 [erratum in Clin Genet 1977; 12: 254]
[14] Göring HHH, Terwilliger JD. Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. Am J Hum Genet 2000; 66: 1095 – 1106
[15] O'Connell JR, Weeks DE. PedCheck: a program for identification of genotyping incompatibilities in linkage analysis. Am J Hum Genet 1998; 63: 259

[16] Löwel H, Döring A, Schneider A et al. The MONICA Augsburg surveys – basis for prospective cohort studies. Gesundheitswesen 2005; 67 S1: S13 – S18
[17] Holle R, Happich M, Löwel H et al. KORA – A research platform for population based health research. Gesundheitswesen 2005; 67 S1: S19 – S25
[18] Wichmann HE, Gieger C, Illig T et al. KORA-gen – Resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 2005; 67 S1: S26 – S30
[19] 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
[20] Illig T, Bongardt F, Schöpfer-Wendels A et al. Genetics of type 2 diabetes: impact of Interleukin-6 gene variants. Gesundheitswesen 2005; 67 S1: S122 – S126
[21] Lamina C, Steffens M, Mueller J et al. Genetic diversity in German and European populations: looking for substructures and genetic patterns. Gesundheitswesen 2005; 67 S1: S127 – S131