Über Korrelationsstrukturen bei SNP-Assoziationsanalysen

Arnd Groß

Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE)

Dissertationsverteidigung

Inhalt

Motivation

Isolatpopulationen interessant für Disease-Mapping

- Geringere Anzahl kausaler Varianten
- Homogenere Umwelteinflüsse
- Homogenere Bereiche im Genom

Studien

- Sorben (N=977)
- KORA (N=1644)

Ziele

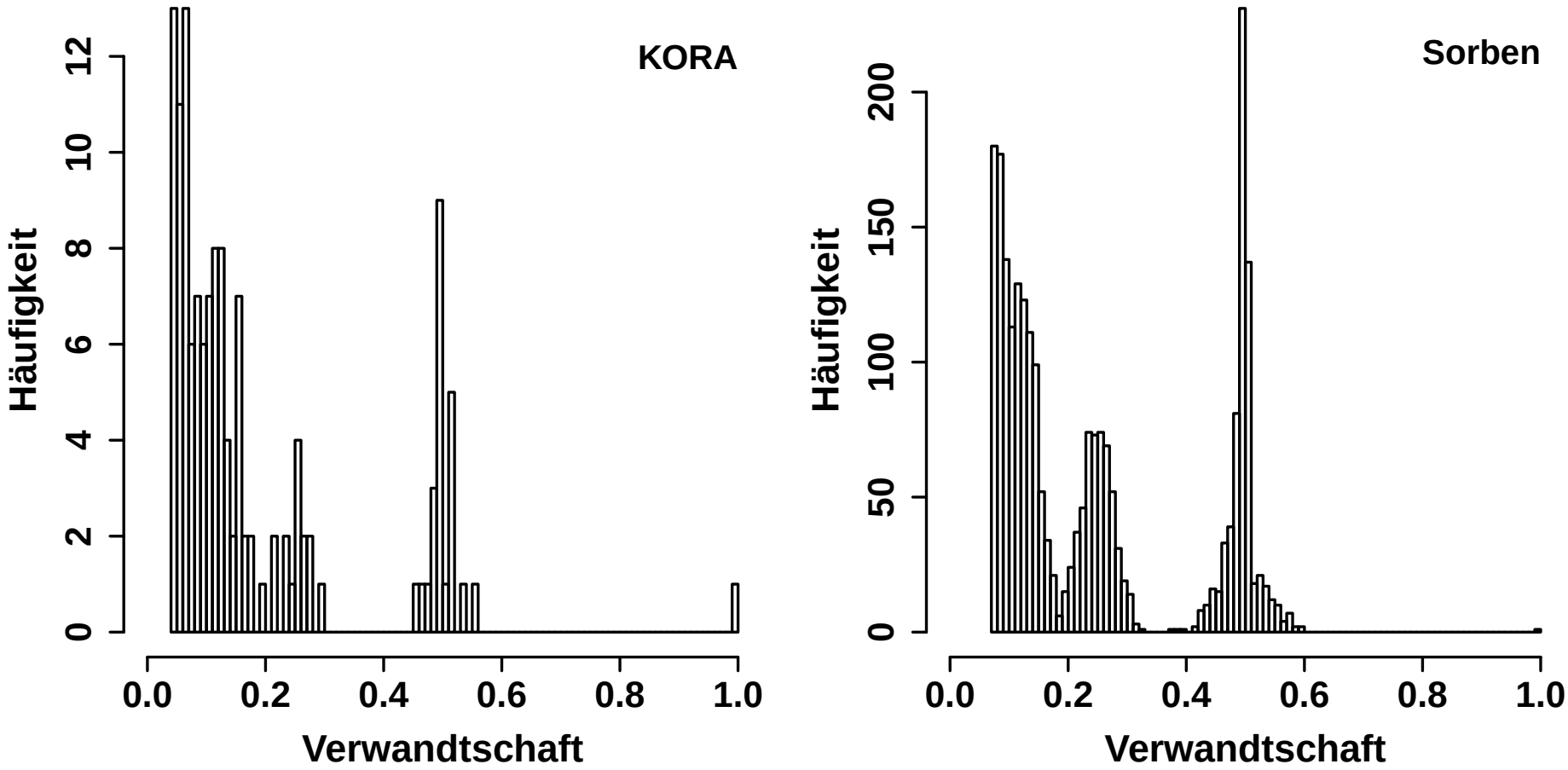
- Analyse Isolatcharakter der Sorben
- Bedeutung für genetische Assoziationsanalysen

Methoden

Analysen auf Grundlage von SNP-Arrays

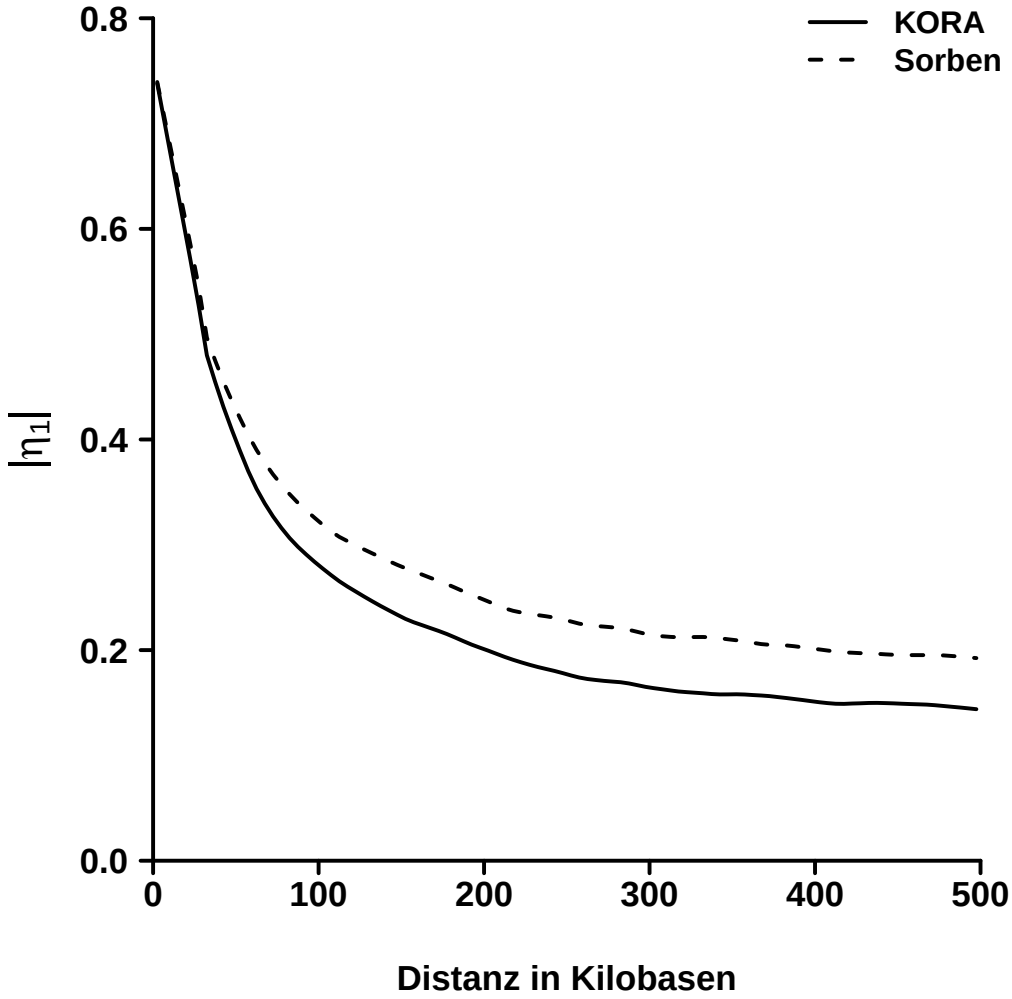
- Hauptkomponenten (PCA)
- Seltene SNPs
- F-Statistiken
- Runs of homozygosity (ROH)
- **Paarweise Verwandtschaft**
- **Kopplungsungleichgewicht (LD)**
- **Power von Assoziationsanalysen**

Paarweise Verwandtschaft



Top 0,01% der paarweisen Verwandtschaften von KORA und Top 0,5% der Sorben

Kopplungsungleichgewicht (LD)



η_1 gemittelt über alle SNP-Paare von Chromosom 22

Power von Assoziationsanalysen

LD wichtig für indirekte genetische Assoziation

- Simuliere Phänotyp für kausalen SNP
- Teste Phänotyp mit SNP im maximalen LD zum kausalen SNP
- Bestimme Power des Tests

Erklärte Varianz	Signifikanzniveau	Studie	1. Quartil	Median	3. Quartil
2%	10^{-5}	KORA	6.70	37.10	48.40
2%	10^{-5}	Sorben	10.08	38.95	48.90
5%	10^{-7}	KORA	24.78	88.30	95.12
5%	10^{-7}	Sorben	27.30	83.60	91.80

Simulation der Power für alle SNPs von Chromosom 22

Zusammenfassung

- Sorben zeigen moderate Merkmale genetischer Isolation
- Slawischer Ursprung der Sorben erkennbar
- Im Mittel höheres LD führt nicht zu klarem Vorteil bei Power
- Verwandtschaftsstruktur führt zu Varianzinflation des Effektschätzers und beeinflußt Power in komplexer Weise

Untersuche Zusammenhang zwischen Verwandtschaftsstruktur, Varianzinflation und Power analytisch!

Motivation

Analytische Beschreibung der Zusammenhänge aus empirischen Studien

- Verwandtschaftsstruktur, Heritabilität und Effektschätzer
- Allelfrequenz
- Fehler erster Art
- Power des Tests
- Verletzung Grundannahme unabhängiger Beobachtungen

Studien

- HapMap Trios (N=129)
- Sorben (N=977)
- Synthetische Familienstudien (N=129-999)

Modelle

Wahres Modell des Phänotyps

$$\mathbf{y} = b_1 + b_2\mathbf{s} + \mathbf{g} + \mathbf{e}$$

Analysiertes Modell des Phänotyps

$$\mathbf{y} = \beta_1 + \beta_2\mathbf{s} + \boldsymbol{\epsilon}$$

Es läßt sich zeigen

$$\begin{aligned} \mathrm{E}(\hat{\beta}_2) &= b_2 \\ \mathrm{V}(\hat{\beta}_2) &= \frac{\lambda}{1 - R^2_{\mathrm{h}}} V_\beta \end{aligned}$$

mit Inflationsfaktor λ, Heritabilität R^2_{h}, Varianz V_β ohne Heritabilität

Varianzinflation

Erwartete Varianzinflation

$$\lambda = 1 + R_{\mathrm{h}}^2 \frac{\sum_i \sum_{j \neq i} G_{ij}^2 - \frac{2}{n} \sum_i \left(\sum_{j \neq i} G_{ij} \right)^2}{n - 1}$$

Eigenschaften

- Stärkere Verwandtschaften **G** erhöhen Inflation
- Größere Heritabilität R_{h}^2 erhöht Inflation
- Inflation ist unabhängig von Allelfrequenz
- Ähnliches λ durch verschiedenes **G** und R_{h}^2

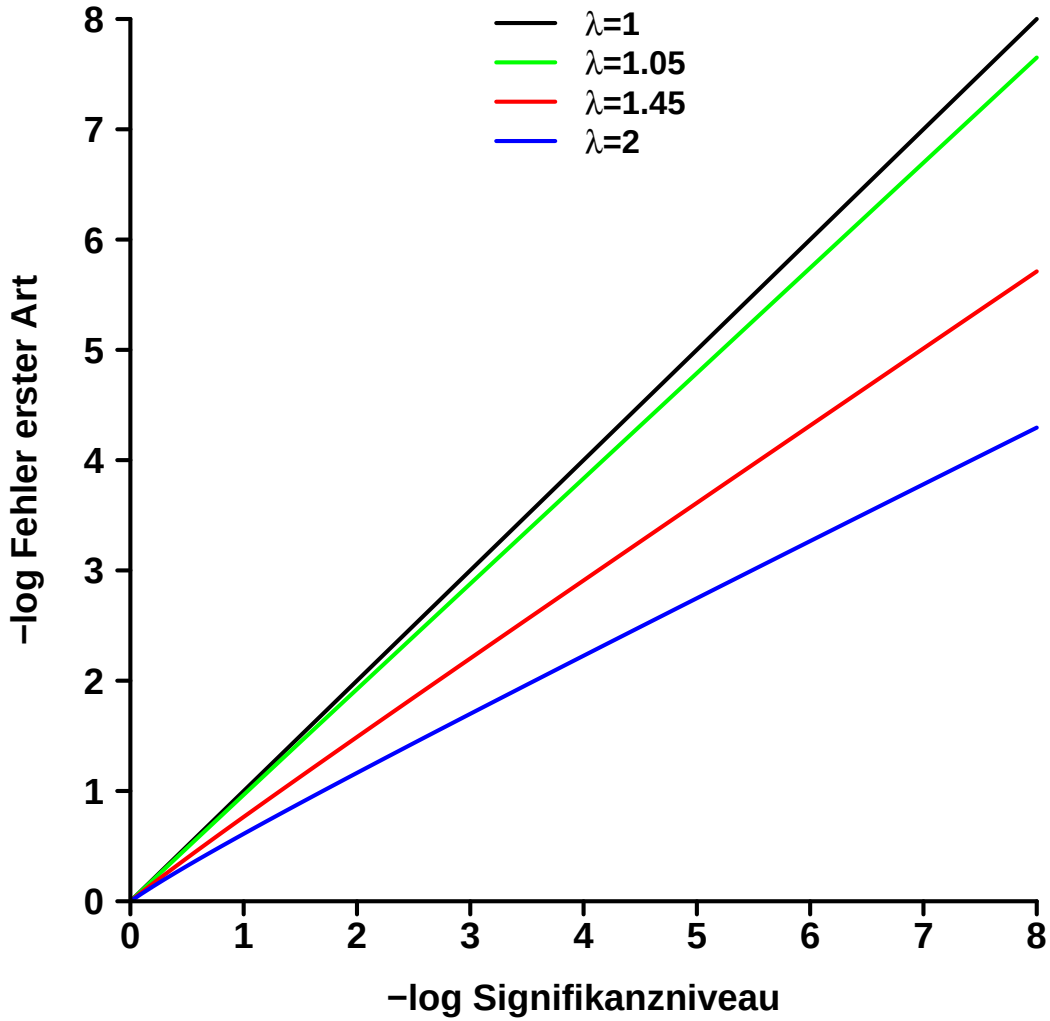
Fehler erster Art

Unter Nullhypothese
$b_2 = 0$ gilt näherungsweise

$$T \sim \mathrm{N}(0, \lambda)$$

für Teststatistik

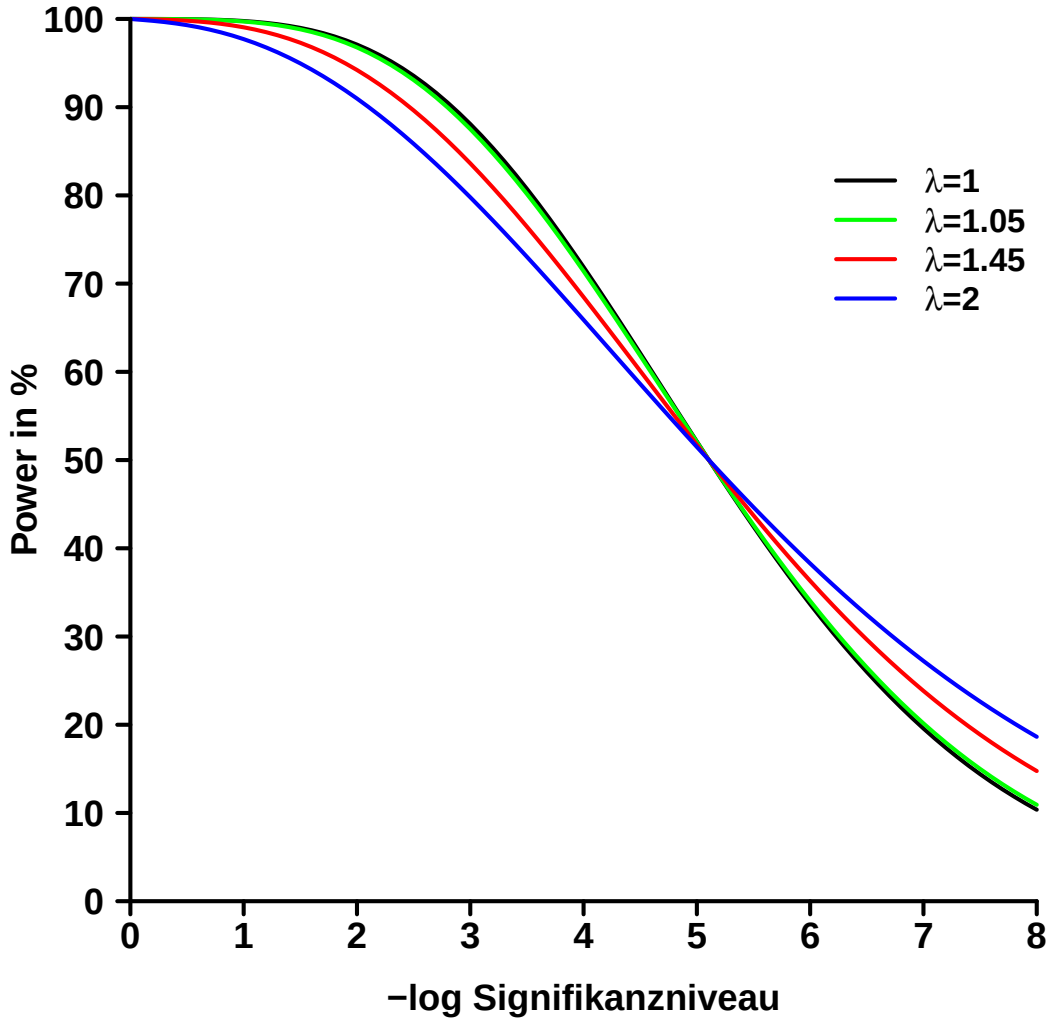
$$T = \frac{\hat{\beta}_2}{S_\beta}$$



Power des Tests

Unter Alternativhypothese $b_2 \neq 0$ gilt näherungsweise

$$T \sim \mathrm{N}\left(\sqrt{(n-1)R_{\mathrm{s}}^2}, \lambda\right)$$



$n = 1000$ und erklärte Varianz $R_{\mathrm{s}}^2 = 0.02$

Zusammenfassung

Herleitung analytischer Formeln für empirische Beobachtungen

Einfluß auf Inflationsfaktor λ

- λ wird größer durch stärkere Verwandtschaften
- λ wird größer durch größere Heritabilität
- λ ist unabhängig von Allelfrequenz

Inflationsfaktor λ beeinflußt Testen

- Größeres λ führt zu größerem Fehler 1. Art
- Power wird größer oder kleiner in Abhängigkeit vom Signifikanzniveau
- Aber Inflationsfaktor < 1.05 vernachlässigbar, sonst Analyse und Test mit gemischtem Modell

Motivation

Vergleich klassischer und bayesianischer Assoziationsanalyse an Kinderstudie (N=594) zu genetischen Ursachen für Stoffwechselparameter

- Lipidkonzentrationen
  - High density lipoprotein cholesterol (HDL-C)
  - Low density lipoprotein cholesterol (LDL-C)
  - Total cholesterol (TC)
  - Triglyceride (TG)
- Kandidaten-SNPs
  - rs599839 (SORT1)
  - rs3846663 (HMGCR)
  - rs3812316 (MLXIPL)
  - rs174570 (FADS2)
  - rs4420638 (APOE)
  - rs6102059 (MAFB)
- Kovariablen Alter, BMI und Geschlecht

Klassische (frequentistische) Analyse

Eigenschaften

- Analyse jeder Kombination aus SNP, Phänotyp, genetischem Modell
- Große Zahl von Tests, Korrektur für multiples Testen
- Unübersichtliche Ergebnistabellen

Phenotype	Variant	N	Add Beta	Add SE	Add p-val	Dom Beta	Dom SE	Dom p-val	Rec Beta	Rec SE	Rec p-val
HDL-C	...	...	...	...	...	...	...	...	...	...	...
LDL-C	**rs599839**	**576**	**-0.2996**	**0.0668**	**8.82e-06**	**-0.4266**	**0.1725**	**0.0137**	**0.3555**	**0.0824**	**1.90e-05**
LDL-C	rs3846663	572	0.1196	0.0619	0.0539	0.3028	0.1262	0.0167	-0.0875	0.0848	0.3025
LDL-C	rs3812316	564	-0.0425	0.0935	0.6494	-0.1414	0.4042	0.7267	0.0406	0.1008	0.6872
LDL-C	rs174570	578	0.0212	0.0913	0.8165	-0.2008	0.3545	0.5714	-0.0415	0.1002	0.6785
LDL-C	**rs4420638**	**584**	**0.3819**	**0.0783**	**1.38e-06**	**0.6728**	**0.2789**	**0.0161**	**-0.4057**	**0.0873**	**4.18e-06**
LDL-C	rs6102059	575	0.0180	0.0648	0.7811	0.3394	0.1493	0.0233	0.0739	0.0827	0.3715
TC ...	...	...	...	...	...	...	...	...	...	...	
TG	rs599839	576	-0.1135	0.0650	0.0811	-0.2432	0.1659	0.1433	0.1157	0.0801	0.1493
TG	rs3846663	572	-0.0065	0.0593	0.9130	0.0888	0.1209	0.4631	0.0519	0.0810	0.5218
TG	rs3812316	564	-0.1342	0.0878	0.1269	-0.8484	0.3784	0.0254	0.1030	0.0947	0.2770
TG	rs174570	578	0.1367	0.0868	0.1159	0.1770	0.3378	0.6005	-0.1504	0.0953	0.1148
TG	rs4420638	584	0.1352	0.0760	0.0758	0.3752	0.2671	0.1607	-0.1299	0.0847	0.1258
TG	rs6102059	575	0.0754	0.0620	0.2245	0.1639	0.1436	0.2544	-0.0732	0.0793	0.3564

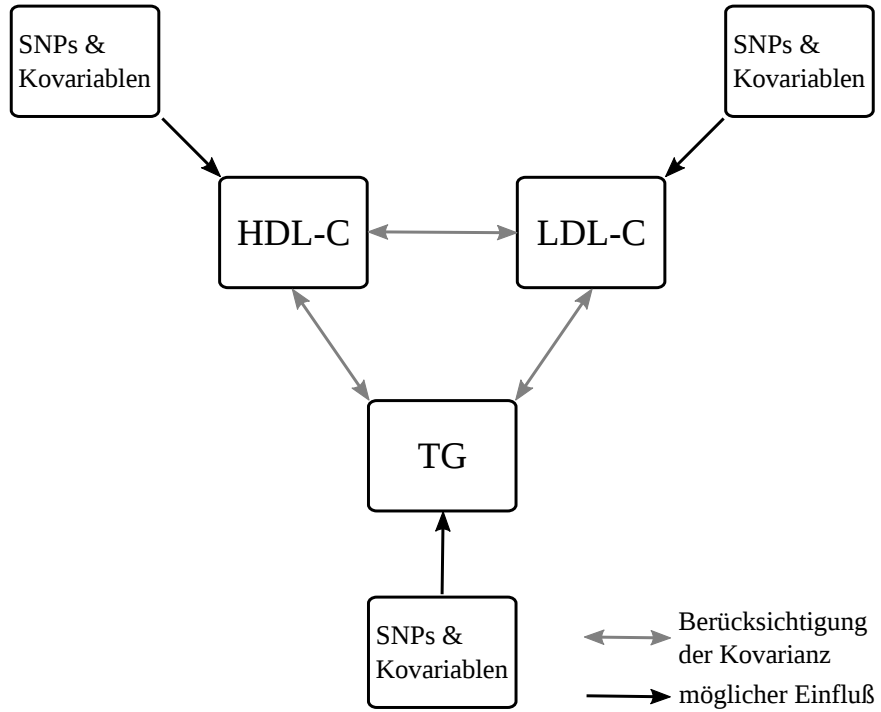
Bayesianische Analyse

Eigenschaften

- Bayesianisches Paradigma
- Modellierung Phänotyp-Korrelation

$$(\text{HDL-C}, \text{LDL-C}, \text{TG}) \sim \text{N}_3(\mu, \Sigma)$$

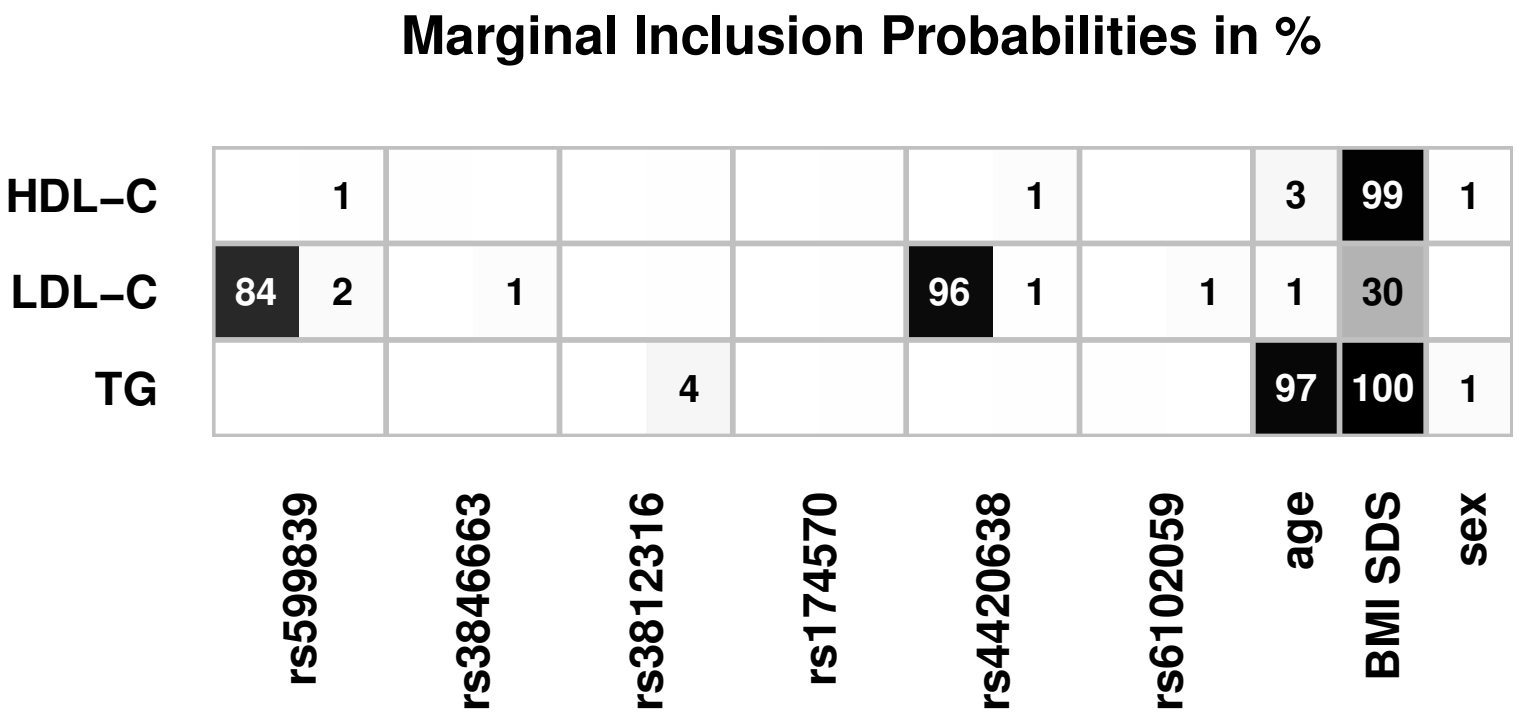
- Bayesianische Modellauswahl



Modelle

Phenotype	Model	Probability	Bayes factor
HDL-C	BMI SDS	91.89	371265
HDL-C	age, BMI SDS	3.08	1041
HDL-C	$\text{rs4420638}_{\text{dom}}$, BMI SDS	0.99	329
LDL-C	$\text{rs599839}_{\text{rec}}$, $\text{rs4420638}_{\text{rec}}$	53.49	37691
LDL-C	$\text{rs599839}_{\text{rec}}$, $\text{rs4420638}_{\text{rec}}$, BMI SDS	22.88	9720
LDL-C	$\text{rs4420638}_{\text{rec}}$	7.65	2714
LDL-C	$\text{rs4420638}_{\text{rec}}$, BMI SDS	4.62	1586
LDL-C	$\text{rs599839}_{\text{rec}}$	2.54	855
LDL-C	$\text{rs599839}_{\text{dom}}$, $\text{rs4420638}_{\text{rec}}$	1.03	340
LDL-C	$\text{rs599839}_{\text{rec}}$, $\text{rs4420638}_{\text{rec}}$, age	0.80	266
LDL-C	$\text{rs599839}_{\text{rec}}$, $\text{rs4420638}_{\text{rec}}$, $\text{rs6102059}_{\text{dom}}$	0.77	254
LDL-C	$\text{rs599839}_{\text{rec}}$, BMI SDS	0.74	244
LDL-C	null	0.56	186
TG	age, BMI SDS	90.47	311171
TG	$\text{rs3812316}_{\text{dom}}$, age, BMI SDS	3.66	1247
TG	BMI SDS	2.55	856

Variablen


Marginal Inclusion Probabilities in %
HDL–C
LDL–C
TG
1
1
3
99
1
84
2
1
96
1
1
1
30
4
97
100
1
rs599839
rs3846663
rs3812316
rs174570
rs4420638
rs6102059
age
BMI SDS
sex

Zusammenfassung

Zusammenhang von rs599839 (SORT1) und rs4420638 (APOE) mit LDL-C bei Kindern

Vorteile Bayesianische Analyse

- Plausible Auswahl von Einflußfaktoren aus SNPs und Kovariablen
- Berücksichtigung Korrelationsstrukturen in Phänotypen und SNPs
- Berücksichtigung verschiedener genetischer Modelle
- Verbesserte Identifikation von Phänotyp-Genotyp-Beziehungen
- Präsentation der Ergebnisse in verständlicher Form

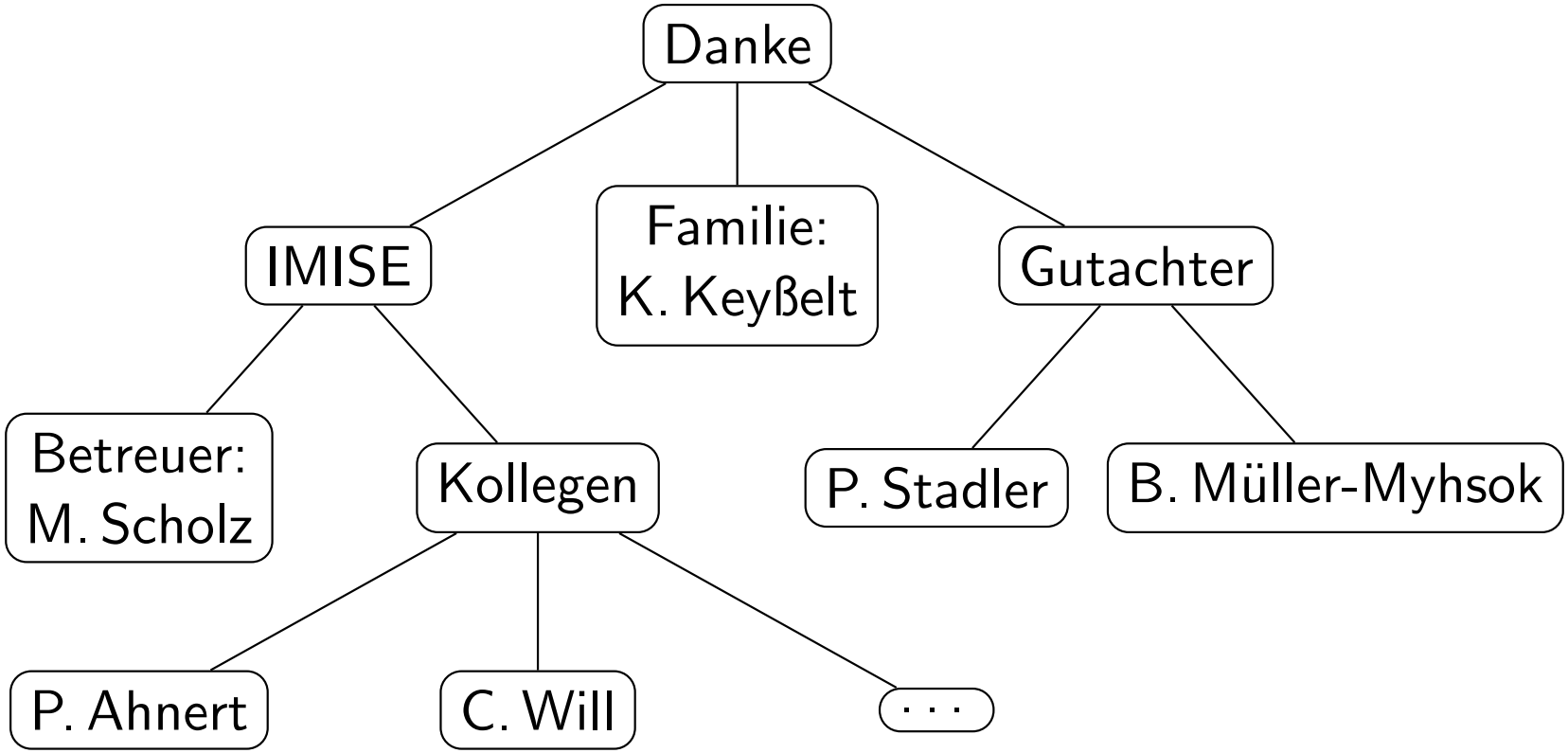
Ausblick

Vergleich klassische und bayesianische Analyse

- Einfluß Korrelationsstrukturen auf Identifikation genetischer Effekte
- Einfluß Anzahl und Stärke genetischer Effekte auf Identifikation
- Einfluß Fehlwerte auf Identifikation

Sind Korrelationsstrukturen ein

Bug oder Feature?


Danke
IMISE
Familie:
K. Keyßelt
Gutachter
Betreuer:
M. Scholz
Kollegen
P. Stadler
B. Müller-Myhsok
P. Ahnert
C. Will
· · ·