

POPULATION APPROACH IN HIGH-THROUGHPUT PHARMACOGENETICS: CHALLENGING THE MAXIMUM LIKELIHOOD APPROACHES AND EXPLORATION OF A BAYESIAN ALTERNATIVE

Julie Bertrand¹, Maria de Iorio², David J. Balding¹

¹ Genetics Institute, University College London, London, UK
² Statistical Science department, University College London, London, UK



PHARMACOGENETICS

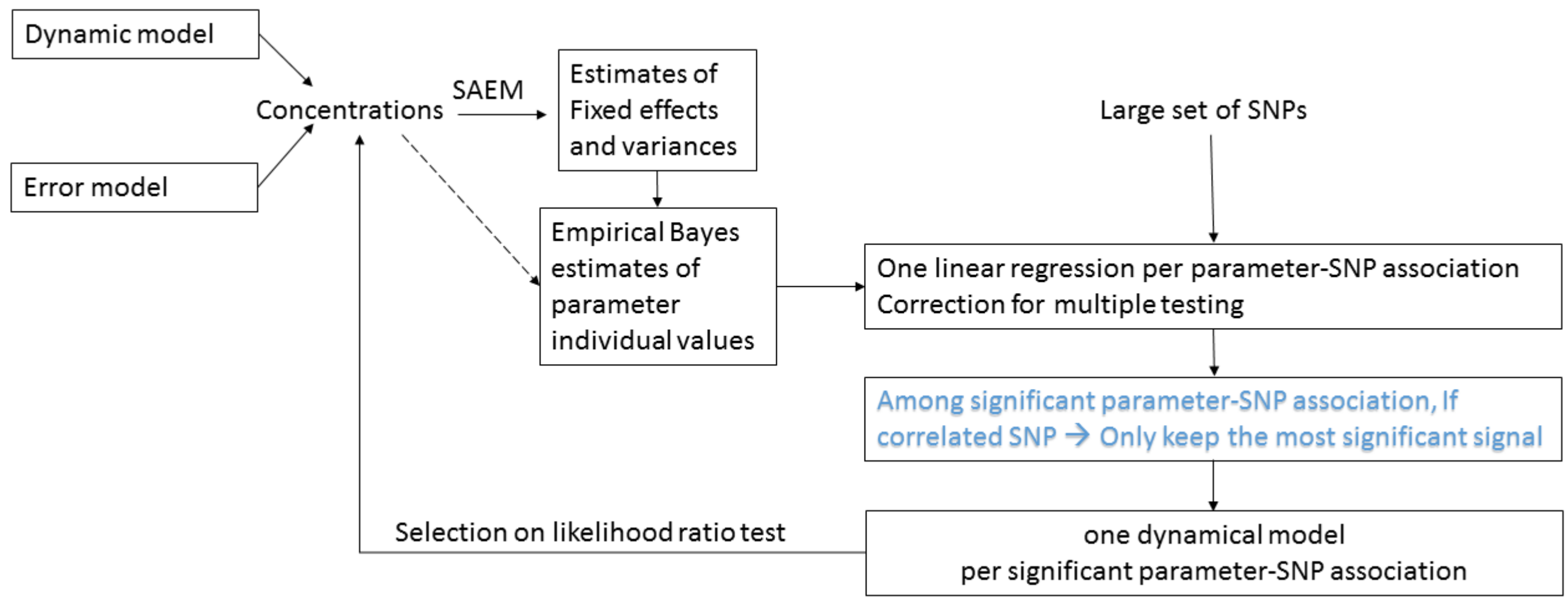
- Genetic part of pharmacokinetic/pharmacodynamic (PK/PD) variability
 - genetic variation coding for proteins involved in PK/PD processes
 - single nucleotide polymorphism (SNP)
- [Stratified medicine](#)

METHODOLOGICAL CHALLENGES IN PHARMACOGENETICS

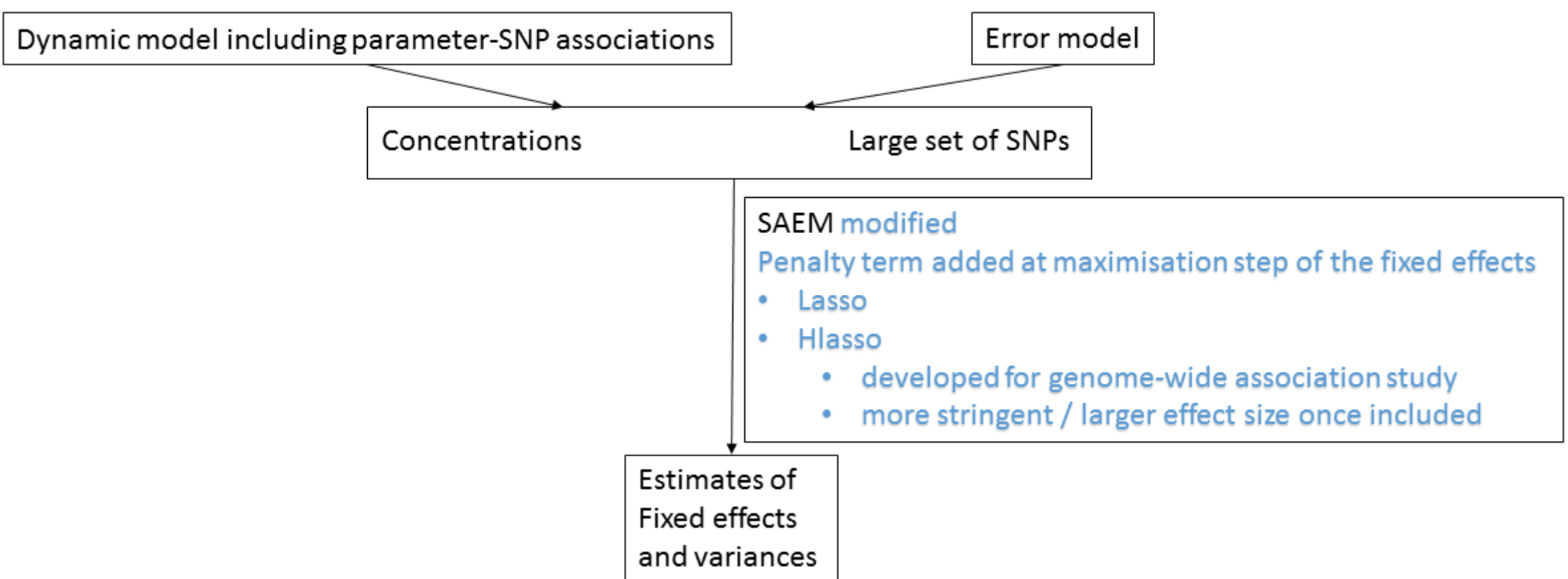
- PK/PD phenotype → not observed
 - data : plasma or insulin concentrations, ...
→ [dynamical models](#)
- Variable informativeness of genetic markers
 - uneven distribution, small sample size of some genotypes
→ [mixed effect models](#)
- Large genetic data sets in high throughput screening
 - dimensionality curse $N \ll p$
 - structural correlation along the genome (linkage disequilibrium)
→ [Need for new approaches bridging statistical genetics with dynamical mixed effect models](#)

MAXIMUM LIKELIHOOD APPROACHES

STEPWISE PROCEDURE



INTEGRATED APPROACH



SIMULATION STUDY

- Two hundred data sets under the null H_0 and the alternative H_1 hypothesis

Genetic variation

- generation of genotypes using HAP-GEN (Su, 2011)
- $N_s=1227$ SNPs on 171 genes from the DMET Chip (Daly, 2007)
 - 6 [1-56] snps per gene
- HAPMAP caucasian reference haplotypes

Pharmacokinetic profiles

- inspired from real study (Kappelhoff, 2005)
 $T_{lag}(h) = 1$ (30%)
 $K_a(h) = 3.07$ (40%)
 $Q/F(L/h) = 62.9$ (60.8%)
 $V_c/F(L) = 189$ (40.1%)
 $V_p/F(L) = 233$ (0%)
 $CL/F(L/h) = 11.7$ (39.7%)
Dose = 600 (mg) once daily
- diagonal variance matrix of random effects and combined residual error model

- Scenarios

	Sample size	Sampling times	Alternative hypothesis
1: phase II study	N=300	n=6	6 unobserved causal variants ↓ CI
2: phase II study with additional samples collected in clinical routine	N=300 N=700	n=6 n=1	$R_{GC}=1,2,3,5,7,12\%$
3: phase II study	N=300	n=6	4 correlated unobserved causal variants SNP1 and 3 ↓ CI, SNP2 and 4 ↓ Vc $R_{GC}=15,15,15,15\%$

→ Genetic association explored on V_c , CI and Q

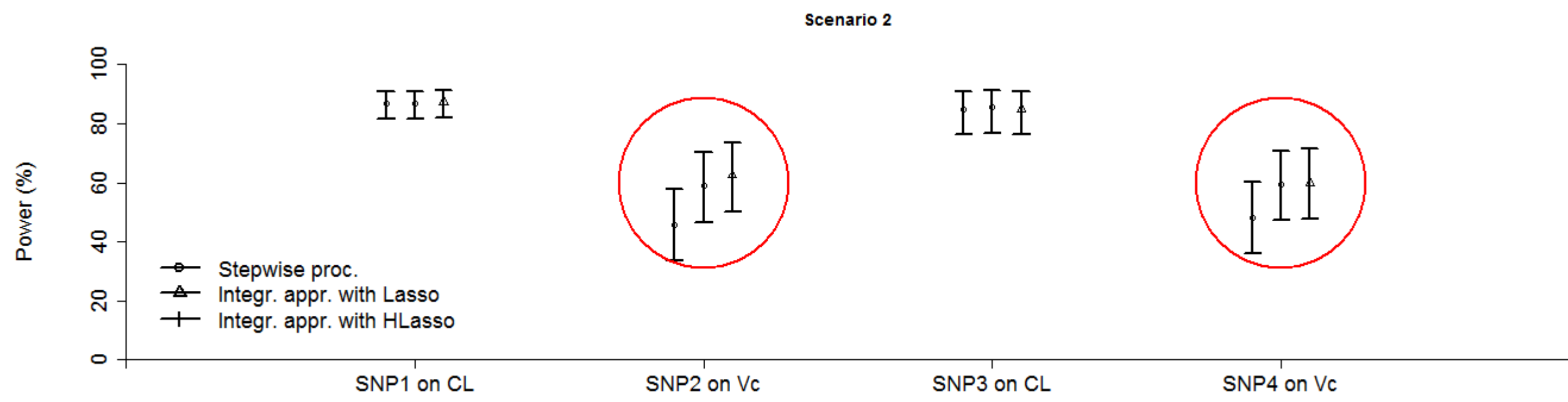
RESULTS								
Method	FWER (%)	TP	FP _{CL}	FP _V	FP _Q	Computing times (h)		
						H ₀	H ₁	
Scenario 1								
Stepwise procedure	19.0	366	15	7	5	0.05 [0.05-0.24]	0.24	[0.06-1.09]
Integrated approach with Lasso	20.0	340	7	6	0	1.04 [0.81-1.48]	1.14	[0.83-1.61]
Integrated approach with HLasso	22.5	335	4	7	1	1.08 [0.81-1.57]	1.19	[0.84-1.60]
Scenario 2								
Stepwise procedure	22.5	736	40	11	5	0.16 [0.11-0.93]	2.31	[0.31-8.53]
Integrated approach with Lasso	21.0	722	12	4	0	1.79 [1.40-2.48]	1.90	[1.47-3.73]
Integrated approach with HLasso	21.5	720	15	6	0	1.91 [1.44-3.14]	1.94	[1.58-4.19]
Scenario 3								
Stepwise procedure	-	531	18	7	22	-	0.84	[0.06-2.56]
Integrated approach with Lasso	-	582	2	13	22	-	1.51	[1.23-9.05]
Integrated approach with HLasso	-	590	0	10	24	-	1.62	[0.79-1.83]

Family wise error rate (FWER), expected value of 20[14.5–25.5]%
True positive (TP) = count of detected SNP in $r^2 \geq 0.05$ with a causal variant, with a maximum of $200 \times 6 = 1200$ on scenario 1 and 2 and $200 \times 4 = 800$ on scenario 3
False positive (FP)

→ [Scenario 1: more TP and FP with stepwise procedure](#)

→ [Scenario 2: similar TP across methods but still more FP with stepwise procedure](#)

→ [Scenario 3: more TP with integrated approach](#)



→ [More TP on SNP associated to \$V_c\$ with the integrated approach](#)

DISCUSSION

- Not enough information to detect realistic genetic effect** in classical phase II study designs

	Stepwise procedure	Integrated approach
SNP selection	After model parameter estimation	simultaneous to model parameter estimation
Parameter-SNP association	independent	altogether

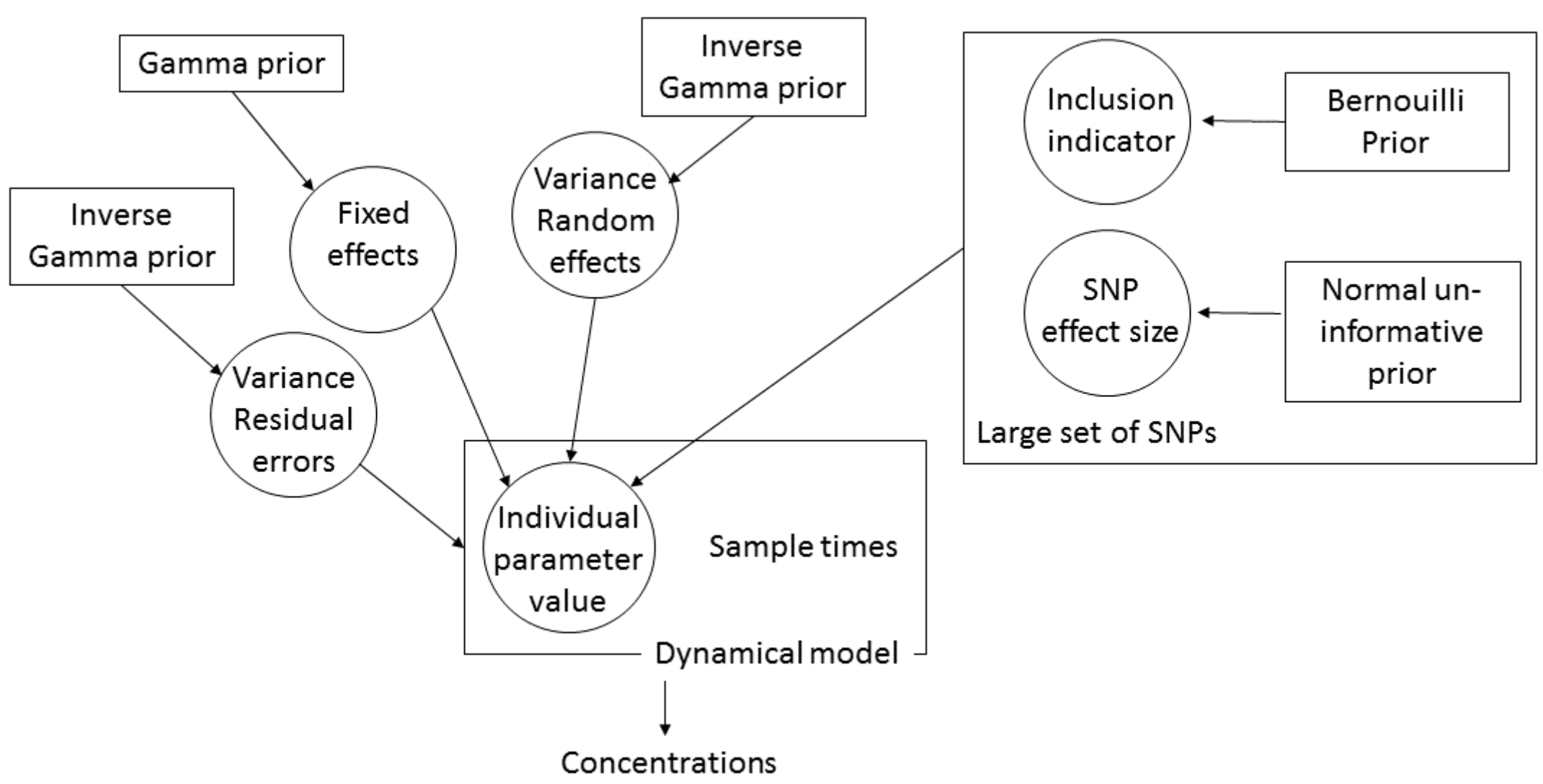
→ **Better performance of the integrated approach** with penalised regression on

- design combining full-profile and sparse sampling
- multiple parameter-SNP associations
- cost in computing time** equivalent when increasing sample size

- No difference between Lasso and Hlasso, because not on genome scale

BAYESIAN ALTERNATIVE

BAYESIAN VARIABLE SELECTION



- R package R2jags: 1 chain / 20000 total with 4000 burning and a thinning of 20 = 320 iterations
- SNP selection (positive) = mean posterior probability of inclusion indicator ≥ 0.75 (Eicher, 2007)

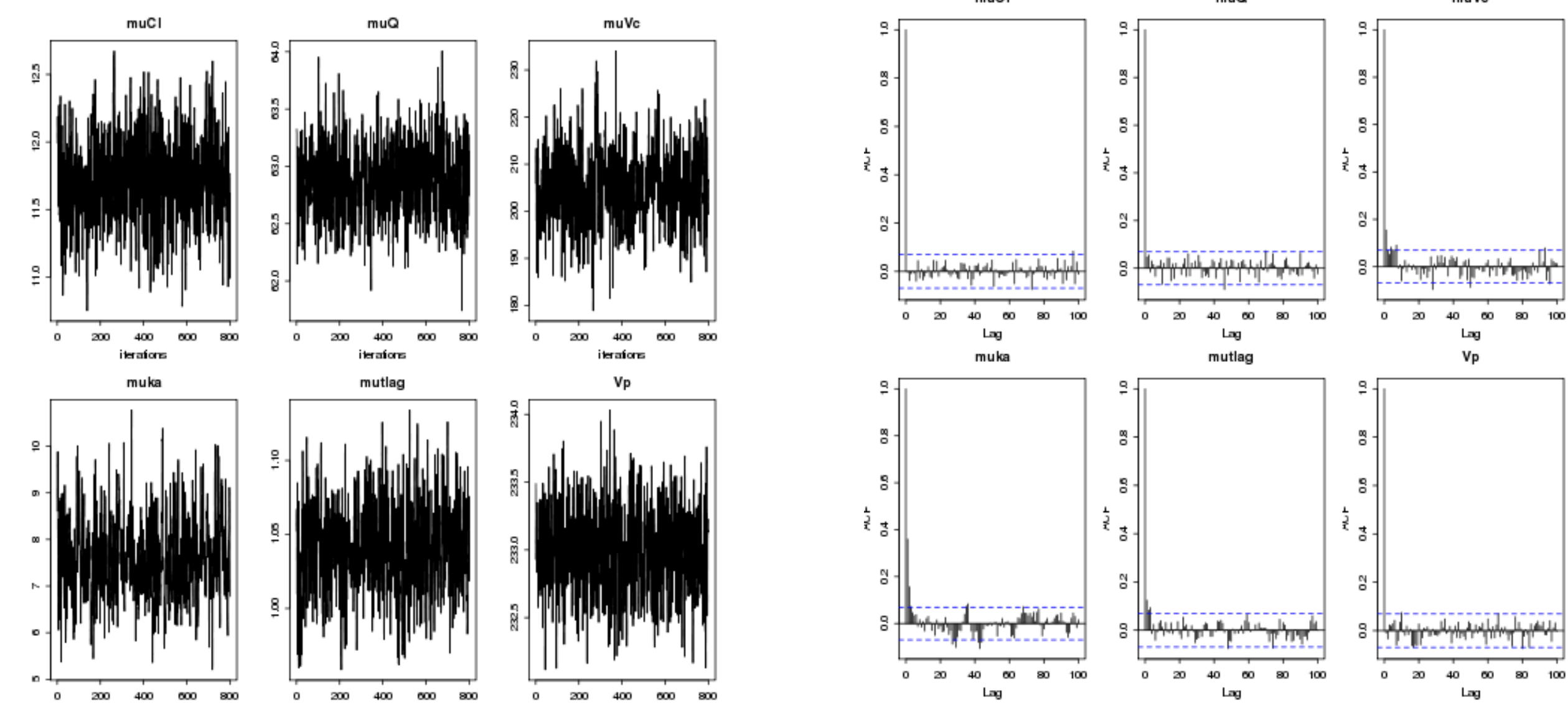
SIMULATION STUDY

- Two hundred data sets under the alternative H_1 hypothesis
- Genetic variation:** $N_s=106$ SNPs

	Sample size	Sampling times	Alternative hypothesis
Scenario 1	N=300	n=6	6 observed causal variants ↓ CI $R_{GC}=1,2,3,5,7,12\%$

- Bernoulli prior parameter reflecting the expected number of signals = $6/(N_s \times 3)$

RESULTS



→ [Stable center and small fluctuations](#)

→ [Good exploration of area of low density](#)

→ k_a and V_c requiring lag=20 for an autocorrelation < 10%

Method	TP	FP _{CL}	FP _V	FP _Q	Computing times (h)
Bayesian variable selection	388	10	0	3	1.82 [1.02-2.81]

→ [More TP than the maximum likelihood approaches, but \$N_s = 106\$ + causal variants observed](#)

→ [Computing time already longer than maximum likelihood approaches with \$N_s = 106\$ only](#)