Institute of
Epidemiology

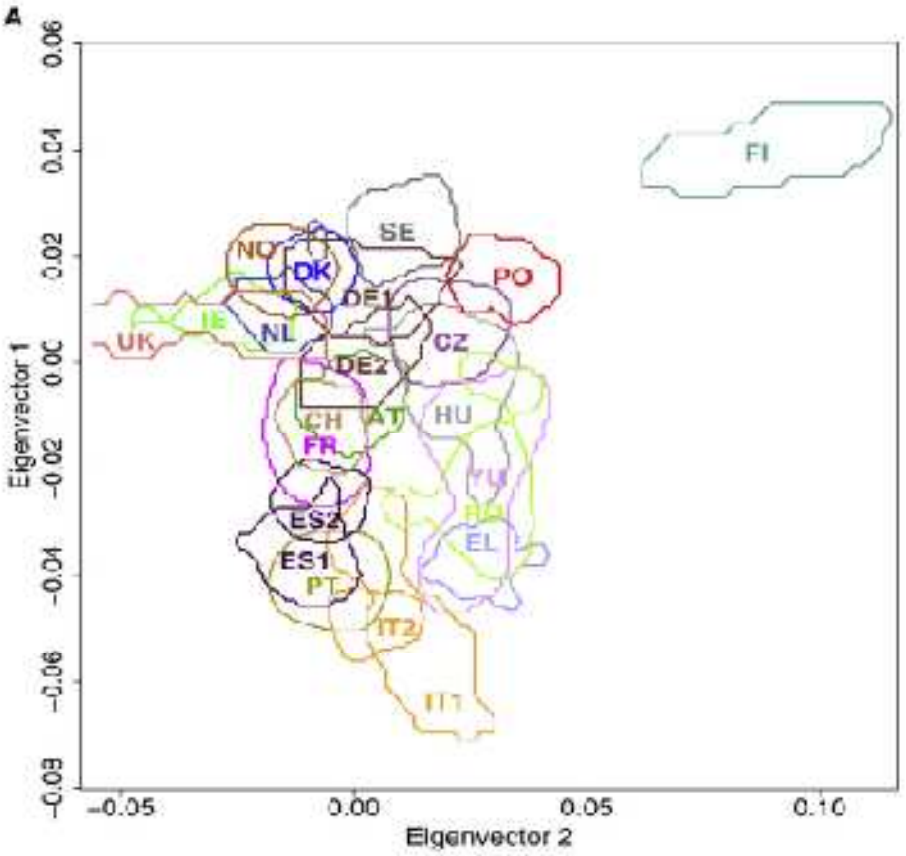
Phenotypic and genotypic information associated to a biobank

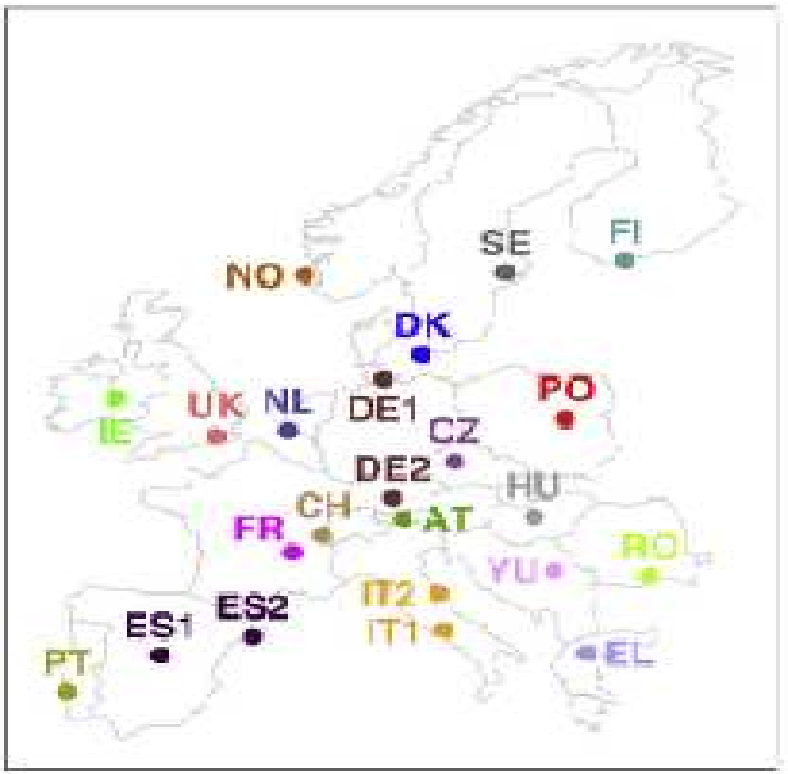
Erich Wichmann

Helmholtz Center Munich, University of Munich

Presentation at ESF-UB Conference in Biomedicine - Biobanks: Introduction and Next Steps
Session 2: Data Archiving and Management: the IT Infrastructure
Sant Feliu de Guixols (Costa Brava), Spain, 1-6 November 2008

Correlation between Genetic and Geographic Structure in Europe





N=2514, 23 subpopulations, Affy 500k, PCA, small genetic variation, strong correlation of geographic and genetic distance

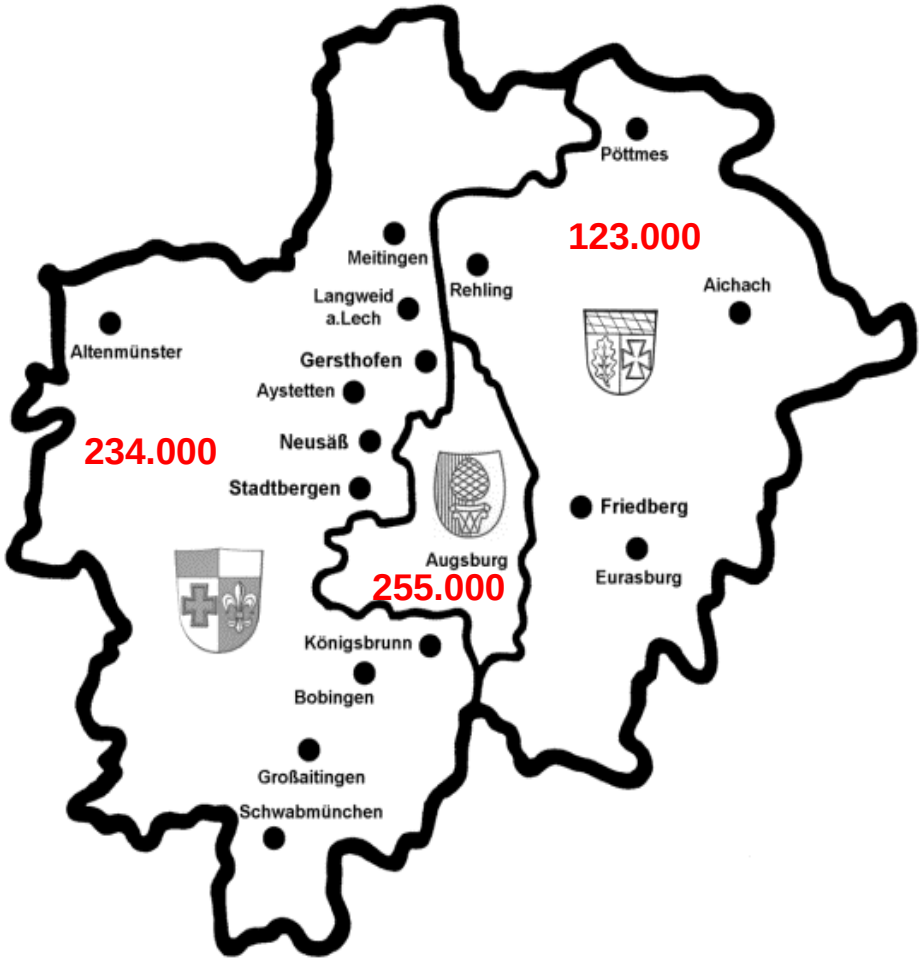
Lao et al, Current Biology, 2008

Content

- **Research examples**
  - Epidemiology
  - Genetic epidemiology (GWAS)
  - Gene-environment Interaction
- **From a study to a biobank**
  - Rules of access
  - Phenotypes
  - Genotypes
  - Biorepository
- **Future: New developments**
  - Genomics, transcriptomics, proteomics, metabolomics, sequencing
- **Summary**

KORA - Study

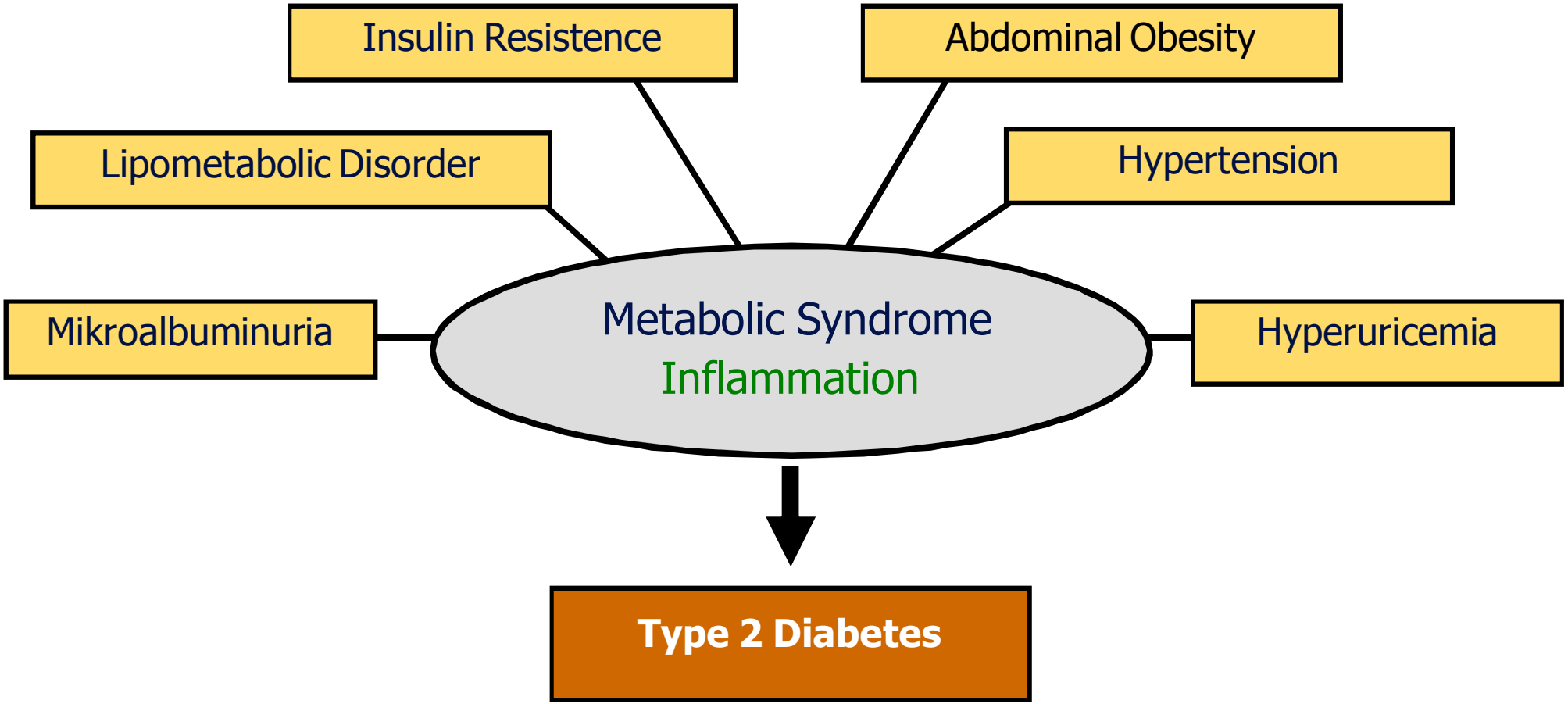
- KORA = Cooperative Health Research in the Region of Augsburg
- Population based cohort study (18,000 individuals)
- Age range 25-74 years at recruitment
- Follow up investigations for more than 20 years



KORA Phenotyping

- Interview: Socio-demographic variables, smoking, nutrition, physical activity, medication use, self-reported health status, …
- Medical examinations: Blood pressure, anthropometry, ECG, echocardiography, OGTT, lung function, endothelial dysfunction, thyroid sonography, skin …
- Laboratory examinations: Total cholesterol, HDL, LDL, uric acid, blood cell counts, HbA1c, glucose, inflammatory and many other parameters

Epidemiology: Type 2 diabetes
role of inflammation



Epidemiology: Type 2 diabetes
investigated parameters

Cytokines/ Chemokines

- **IL-6**; solv. IL-6R
- **IL-18**, TNF-α
- solv. TNF-R1 & TNF-R2
- MIF, MCP-1
- RANTES, Eotaxin
- IL-8, IP-10

Blood-/Endothel-Activation

- solvable ICAM-1
- solvable E-Selectin
- Von-Willebrand-Factor

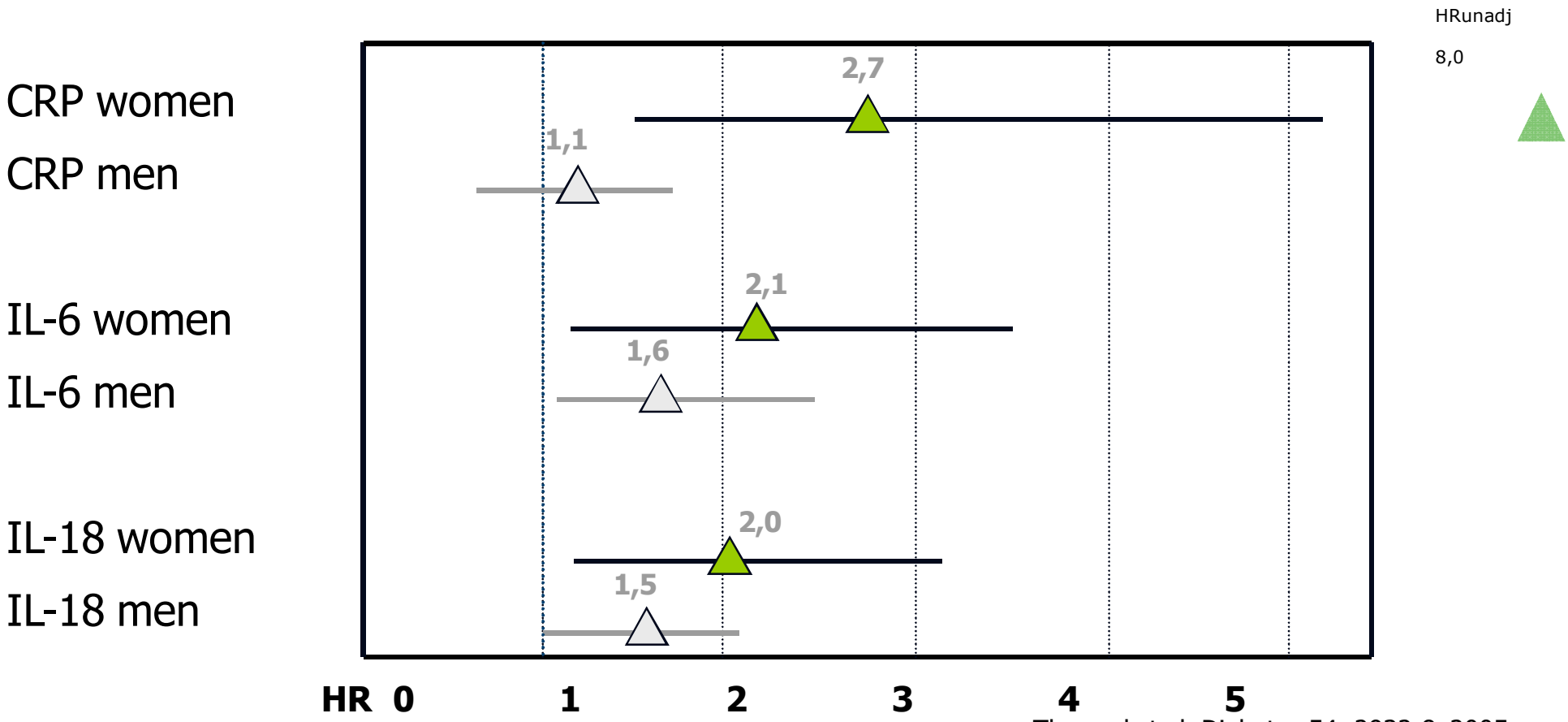
Acute-Phase Proteins

- **CRP** (C-reactive Protein)
- Serum Amyloid A
- Fibrinogen

KORA S123 case cohort study 1984-2002

KORA S4 case control study T2D, IGT, NGT

Epidemiology: Type 2 diabetes
results: role of inflammation



HRadj highest vs lowest tercentile

Thorand et al. Diabetes 54: 2932-8, 2005;
Thorand et al. Diabetes Care 30: 854-60, 2007

Genetic Epidemiology: Type 2 diabetes
Genome-wide analysis

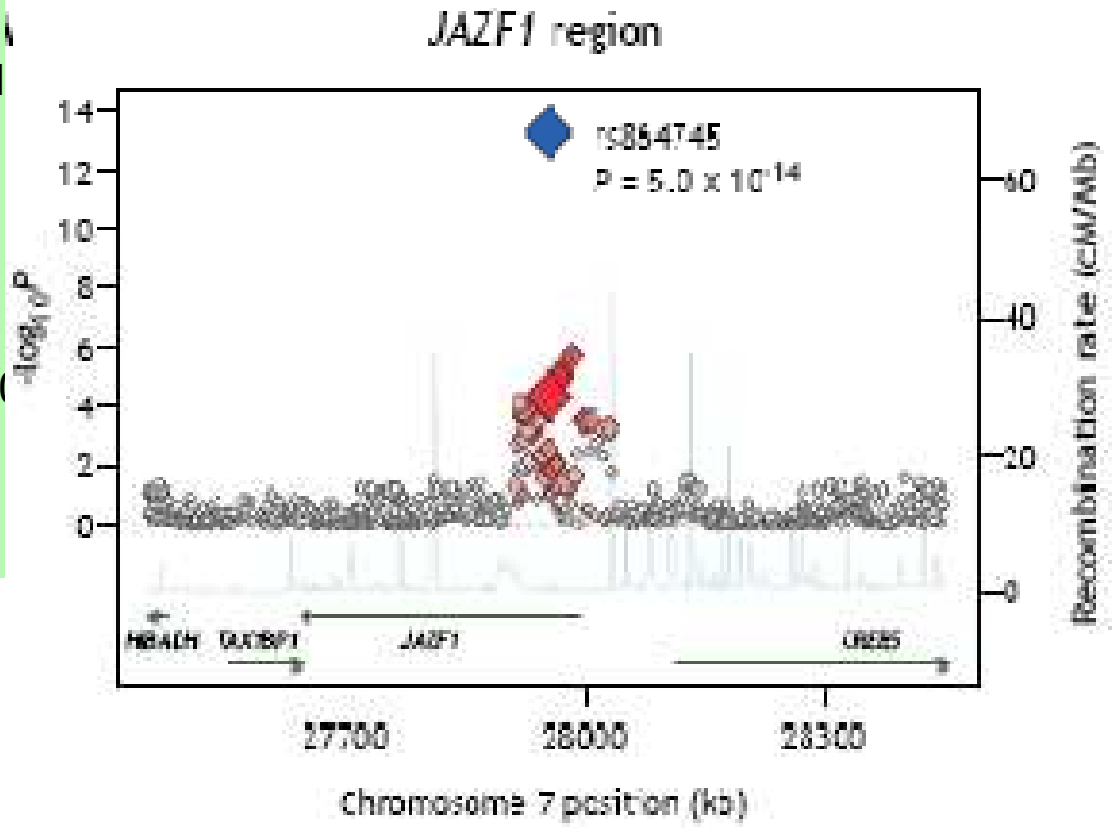
12 study populations (WTCCC, DGI, FUSION, DeCode, **KORA**, Danish, Hunt, EPIC, NHS, ADDITION/Ely, GEM, METSIM

Stage 1: GWA meta-analysis (4,500 T2D cases, 5,500 controls)

…

Stage 3: replication (14,000 cases, 43,000 controls)

> 6 new susceptibility genes for T2D



Zeggini et al. Nat Genet, 2008

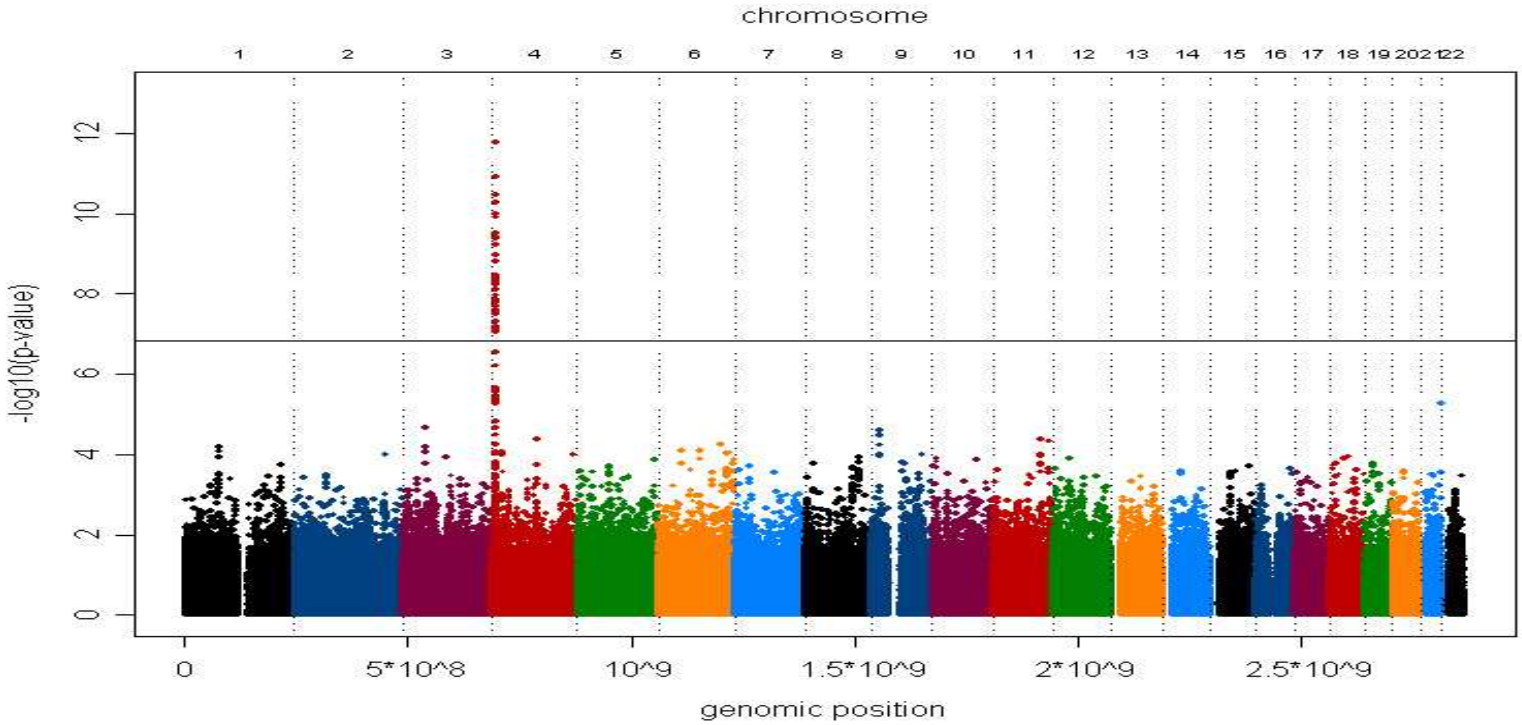
Genetic Epidemiology: GWAS of Uric Acid (UA)
Genome-wide screen (n=1644)

Institute of Epidemiology

Distribution of p-values on the genome in an additive model
SNPs above the line are genome-wide significantly correlated with uric acid

The x-axis represents the genomic position (in Gb) of 335,152 SNPs, and the y-axis shows –log10 (p).

Bonferroni corrected significance level $1.5*10^{-7}$, (nominal significance 0,05)

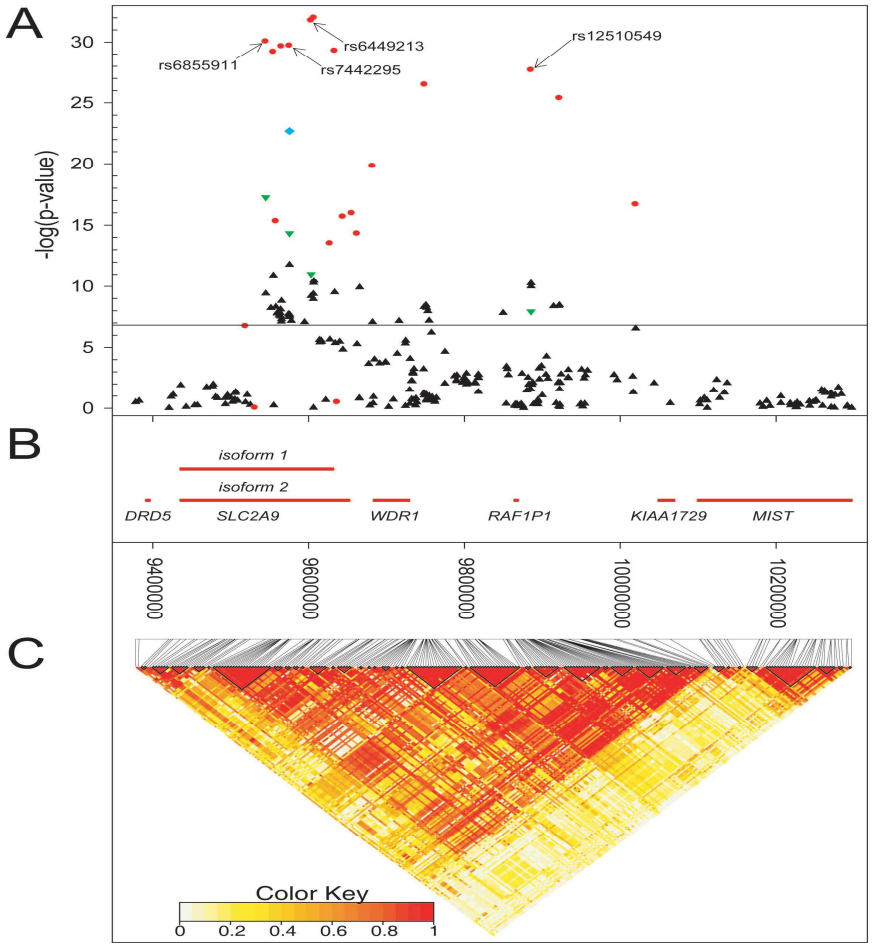


Genetic Epidemiology: GWAS of Uric Acid
replication (n=11,591) LD-Plots Chr 4

A: P-value distribution on chromosome 4:
The y-axis shows –log10 (p) - values of KORA F3 500K (black), KORA S4 (red), SAPHIR (green), SHIP (blue)

B: Gene regions are indicated by bars.

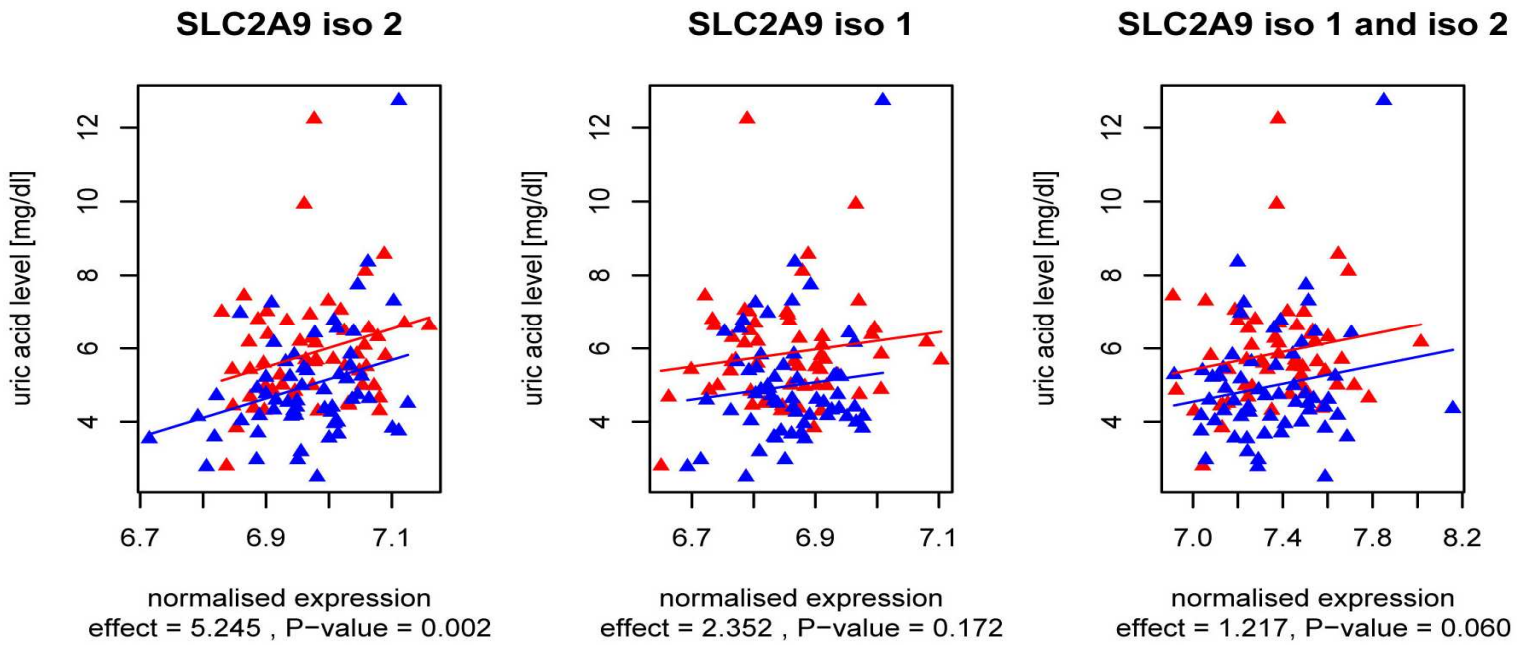
C: Pairwise linkage disequilibrium diagram of the region on chromosome 4



Döring et al. Nat. Genet. 2008

Genetic Epidemiology: GWAS of Uric Acid
Transcription analysis (n=117)

Institute of Epidemiology



The *SLC2A9* gene is represented with three probes detecting the alternative first exons of isoforms 1 (iso 1) and 2 (iso 2) as well as both isoforms. The regression line is shown for females (blue) and males (red).

KORA GWAS publications

2006 (2)

Herbert et al. **Science** (2006)

Arking et al. **Nat Genet** (2006)

2007 (4)

Winkelmann et al. **Nat Genet** (2007)

Lyon et al. **PLoS Genet** (2007)

Moffatt et al. **Nature** (2007)

Samani et al. **N Engl J Med** (2007)

2008 (20)

Döring et al. **Nat Genet** (2008)

Zeggini et al. **Nat Genet** (2008)

Lettre et al. **Nat Genet** (2008)

Loos et al. **Nat Genet** (2008)

Aulchenko YS et al. **Nat Genet** (2008)

Schormair B et al. **Nat Genet** (2008)

Lasky-Su et al. **Am J Hum Genet** (2008)

Luca et al. **Am J Hum Genet** (2008)

Sue et al. **Am J Hum Genet** in press

Gibson et al. **Proc Natl Acad Sci** (2008)

Schunkert et al. **Circulation** (2008)

Lieb et al. **Circulation** (2008)

Sinner et al. **Eur Heart J** (2008)

Hinterseer et al. **Eur Heart J** (2008)

Weidinger et al. **Plos Genetics** (2008)

Linsel-Nitschke et al. **Plos One** (2008)

Gieger et al. **Plos Genetics** in press

Lao O et al. **Curr Biol** (2008)

Herder C et al. **Horm Metab Res** (2008)

Heid I et al. **Circ Cardiovasc Genet** (2008)

60 further KORA GWAS ongoing

KORA phenotypes:

Height, weight, BMI, body fat, lean body mass, type A, type D, nicotin, alcohol, lung function, kidney function, metabolic syndrome, type 2 diabetes, micro/macrovascular complications of diabetes, Hba1c, insulin, glucose, atopic dermatitis, IgE, myopia, myocardial infarction, left ventricular hypertrophy, endothelial dysfunction, blood pressure, pulse pressure, ABI, ECG (QT, PQ, QRS), atrial fibrillation, Ca, K, Mg, cholesterol, HDL, LDL, triglycerides, CRP, phytosterols, MCP-1, fibrinogen, MPV, leptin, adiponectin, uric acid, liver enzymes, phosphate, Fe, BNP, aldosteron, renin, APOE

KORA as controls: 1600 Affy 500k, 2000 Affy 1000k, 900 Ilumina 550k

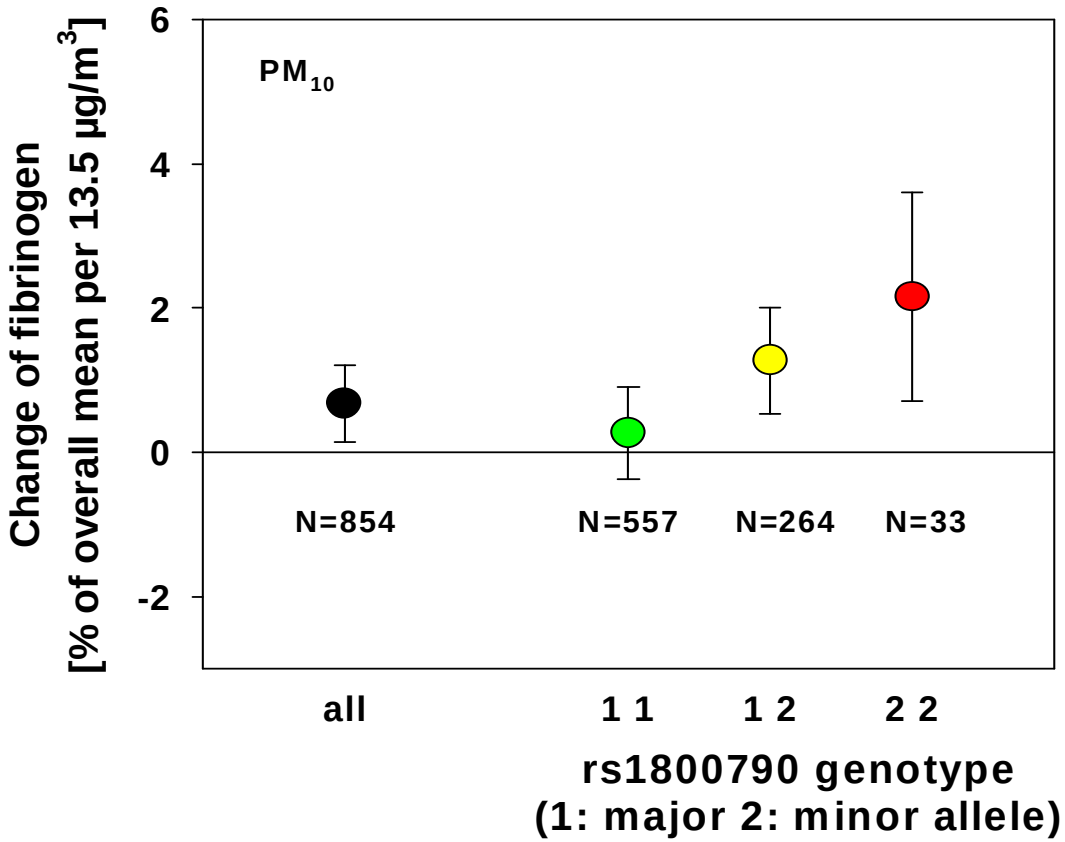
Collaborations/consortia:

MORGAM, ENGAGE, CARDIOGENICS, GIANT, IQWANA, MOLPAGE, DIAGRAM, NGFN (German National Genome Network)

Gene-Environment Interaction
Effect of ambient air pollution on fibrinogen in susceptible groups

- **AIRGENE:** European Multi-center Study in 1003 myocardial infarction survivors (Coordination: Annette Peters, HMGU)
- 5813 measurements of blood markers of inflammation
- Particulate Matter and genes involved in regulation of inflammation

> Genetic Susceptibility modifies the influence of Particulate Matter (PM10)



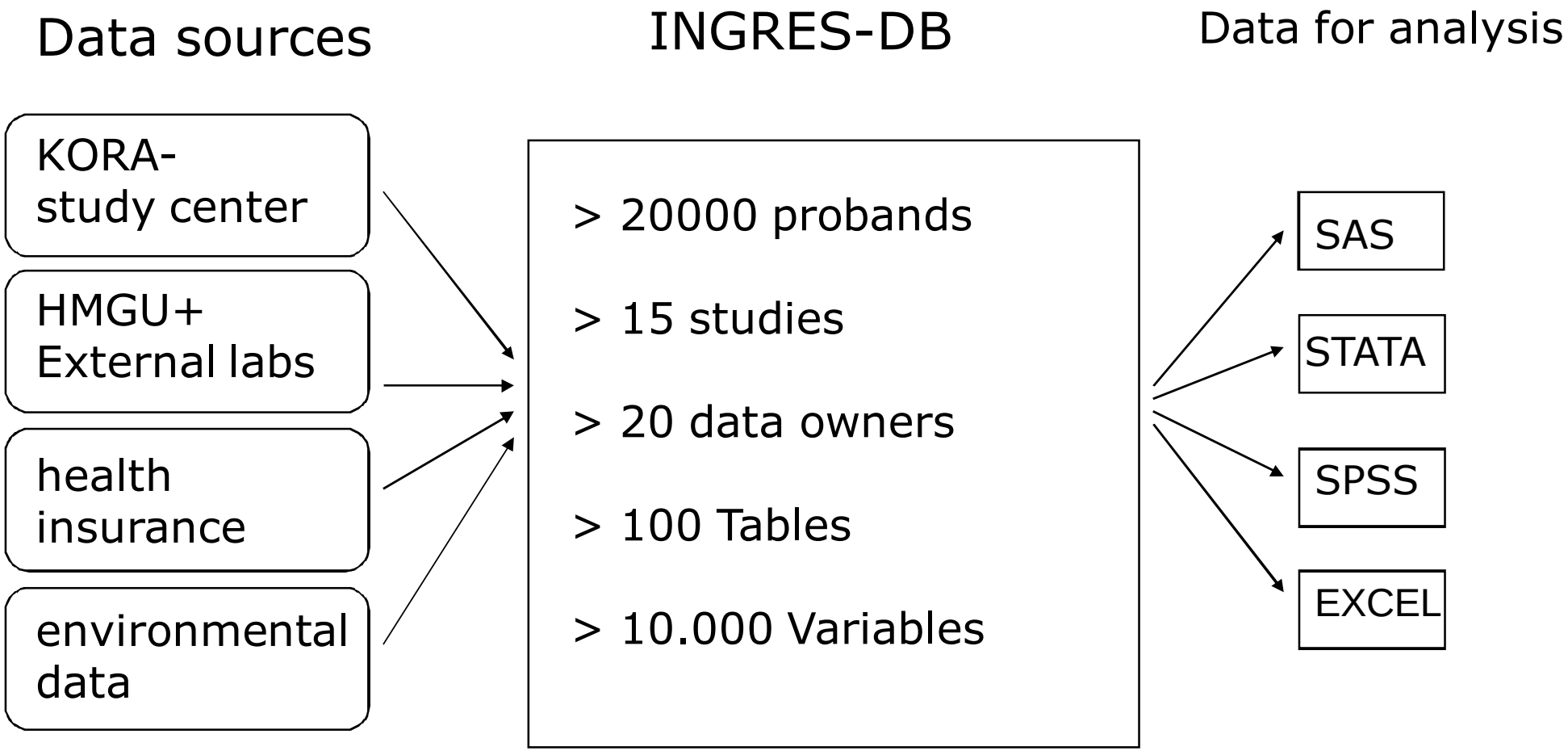
Peters et al. 2008 in press

KORA -> KORA-gen
from a study to a biobank

Rules of access for external partners

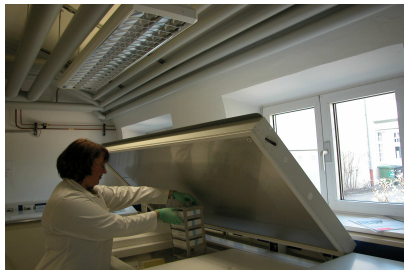
- Access to DNA, blood, serum, plasma samples, phenotypes and genotypes
- KORA-gen internet portal (http://epi.gsf.de/kora-gen/)
- Project agreement contracts: delineation of the planned project, scientific background, aim of publications, variables and data
- Co-operation agreement with data owners who become co-authors in resultant publications
- To date > 400 transfer agreements for KORA data and biosamples have been contracted

Databases for phenotypes



Biorepository

- Samples of 50 000 individuals from birth to 84 years (KORA, GINI, LISA, case samples)
- Stored materials: blood, plasma, serum, urine, DNA, RNA, EBV-immortalized lymphocytes
- Storage conditions 30 x –80°C freezers and 12 liquid nitrogen tanks in an air conditioned room
- More than 50 000 samples retrieved in 2007

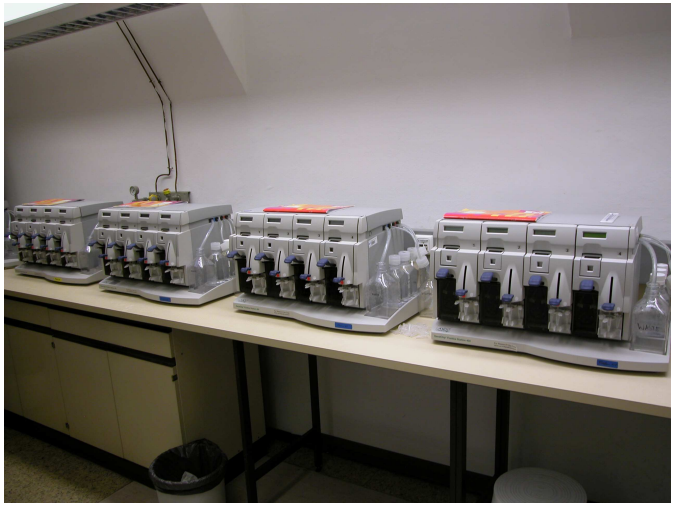
Data generated by genotyping

- **Sequenom**
  - MALDI-TOF system
  - ca. 1-50 SNPs per proband
  - ca. 500K to 1M data sets per analysis
- **Affymetrix**
  - 500K, 1000K
  - 500K to 1M SNPs per proband
  - ca. 50M to 10G data sets per analysis
- **Illumina**
  - Golden Gate, 50K, 550K, 1000K
  - 30K to 1M SNPS per proband
  - ca. 50M to 10G data sets per analysis
- **Illumina Solexa**
  - Whole genome sequencing
  - raw data: ca. 1TB per proband
  - N * 3,5G data sets per analysis

Sequenom


BRUKER
SEQUENOM
autoflex
autoflex
BRUKER DALTONICS
SEQUENOM

Affymetrix



Illumina

Illumina Solexa

Information management

- Modern high-throughput technologies generate an enormous amount of data that must be stored and processed.
- 3.500 probands, genotyped with Affymetrix GeneChip technology, yield up to 3.5 Billion genotypes.
- Thus, an efficient, robust and scaleable information management system is required.

Challenges in establishing and handling Bio-(Data-)Banks 1

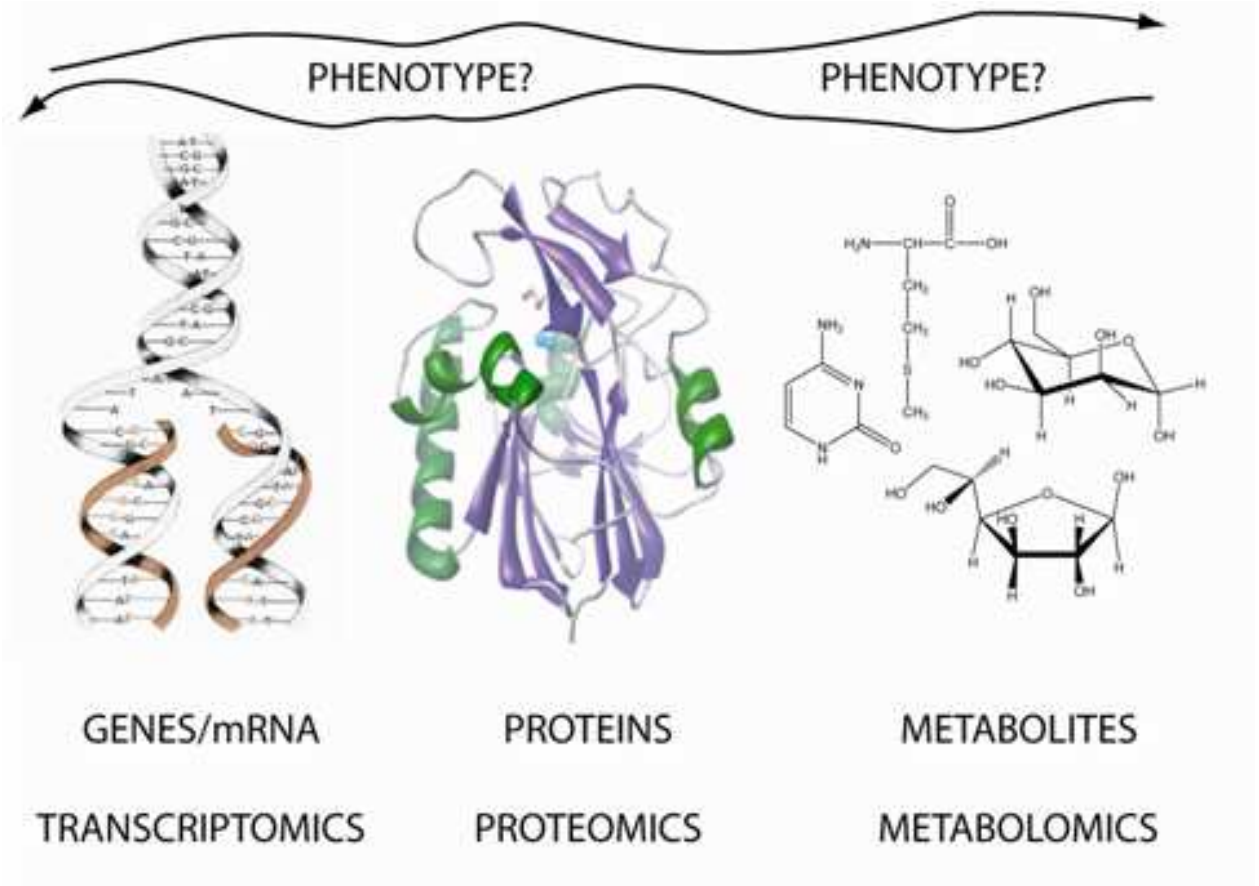
- High storage capabilities for raw data (incl. backups) and their availability for repeated analyses (KORA-gen: 40TB storage system)
- Access to powerful computers for data processing and data analysis (KORA-gen: Linux based systems/cluster environment)
- Integration of various technologies and data types (SNPs, CNVs, haplotypes, …) in one combined data base

Challenges in establishing and handling Bio-(Data-)Banks 2

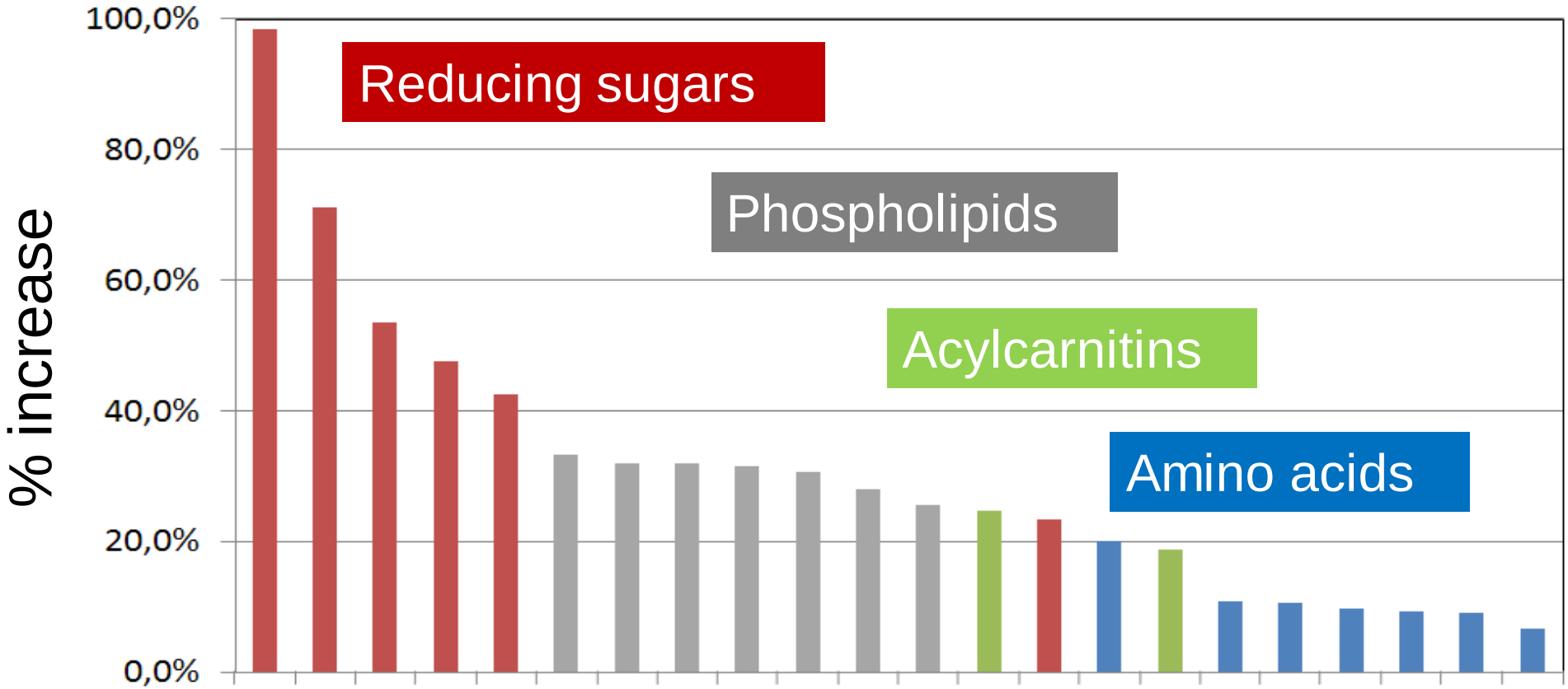
Availability of the data for cross-study projects:

- Data harmonization in networks of biobanks (BBMRI)
- Decentralized data handling in consortia (ENGAGE, GIANT, CHARGE) for combined analyses and meta-analyses
- Definition of data exchange and exchange formats
- Systems: BIMS / Karolinska Institutet; AIMS/EMBL-EBI

Future: New sources of intermediate phenotypes: Transcriptomics, Proteomics, Metabolomics



Future: Metabolomics
Type 2 diabetes pilot study



40 KORA diabetes patients compared to 296 controls

expected: reducing sugars, amino acids;

new: Phospholipids, Acylcarnitines; Mechanism?

Replication and validation necessary!

Future: KORA F3 metabolomics study

Targeted Quantitative Metabolomics

API 4000 LC-MS-MS, Hamilton robotics

Quantitative and reproducible, High throughput (hundreds of samples)

Large set of relevant metabolite markers: 363 metabolites were detected in 300 males:

- Sugars (9x)
- BiogenicAmines (7x)
- Prostaglandines (7x)
- Acylcarnitines & Amino Acids (47x)
- Sphingolipids (85x)
- Glycerophospholipids (208x)

Future: Biosample repository
What do we need?

All HMGU populations (including KORA) :

- Sample size N = ca. 60.000, which translates in ca. 1.600.000 serum aliquots, 30.000 DNA samples, 4.000 cell lines

Helmholtz Cohort:

- Sample size N = ca. 200.000 adult subjects, 4 million bio samples

Biosample repository:

- -80°C or liquid nitrogen automated storage and retrieval system for several million samples
- Automated preparation using robotic systems
- Professional „Laboratory Information System“ (LIMS)
- Back up storage (catastrophic loss)
- Retrieval of more than 200 000 samples per year

Summary
with respect to data archiving and management

- broad phenotyping and detailed data on risk factors: important
- Genotyping for large numbers of subjects: needed
- Expansion to other –omics: produces even more data

-> complex IT solutions, exchange of data, interaction in networks is crucial

Acknowledgements

KORA phenotyping: Angela Döring, Christa Meisinger, H-Erich Wichmann

KORA genotyping: Peter Lichtner, Gertrud Eckstein, Norman Klopp, Thomas Illlig, Thomas Meitinger

KORA statistics: Christian Gieger, Iris M. Heid

KORA IT: Guido Fischer, Christian Gieger