**H.-E. Wichmann**
**C. Gieger**
**T. Illig**
**for the MONICA/KORA Study Group**

# KORA-gen – Resource for Population Genetics, Controls and a Broad Spectrum of Disease Phenotypes

*KORA-gen – Ressource für Bevölkerungsgenetik, Kontrolle und ein breites Spektrum an Krankheitsphänotypen*

## Zusammenfassung

KORA-gen ist eine Ressource für die genetisch-epidemiologische Forschung, die sich auf die KORA-Plattform (Kooperative Gesundheitsforschung in der Region Augsburg) stützt. Bioproben sowie phänotypische Information und Daten zu Umwelteinflüssen von ca. 18 000 Erwachsenen aus Augsburg und Umgebung sind vorhanden. Das Alter der Teilnehmer betrug bei der Rekrutierung 25 bis 74 Jahre und liegt im Jahr 2005 bei 30 bis 90 Jahren. KORA-gen kann auch von externen Partnern zur Erforschung genetischer Krankheitsursachen genutzt werden. Interessenten haben die Möglichkeit, sich interaktiv im Internet einen Überblick über die vorhandenen Daten und die Nutzungsregeln zu verschaffen. Die Genotyp-Datenbank steht allen beteiligten Partnern gemeinsam zur Verfügung.

### Schlüsselwörter

Epidemiologie · Genetik · Surveys · Kohorten · Bevölkerungsstudien

## Abstract

KORA-gen is a resource for genetic epidemiological research, based on the KORA platform (Cooperative Health Research in the Region of Augsburg). Biosamples and phenotypic characteristics as well as environmental parameters of 18,000 adults from Augsburg and the surrounding counties are available. The age range of the participants was 25 to 74 years of recruitment and is 30 to 90 years in 2005. KORA-gen can be used by external partners. Interested parties can inform themselves interactively via internet about the available data and the rules of access. The genotypic data base is a common resource of all partners.

### Key words

Epidemiology · genetics · surveys · cohorts · population studies

## What is needed for successful research in genetic epidemiology

Basic tools to perform genetic epidemiology are detailed information about the disease phenotype, and biological samples for molecular research. However, these minimum requirements might be not sufficient for successful results. Genetic epidemiology mainly deals with complex diseases. On the one hand, most of them are relatively frequent in the population. On the other hand, they are polygenic in origin, with several genes involved, in addition to environmental influences like lifestyle and contact to toxic or carcinogenic substances. Typically, the contribution of single genes as well as of single external risk factors is small to moderate. Therefore, in addition to large sample numbers a

high quality of data as well as sound epidemiological methodology are prerequisites of successful research. This is independent of the study type used (e. g. case control study, case only study, cohort study, family-based linkage or association studies).

To study the genetic basis of complex diseases, typically thousands rather than hundreds of patients with the disease of interest are needed. Therefore, the most common way is to recruit them via hospitals or doctor's offices. However, the quality of phenotypic characterization might be variable. It is relatively easy to recruit large numbers if the standards for diagnostic criteria are low, but if well defined patients are needed, the available numbers are much smaller. To find a strong genetic effect, a more crude phenotypic information might be sufficient, but for weak effects, thorough phenotypic characterization is important.

The simplest way of phenotyping for the disease of interest is to take doctors' diagnoses from the records of the patients. However, it is quite clear for many diseases that under the umbrella of one diagnosis several distinct pathophysiological entities are summed up, which might have a different genetic origin. Therefore it is crucial for genetic epidemiology to be based on well defined, standardized phenotyping, which then allows us to define subgroups of patients that have more specific clinical or subclinical properties. This makes it necessary to use specific laboratory parameters as well as refined diagnostic tools.

Another strong argument for the importance of sound phenotyping is the matter of intermediate phenotypes. Many parameters show an early reaction of the body, and they are associated with the disease after its manifestation. They do, however, not yet represent a status of disease. These parameters like specific IgE, elevated cholesterol, elevated C-reactive protein or obesity are often called intermediate phenotypes. The chance of success might be higher if we try to identify genes that influence intermediate phenotypes instead of trying to identify the genetic influence on an endpoint which may be influenced by dozens of intermediate phenotypes and therefore hundreds of genes.

As long as genetic epidemiology is „only" interested in the identification of genes, most environmental factors may be ignored. However, the situation is different if one is interested in gene environment interaction. In complex diseases, it is likely that a combination of genes predisposing for the disease and environmental factors exacerbating the impact of these genes are jointly responsible for disease development in populations. In addition, environmental factors which seem to have only a moderate impact at the population level might have larger relative risks in subpopulations with certain genetic predispositions. Classical epidemiology has always been dealing with these „environmental" risk factors, but only today we are able to combine knowledge on the genetic background with classical epidemiological research, and we have tools to investigate the interaction of genes and the environment whose applications help to understand diseases [1].

## International development of biobanking

In many countries large biobanks with more than 50,000 participants are planned or have already been established (planned numbers in brackets). The first national biobank of this size was realized in Iceland (270,000; [2]), followed by the Estonian project (100,000; [3]). The planning of the UK Biobank (500,000, [4]) is advanced, the Japanese biobank (500,000; [5]) has been started, and now in the US a biobank also is discussed (500,000 or more; [6]). As one example the UK biobank will be described in the following.

The aim of the UK Biobank is to investigate the separate and combined effects of genetic and environmental factors (including lifestyle, physiological and environmental exposures) on the risk of common multifactorial diseases of adult life. At least 500,000 men and women aged 45 to 69 years from the general population of the United Kingdom will be investigated prospectively. People registered with participating general practices will be asked to join the study by completing a questionnaire and attending an interview and examination by a research nurse. Moreover, they will be asked to give a blood sample and to provide written consent for participation and follow-up. Follow-up information on cause-specific mortality and cancer incidence will be obtained from the office of national statistics. Data regarding incident morbidity will be received via regular follow-up hospitalization data and general practice records with confirmation of diagnoses using standard criteria. Every second year a subset of around 2,000 participants will be re-surveyed to allow for correction of regression dilution and the entire cohort will be re-surveyed by postal questionnaire at 5 years to update exposure data and to ascertain self-reported incident morbidity. The UK Biobank is funded by the Wellcome Trust, together with the Medical Research Council (MRC) and the Department of Health. The Coordinating Center is in Manchester and there are several regional centers all over the country. Pilot studies are being performed now, and it will last several years until UK Biobank can be used [1, 4].

To foster collaboration between researchers in the field of population genomics, an international consortium called Public Population Project in Genomics (P3G) has been founded. The aim of P3G is the establishment of standards, nomenclatures, communication tools and sharing of technological know-how. This will allow efficient sharing of data between projects and with the international human genetics community. Ultimately, it favours the study of important biomedical research questions that are beyond the scope of a single effort [7].

It should also be mentioned that there are many scientific arguments for large biobanks [8, 9]. However, some scientists have concerns about the upcoming boom. The databases will only be as good as the individual clinical and exposure information they contain. Opinions vary whether the standard procedure – a routine examination and a patient's health record – is sufficient. A more fundamental critique is, that we already know that most variation in human disease is due to diet and lifestyle factors, and quantifying how the risks vary with one's genetic make-up

usually won't change the solution: encouraging healthier lifestyles. Thus, the enormous investment in genomic medicine might divert resources from prevention [10, 11].

## Situation in Germany – POPGEN and KORA

In Germany promising developments took place recently. In the past, the ethical rules of genetic epidemiological studies were strongly dependent on the local ethics committees and could be quite restrictive. Meanwhile, the situation is better since common rules have been agreed upon [12, 13]. The situation has improved further in 2004, when the German National Ethics Council has published its Opinion on Biobanks for Research. In this Opinion new and research-friendly proposals for ethical regulations are made. It is suggested that in the future it should be possible to perform research without informed consent if the samples and data are anonymized or pseudonymized. It should be possible that the donor gives generalized consent for medical research, including unlimited storage. The use of old collections of biosamples should be possible under specific conditions [14].

Two major German biobanking activities are ongoing. In Northern Germany POPGEN is being established. The concept of POPGEN is to recruit patients for 8 selected diseases in the age range of 18 to 80 years in practices and outpatient offices where the patients are identified from treatment records and diagnoses on the file. The diagnoses will be verified on the basis of available documentation. It is planed to recruit 15,000 patients affected by one of eight disease phenotypes, and a random sample of 10,000 controls. Regular follow-up of 50% of the patients is planned [15].

In Southern Germany, KORA [16, 17] has been used for collaborative genetic epidemiological research, mostly within the National Genome Research Network (NGFN, [18]) since 2001. From altogether 4261 participants of the last KORA survey S4 (S2000), 2,200 probands have been selected as a control pool for genetic analyses. If more subjects were needed the whole study group would be at hand. The control population pool has been used for more than 30 studies on cardiovascular diseases, obesity, type 2 diabetes, allergies, asthma, neurologic disorders, different forms of cancer, and rare mendelian diseases, as well as for projects dealing with population genetics. Here, KORA S4 has served to look for population stratification within Germany and for linkage disequilibrium patterns among European populations [19–32].

## Future development – KORA-gen

In the framework of MONICA (Monitoring of trends and determinants in cardiovascular diseases) and KORA (Cooperative Health Research in the Region of Augsburg), four large population-based cross-sectional studies have been carried out, and a biological specimen bank was established in order to enable the KORA researchers to perform epidemiologic research with respect to molecular and genetic factors. The KORA study center conducts regular follow-up investigations and has collected a wealth of information on sociodemography, general medical history, envir-

onmental factors, smoking, nutrition, alcohol consumption, and various laboratory parameters. This unique resource will be increased further by follow-up studies of the cohort.

Now this collection will be opened also for external researchers, under the name KORA-gen [33]. The objective of KORA-gen is to provide access to information about available population controls for genetic studies as well as provision of DNA samples, genotypic and phenotypic data. The KORA-gen infrastructure will be instrumental in questions of study design, sampling, and matching, of DNA handling and determination of genetic markers, and of data structures and formats. This will be supported by an internet based information resource and by providing competent individual counselling and assistance.

A web-portal for genetic control populations will be available. This portal will resemble modern online libraries that can be searched in great detail. Partners can choose genetic controls based on age, sex and basic phenotype information. This automatic pre-filtering will allow a more informed choice of controls that will further be detailed through individual and person-based counselling.

The biological samples can be genotyped directly at the GSF facilities. Genotyping also can be performed at other genotyping centers which fulfill predefined quality and security standards. There will be an amount-dependent fee per sample. The underlying KORA-gen database will be fed with all relevant information and linked with the KORA database of the GSF. The static phenotype and biosamples' databases will be transformed into a dynamic online system. All genotypes are fed back into a common database for permanent storage. This gradual accumulation will add significant value to the overall dataset. In keeping and administrating the central KORA-gen data base, the GSF acts as a trustee. For all genetic and phenotype data items ownership is defined to those scientists who provided these data. Access to these data for scientific analyses is only granted with permission of the data owners. Rules for data ownership and data access have been formulated and documented as Standard Operating Procedures (SOPs). The genetic database will adhere to existing standards for data communication with other partners.

KORA-gen provides data and biosamples of about 18,000 adults from the general population. It is based on 4 surveys of 4,000–5,000 participants each, performed in the city of Augsburg and the two neighbour counties, with a population of 600,000 inhabitants [17]. The age range was 25 to 74 years at recruitment and is 30 to 90 years in 2005. The first three surveys have been part of the WHO MONICA project. The number of participants is shown in Table **1**.

The available data and biosamples are described in Table **2**. However, not all parameters are available for all participants.

For the surveys S1 to S3 two follow-up interviews with self-administered questionnaires and mortality follow-up have been performed in 1997/98 and 2002/03. Since 2004, the KORA study centre conducts regular follow-up investigations of the original survey population. For details see [16, 17].

Table **1** Sample size of KORA-gen: n = 18,079 participants of the MONICA/KORA surveys S1 to S4 in Augsburg. The age range was 25 to 74 years at recruitment and is 30 to 90 years in 2005

|  | survey S1 (1984/85) | survey S2 (1989/90) | survey S3 (1994/95) | survey S4 (1999/2001) |
|---|---|---|---|---|
| men | 2 023 | 2 482 | 2 405 | 2 090 |
| women | 1 999 | 2 458 | 2 451 | 2 171 |

Table **2** Data and biosamples of the MONICA/KORA surveys which can be used in KORA-gen (some variables are only available for subgroups)

| interview | medical examinations |
|---|---|
| – socio-demographic variables | – systolic and diastolic blood pressure |
| – utilization of the medical health care system | – anthropometrical measurements (body height and weight, waist and hip circumference) |
| – smoking | – bioelectrical Impedance Analysis (BIA) |
| – nutrition | – echocardiography |
| – physical activity | – electrocardiogram |
| – medication use | – pulse wave analysis |
| – family history | |
| – women specific variables | |
| – self-reported health status | |
| – psychosocial variables | |

| laboratory examinations | biosamples |
|---|---|
| – cholesterol (total, HDL, LDL) | – serum |
| – uric acid | – plasma |
| – glucose (partly fasting, partly OGTT) | – DNA |
| – triglycerides (partly fasting) | – immortalized lymphocytes |
| – creatinine | – urine |
| – blood cell counts (WBC, platelets, RBC, MCV, MCH, MCHC platelet volume) | |
| – HbA1 c | |
| – haemostasiological and inflammatory parameters | |

Certain conditions have to be fulfilled when using KORA-gen. The rules set by the responsible ethics committee and the office for privacy/data protection have to be followed. Quality standards have to be met with respect to scientifically sound research questions, study designs have to be based on realistic sample size calculation, quality of lab tests and genotyping has to be fulfilled according to internationally accepted standards.

Furthermore the rights and scientific interests of the KORA researchers have to be taken into account in a fair manner, since they performed the field work and invested a lot of costs and energy to accumulate the data and biosamples.

The KORA-gen platform will start providing data in spring 2005.

**Reference**

[1] Wichmann HE. Genetic epidemiology – from biobanking to genetic statistics. Methods of Information in Medicine 2005 (in press)

[2] DECODE Genetics Biobank Iceland; www.decode.com

[3] Estonian genome project; www.geenivaramu.ee

[4] UK Biobank; www.ukbiobank.ac.uk

[5] Japan Biobank – The biobank Japan project.; www.atip.org/public/atip.reports.03/atip03.042.pdf

[6] Collins FS. The case for a US prospective cohort study of genes and environment. Nature 2004; 429: 475 – 477

[7] P3G – public population project in genomics; www.p3gconsortium.org

[8] Knoppers BM, Fecteau C. Human genomic databases: a global public good? Eur J Health Law 2003; 10: 27 – 41

[9] Eaton W. The logic for a conception-to-death cohort study. Ann Epidemiol 2002; 12: 445 – 451

[10] Kaiser J. Population databases boom, from Iceland to the U.S. Science 2004; 296: 1158 – 1161

[11] Willett WC. Balancing life-style and genomics research for disease prevention. Science 2002; 296: 695 – 698

[12] GEM (Genetic Epidemiological Methods) Munich – Ethics and confidentiality; http://www2.gsf.de:6666/gem/ethik.htm

[13] Wichmann HE, Jäger L, Taupitz J et al. Ethik und genetische Epidemiologie. Deutsches Ärzteblatt 2002; 99: 2 – 4

[14] German National Ethics Council (Nationaler Ethikrat): Biobanks for research – opinion (Biobanken für die Forschung). Berlin: Saladruck, 2004, www.ethikrat.org

[15] POPGEN – Population Genetics in Schleswig-Holstein; www.popgen.de

[16] KORA (Kooperative Gesundheitsforschung in der Region Augsburg); www.gsf.de/KORA

[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] NGFN – National Genomic Research Network (Nationales Genomforschungsnetz); www.rzpd.de/ngfn

[19] Müller S, Martin S, Koenig W et al. Impaired glucose tolerance is associated with increased serum concentrations of interleukin 6 and coregulated acute-phase proteins but not TNF- or its receptors. Diabetologia 2002; 45: 805 – 812

[20] Werner M, Sych M, Herbon N et al. Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF Mass Spectrometry. Hum Mutat 2002; 20: 57 – 64

[21] Herbon N, Werner M, Braig C et al. High resolution SNP scan of chromosome 6p21 in pooled samples from patients with complex diseases. Genomics 2003; 81: 510 – 518

[22] Klopp N, Heon E, Billingsley G et al. Further genetic heterogeneity for autosomal dominant human sutural cataracts. Ophthalmic Research 2003; 35: 71 – 77

[23] Illig T, Bongardt F, Schöpfer A et al. The endotoxin receptor TLR 4 polymorphism is not associated with diabetes or components of the metabolic syndrome. Diabetes 2003; 52: 2861 – 2864

[24] Weedon M, Schwarz P, Horikawa Y et al. Meta-analysis confirms a role for Calpain-10 variation in type 2 diabetes susceptibility. Am J Hum Gen 2003; 73: 1208 – 1212

25 Kääb S, Pfeufer A, Illig T et al. Frequent allelic variants of the myocardial Ikr-alpha-subunit gene KCNH2 modulate the duration of the QTc interval in the general population. Circulation 2003; Suppl IV: 726

26 Geller F, Reichwald K, Dempfle A et al. Melanocortin-4 receptor gene variant I103 is negatively associated with obesity. Am J Hum Gen 2004; 74: 572 – 581

27 Weidinger S, Klopp N, Wagenpfeil S et al. Association of a STAT 6 Haplotype with elevated serum IgE levels in a Population-Based Cohort of Caucasian Adults. J Med Genet 2004; 41: 658 – 662

28 Illig T, Bongardt F, Schöpfer A et al. Significant association of the interleukin-6 gene polymorphisms C-174G and A-598G with type 2 diabetes. J Clin Endocr Metab 2004; 89: 5053 – 5058

29 Illig T, Vollmert C, Gohlke H et al. Hochdurchsatzgenotypisierung mit MALDI TOF MS. BIOspektrum 2004; 5: 670 – 674

30 Mostafazadeh A, Herder C, Haastert B et al. Association of Humoral Immunity to Human Hsp60 With the IL-6 Gene Polymorphism C-174G in Patients With Type 2 Diabetes and Controls. Hormone Metab Res 2005; 37: 257 – 263

31 Lamina C, Steffens M, Wienker T et al. Genetic diversity in German and European populations: Looking for substructures and genetic patterns. Gesundheitswesen 2005; 67 S1: S57 – S61

32 Mueller JC1, Lõhmussaar E, Mägi R et al. Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 2005; 76: 387 – 398

33 KORA-gen; www.gsf.de/KORA-gen

34 Löwel H, Döring A, Schneider A et al. The MONICA Augsburg surveys – basis for prospective cohort studies. Gesundheitswesen 2005; 67 S1: S13 – S18

35 Koenig W, Meisinger C, Baumert J et al. Systemic Low-Grade Inflammation and Risk of Coronary Heart Disease: Results from the MONICA/KORA-Augsburg Cohort Studies. Gesundheitswesen 2005; 67 S1: S62 – S67

36 Schunkert H, Döring A, Kuch B et al. Cardiovascular phenotypes and functional parameters in the general population – results of the MONICA/KORA studies. Kardiovaskuläre Phänotypen und Funktionsparameter in der Normalbevölkerung – Ergebnisse der MONICA/KORA-Studien. Gesundheitswesen 2005; 67: S74 – S78

37 Peters A, für die MONICA/KORA-Studiengruppe. Partikel in der Außenluft erhöhen das Risiko für Herzkreislauferkrankungen. Gesundheitswesen 2005; 67 S1: S79 – S85

38 Thorand B, Schneider A, Baumert J et al. Fall-Kohorten-Studien: Ein effektives Design zur Untersuchung von Biomarkern als Risikofaktoren für chronische Krankheiten – Darstellung am Beispiel der MONICA/KORA Augsburg Fall-Kohorten Studie 1984 – 2002. Gesundheitswesen 2005; 67 S1: S98 – S102

39 Herder C, Illig T, Rathmann W et al. Inflammation and Type 2 Diabetes: Results from KORA Augsburg. Gesundheitswesen 2005; 67 S1: S115 – S121

40 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

41 Heid IM, Lamina C, Bongardt F et al. Wie gut können Haplotypen in den populationsbasierten KORA Studien rekonstruiert werden? Gesundheitswesen 2005; 67 S1: S132 – S136