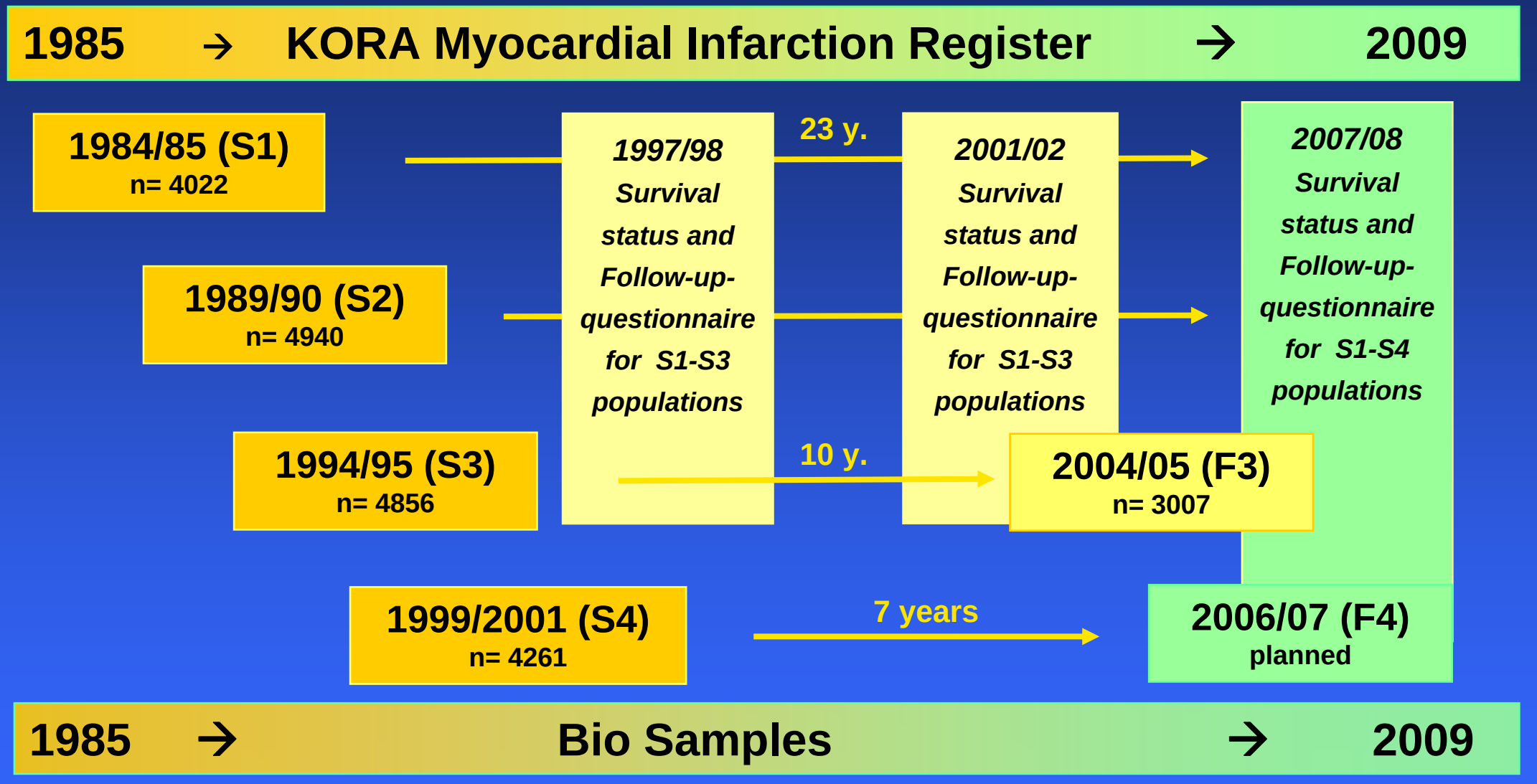
Genome-wide association studies – first experiences with study planning, data management and statistical analysis

Wichmann HE, Gieger C, Heid I, Illig T, Meitinger T (GSF) and 10 collaborators

Presentation at GMDS Conference Leipzig
10. – 14. September 2006

gsf - Institute of Epidemiology

MONICA/KORA populations



KORA-gen: Use of KORA in genetic studies

2003-2006

- More than more than 50 genetic case control studies
- 2 studies in populations genomics
- 3 genetic meta-analyses and pooled analyses
- in total, 80-90 data and sample transfer agreements

Genome-wide association studies

- 100k studies (EKG-QT, BMI)
- KORA 500k project

Genome-wide association studies

Successful proof of principle in the literature:

- **Lymphotoxin-alpha gene and Myocardial Infarction (93,000 SNPs, 1,100 cases, 1,000 controls) Ozaki et al. Nat. Genet. (2003)**
- **Complement factor H gene and Macular Degeneration (116,000 SNPs, 96 cases, 50 controls) Klein et al. Science (2005)**
- **SEMA5A, PARK11 genes and Parkinson‘s Disease (198,000 SNPs, in 443 discordant sibs, 2,000 SNPs in 332 case control pairs) Maraganore et al. Am. J. Hum. Genet. (2005)**

Multiple common genetic variants modulate cardiac repolarization (QT-interval)

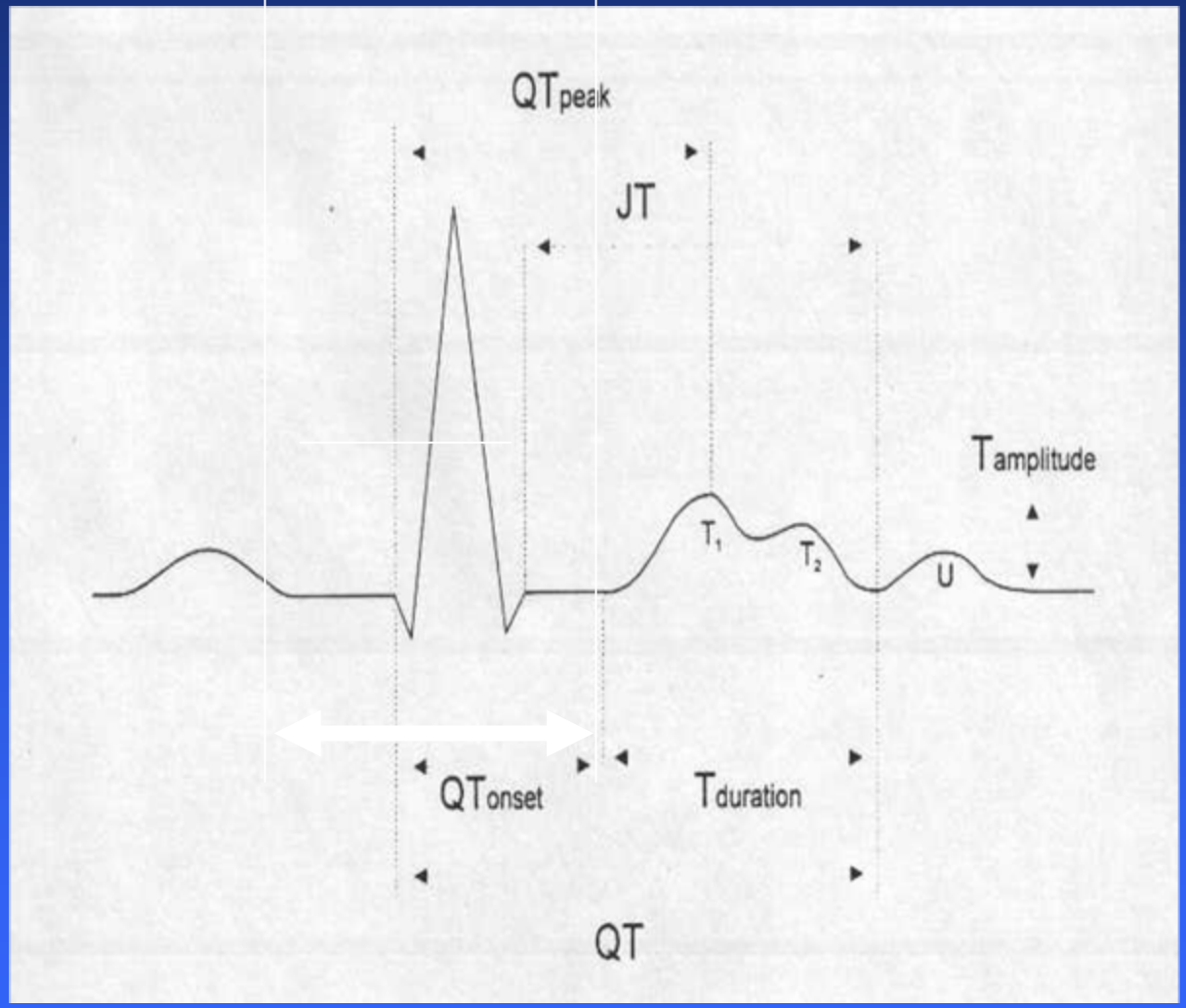
D.E. Arking*, A. Pfeufer*, W. Post, W.H.L. Kao, M. Ikeda,
K. West, C. Kashuk, M. Akyol, S. Perz, S. Jalilzadeh,
T. Illig, H.E. Wichmann, E. Marbán, S. Kaab,
P.M. Spooner, T. Meitinger, A. Chakravarti
**authors contributed equally*

Johns Hopkins University School of Medicine, Baltimore
Department of Epidemiology, Johns Hopkins University
GSF, TUM, LMU

Nature Genetics May 2006

gsf - Institute of Epidemiology

The QT interval



A Common Genetic Determinant of Adult and Childhood Obesity

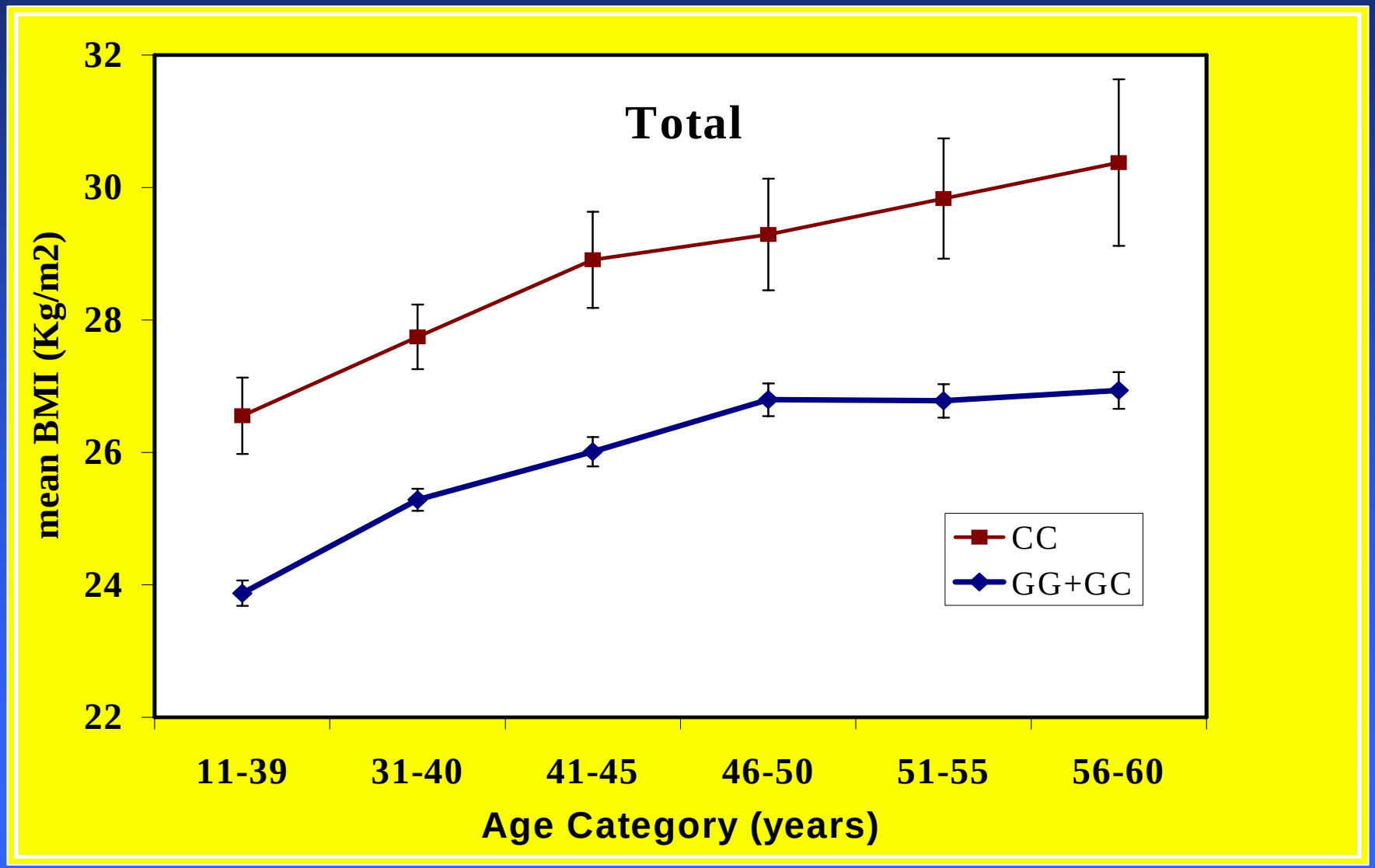
A. Herbert, N.P. Gerry, M. McQueen, I.M. Heid, A. Pfeufer, T. Illig, H.E. Wichmann, T. Meitinger, D. Hunter, F.B. Hu, G. Colditz, J. Hebebrand, X. Zhu, R. Cooper, K. Ardlie, H. Lyon, J. Hirschhorn, N.M. Laird, M.E. Lenburg , C. Lange, M.F. Christman

Boston University Medical School
Harvard School of Public Health
Harvard Medical School
Broad Institute of MIT and Harvard
SeraCare Life Sciences Inc.
Loyola University Medical Center
Essen University, GSF, LMU, TUM

Science May 2006

gsf - Institute of Epidemiology

A Big Effect for BMI



KORA 500k - Design

1800 S3/F3 subjects (survey S3 and 10y follow-up F3)

Phenotypes:

Type 2 Diabetes, metabolic syndrome, hypertension, body mass index (BMI), QT interval, left ventricular hypertrophy (LVH), inflammatory parameters, lipids.

KORA 500k
Qualitative or quantitative endpoint

Phenotype	Partner	Quantitative (linear regression)	Qualitative (logistic regression)
Atopy	**Weidinger/Ring (TU Munich)**		**x**
BMI	**Hebebrand (Essen)**	**x**	
CRP	**Koenig (Ulm)**	**x**	
HDL	**Kronenberg (Innsbruck)**	**x**	
Hypertension	**Laan (Tartu, Estonia)**		**x**
LVMass	**Schunkert (Lübeck)**	**x**	
Met. syndrome	**Hengstenberg (Regensburg)**		**x**
QT interval	**Pfeufer/Kääb (LMU Munich)**	**x**	
Type 2 diabetes	**Scherbaum (Düsseldorf)**		**x**
Phytosterols	**Thiery (Leipzig)**	**x**	

KORA 500k Data Management

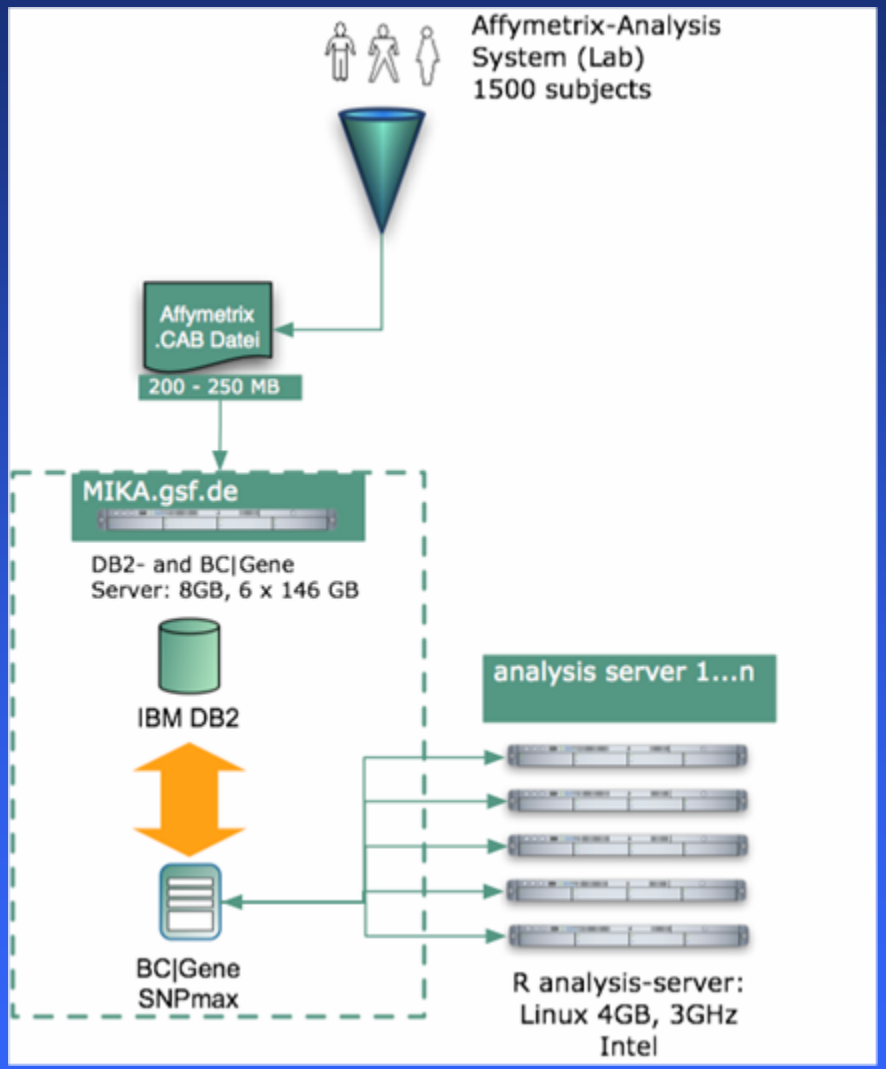
BC SNPmax - Biocomputing Platforms

- Scalable with respect to input and output of data based on IBM DB2
- Combining of genotype and phenotype data
- Management of access rights

KORA 500k Data Management

- Basic project management and data retrieval system
- Defined interface to statistical analysis software, e.g. R
- Possibility to integrate other analysis tools

KORA 500k Data Management
IT-Infrastructure



- Hardware components
- Database
- Data management system
- Data analysis system

KORA 500k Status July 2006
Genotyping and Data Management

- 1138 subjects are completed (NSP-Chip + STY-Chip)
- Chips are called with DM algorithm
- quality controlled:
  - call rates of both chips over 93%
  - called gender of each chip equal to KORA-database
  - 50 overlapping SNP with maximum of 1 error

KORA 500k Status July 2006
Comparison NSP + STY Chip

50 SNPs are located on both chips. The genotype from one patient should only differ in 1 of 50 SNP.

AA	BB	BB	BB	nc	BB	AA	BB	AA	BB	nc
AA	BB	AA	BB	AA	BB	AA	BB	AB	BB	AB

KORA 500k Status July 2006
Database (BC|SNPmax)

To upload one chip:	6-8 min
Maximum uploads per day:	120 Chips

Content:

- 1138 subjects
- 2276 chips
- 500568 SNPs per subject
- In total: 569.646.384 rows (genotypes)

KORA 500k Status July 2006
Data description – Observed genotypes*

A/A (monomorph)	: 27,735	5.6 %
A/A – A/B	: 60,273	12.3 %
A/A – B/B	: 167	0.03 %
A/A – A/B – B/B	: 401,857	82.0 %

(*without X chromosome)

KORA 500k Status July 2006
Data description – minor allele frequencies*

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0004	0.0686	0.1850	0.2047	0.3311	0.5000

9.75% with allele frequency below 1 %

20.5 % with allele frequency below 5 %

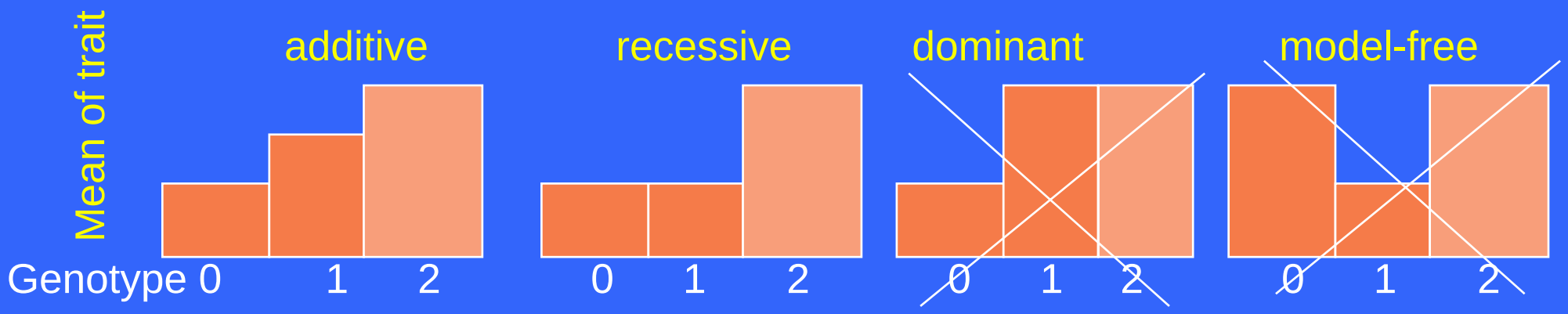
32.5 % with allele frequency below 10 %

(*without X chromosome and monomorph SNPs)

KORA 500k Statistical analysis – inheritance model

- **Recessive**
- **Additive inheritance**
- **Dominant**
- **Model-free (2 df)**

Means of trait by genotype for different inheritance models:

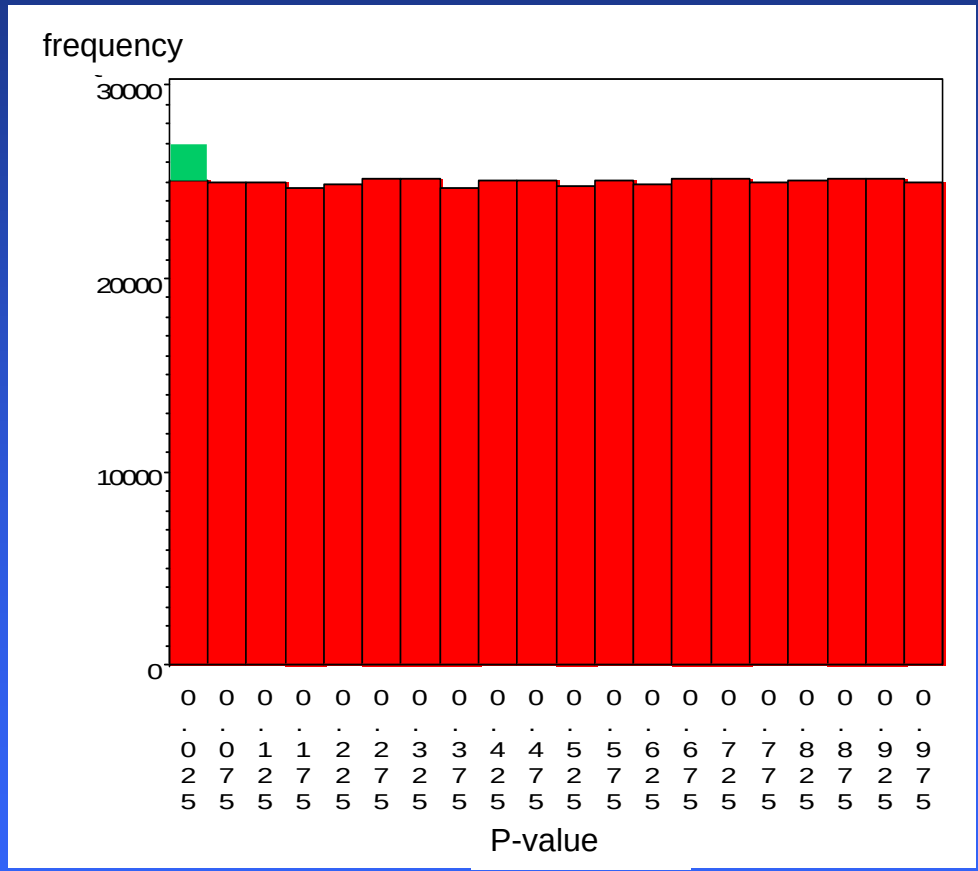


KORA 500k Statistical analysis – The false positives

500 000 SNPs tested

False positives:
25 000 SNPs with p-value < 0.05
2500 SNPs with p-value < 0.005
500 SNPs with p-value < 0.001

Bonferroni-adjusted level:
$0.05/500000 = 10^{-7}$



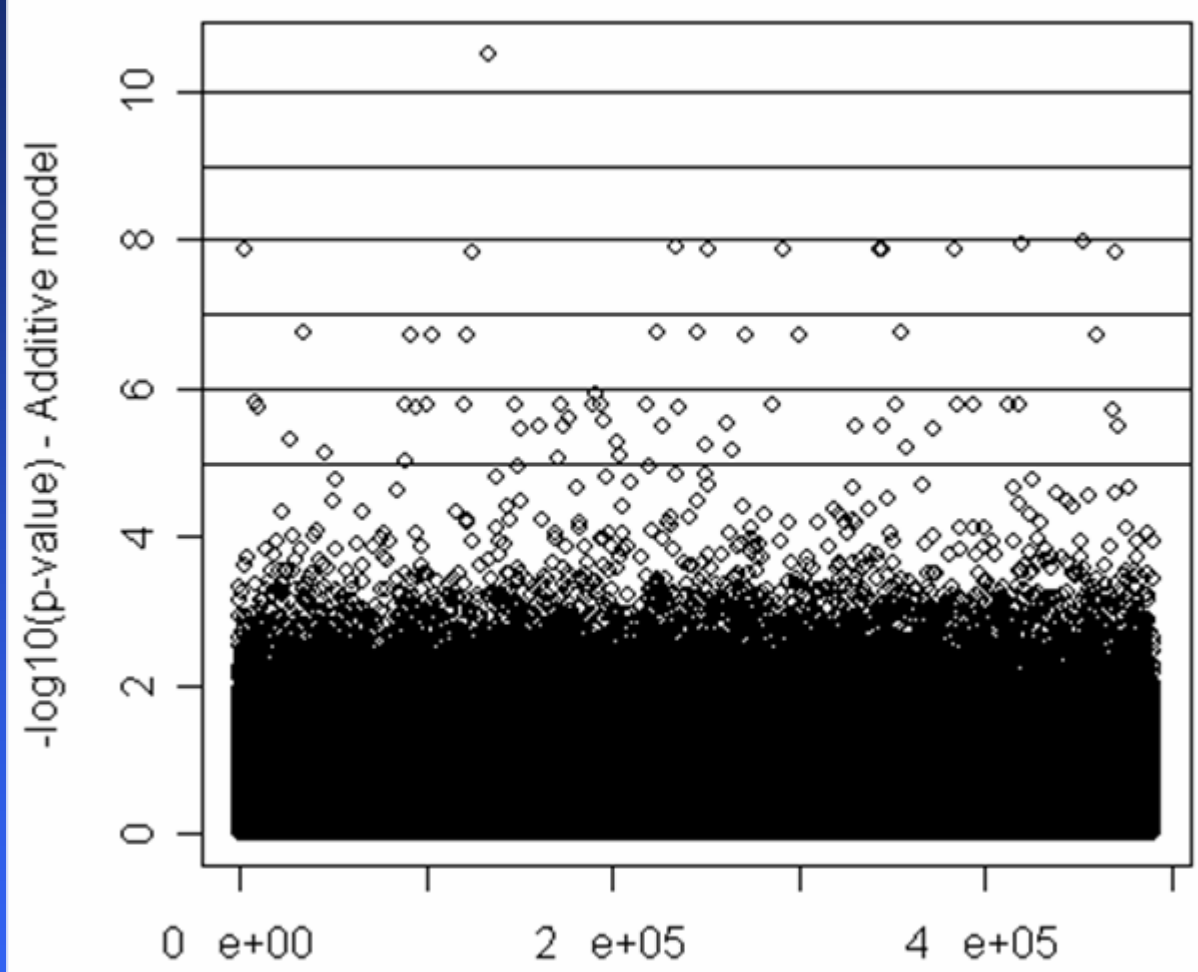
KORA 500k Statistical analysis - a globally significant p-value

- **Bonferroni: $p < 10^{-7}$**
  - But there is LD in the data
  - For less frequent variants (MAF < 10%) difficult
  - For more moderate effects (OR < 1.5) difficult
- **Two-stage design (plus replication = three-stage design):**
  - validation of SNPs with $p > 10^{-7}$ (joint analysis)
  - e.g. 500 SNPs with $p < 0.001$

KORA 500k Status July 2006
Description and first analyses

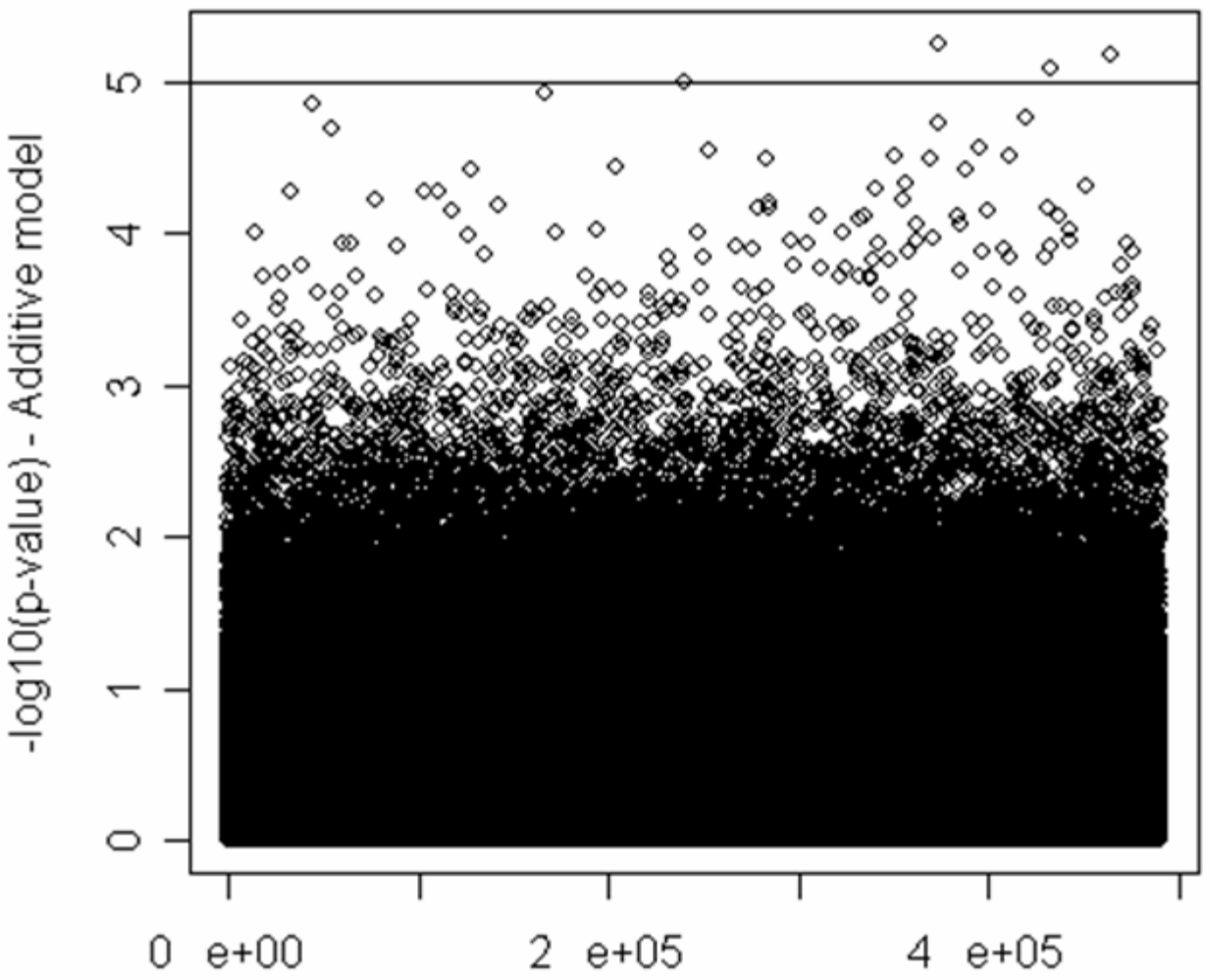
- Comparison between full sample (N=1796) and current sub-sample (N=1138)
- First preliminary analyses with selected phenotypes (N = 1138, including all younger subjects)
- Meanwhile data set complete: N=1644

KORA 500k Status July 2006
First preliminary analysis



Quantitative Phenotype A

KORA 500k Status July 2006
First preliminary analysis



Qualitative Phenotype B

KORA 500k Project Coordination, Genotyping and Data Management Teams

- Coordination
  - H.-Erich Wichmann
  - Thomas Meitinger
- Genotyping (GSF-Human Genetics + Epi)
  - Peter Lichter
  - Thomas Illig
- Data management (GSF-Epi)
  - Christian Gieger
  - Guido Fischer
  - Michael Putz
  - Michael Besler
  - Harald Grallert

KORA 500k Project
Data analysis team

- **GSF-Institute of Epidemiology**
  - Martina Müller
  - Claudia Lamina
  - Cornelia Huth
  - Iris M. Heid
  - Christian Gieger
- **Institutes in Munich (TUM, LMU, MPI), Lübeck, Marburg, Regensburg, Leipzig**