Bayesian Variable Selection for high-throughput genetic association analysis in population pharmacokinetics

Julie Bertrand [1], Maria de Iorio [2], Celine Verstuyft [3], Monidarin Chou [4], Anne-Marie Taburet [5], David Haas [6] and David J Balding [1]

[1] University College London Genetics Institute, London, UK, [2] University College London Statistical Science Department, London, UK, [3] Departments of Molecular Genetics, Pharmacogenetic, and Hormonology, Assistance Publique Hopitaux de Paris, Hopital Bicetre, Paris, France, and Univ Paris-Sud, EA2706, France, [4] Rodolphe Merieux Laboratory, Faculty of Pharmacy, University of Health Sciences, Phnom Penh, Cambodia, [5] Clinical Pharmacy Department, Assistance Publique Hopitaux de Paris, Hopital Bicetre, Paris, France, [6] Center for Human Genetics Research, Departments of Medicine and Pharmacology, Pathology, Microbiology & Immunology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

Context

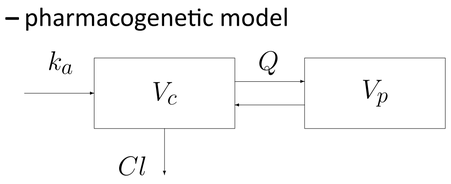
- Previous works have shown the need to increase the sample size of pharmacogenetic studies[1]
  - combination of genetic and pharmacokinetic data from several sources e.g. phase I,II and III studies
- Simultaneous estimation of pharmacokinetic parameters and genetic effect sizes using penalized regression can outperform the standard stepwise procedure[2]
  - implementation in a maximum likelihood framework (saemix[3]), not yet handling ODEs and inter-occasion variability
- Bayesian approaches are growing in importance in high-throughput genetic association studies
  - natural interpretation of penalized regression through prior distribution on effect sizes
  - can manage both complex data structure and missing genetic data
  - fast, robust and cross-platform programs now available such as JAGS [4] and Stan [5]

Objectives

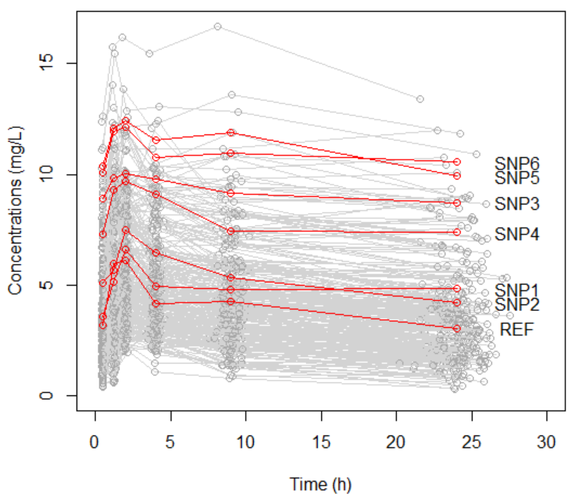
- Simulation study evaluating the selection performance and computing times of several Bayesian approaches
- Motivating real case study
  - PECAN ANRS 12154 study of steady-state nevirapine clearance among HIV-infected Cambodians [6]
  - substudy on additional polymorphisms contribution to variable nevirapine clearance in this cohort [7]

Simulation study

- Simulation settings
  - phase-II like study design with 300 subjects
    - * 6 sampling times (1,2,4,6 and 12 h)
    - * 1200 single nucleotide polymorphisms (SNP) from the DMET chip[8]
  - pharmacogenetic model



* normally distributed inter-individual random effects (η) on all parameters but V_p
* combined residual error model



* effect of 6 unobserved causal variants on logCl explaining 30% of its variability

$\log Cl_i = \log Cl + \sum_{s=1}^{6} \beta_{Cl_s} SNP_{s_i} + \eta_{Cl_i}$

with $SNP_{s_i} = 0, 1, 2$ the number of causal alleles

- Methods to select genetic markers
  - **stepwise procedure (SP)** using `saemix`
    - i) screening of SNPs on empirical Bayes estimates of individual parameters $\hat{\phi}_i$ using a Sidak correction
    - ii) forward model inclusion of significant SNPs on likelihood ratio test
    - iii) return to i), until no more significant SNP found
  - **SAEM with penalized regression (SAEMpr)** using `saemix`

    $\widehat{\mu_{k+1}, \beta_{s_{k+1}}} = argmin_{\mu,\beta_s} \sum_{i=1}^{N} \left(\phi_{ik} - \mu - \sum_{s=1}^{N_s} \beta_s SNP_i\right)^2 + P_\lambda(\beta_s)$

    $P_\lambda(\beta_s) \approx$ double exponential prior on β_s with λ set using an asymptotic approximation
  - **Indicator Model selection (IMs)**[9] using `JAGS`

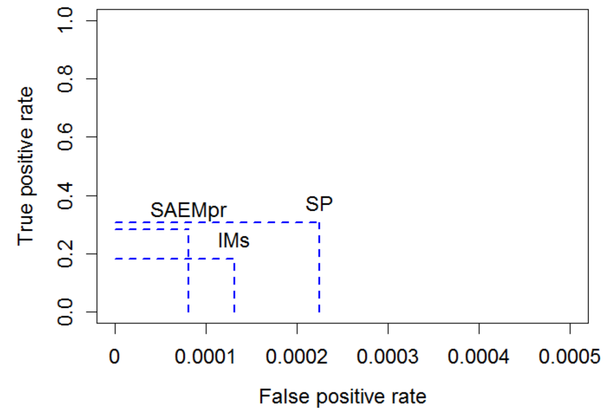
    $P(I_s, \beta_s) = P(I_s)P(\beta_s)$

    $P(\beta_s) = N(0, \sigma_{\beta_s})$ with a large σ_{β_s}

    $P(I_s) =$ Bernoulli(p_{I_s}) an indicator variable with p_{I_s} set empirically

    Note: for **SAEMpr** and **IMs** the SNPs are centered and standardized
- Evaluation
  - explore association on 3 parameters: Cl, V_c and Q on 200 simulated data sets
  - true positive (TP) = any significant SNP which is correlated with a causal variant with an $r^2 \geq 0.05$

- Results

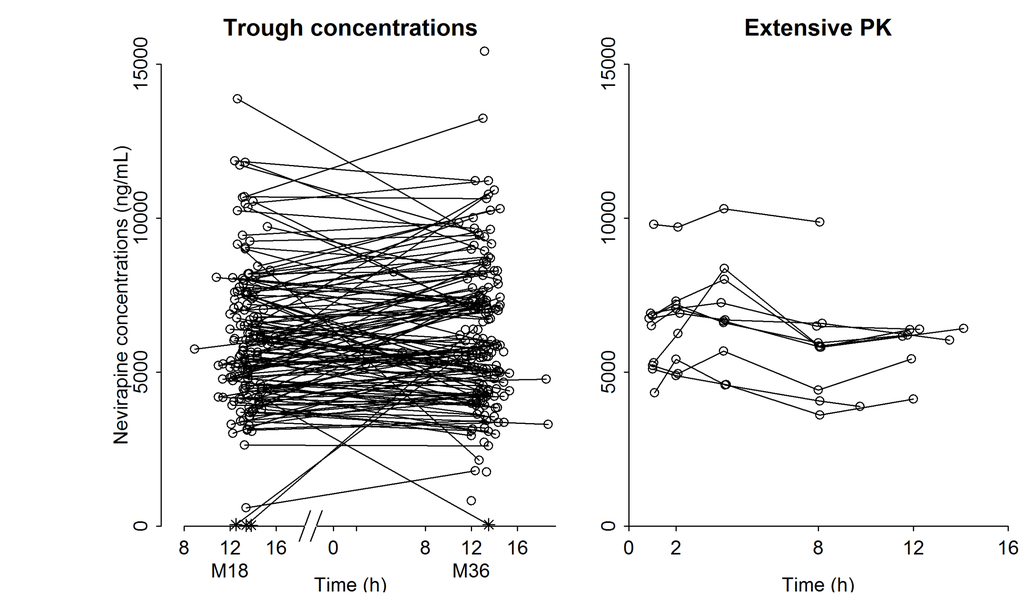


  - mean [range] computing times in hours

Stepwise procedure	**0.24 [0.06 - 1.09]**
SAEMpr	**1.14[0.83 - 1.61]**
IMs	**19.58 [11.51 - 23.12]**

Motivating real case study

- Pharmacogenetic data



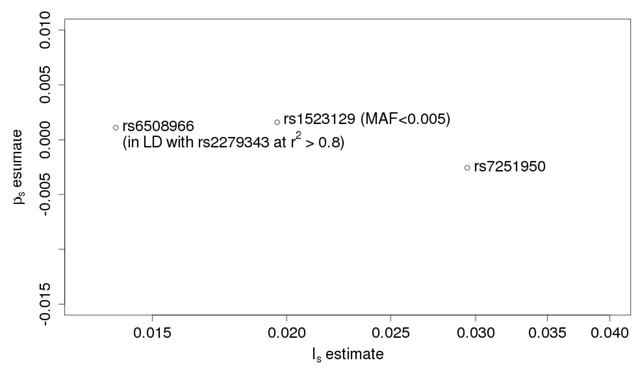
  - 129 patients on up to 3 occasions with 196 markers

Chromosome	3	7			19	
Gene	*NR1I2* (PXR)	ABCB1 (P-gp)	*CYP3A5*	*CYP3A4*	*CYP2A6*	*CYP2B6*
Number of markers	49	63	1	36	1	47

  - 218 missing polymorphisms with a maximum of 7 per subject)
- Analyses
  - one compartment model with 1st-order absorption and elimination
  - inter-individual and inter-occasion variability on clearance
  - adjustment for rs3745274 polymorphism on *CYP2B6*
  - methods to select genetic markers and handle missing data:
    - * stepwise procedure on empirical Bayes estimates with missing data removed [7]
    - * Indicator Model selection with missing data imputed in `JAGS` from Binomial with empirical allele frequency
- Results
  - **stepwise procedure**

rs number	β_s estimate	p-value
rs2279343	**0.835**	**1.66e-5**
rs7251950	**0.404**	**1.09e-4**

  - **Indicator Model selection**



Conclusions

- IMs initial simulation study results not yet competitive

↪ Future works: other indicator-based model selection and shrinkage prior on genetic effect sizes

[1] Tessier, Bertrand, Chenel, Comets. *AAPS J* (2015).
[2] Bertrand, de Iorio, Balding. *PGEN* (2014).
[3] Comets, Lavenu, Lavielle. *PAGE* 20^{th} (2011).
[4] Plummer. `JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling` (2003).
[5] Stan Development Team. `Stan: A C++ Library for Probability and Sampling, Version 2.5.0` (2014).
[6] Bertrand et al. *Antimicrob Agents Chemother* (2010).
[7] Bertrand et al. *Pharmacogenet Genomics* (2012).
[8] Daly et al. *Clinical Chemistry* (2007).
[9] Kuo & Mallick. *Sankhya Ser. B* (1998)

During this work, Julie Bertrand was supported by a fellowship from the UK Medical Research Council.