

**Supplementary material to ‘Combined analysis of phase I and phase II data to enhance the power of pharmacogenetic tests’ by Tessier et al.**

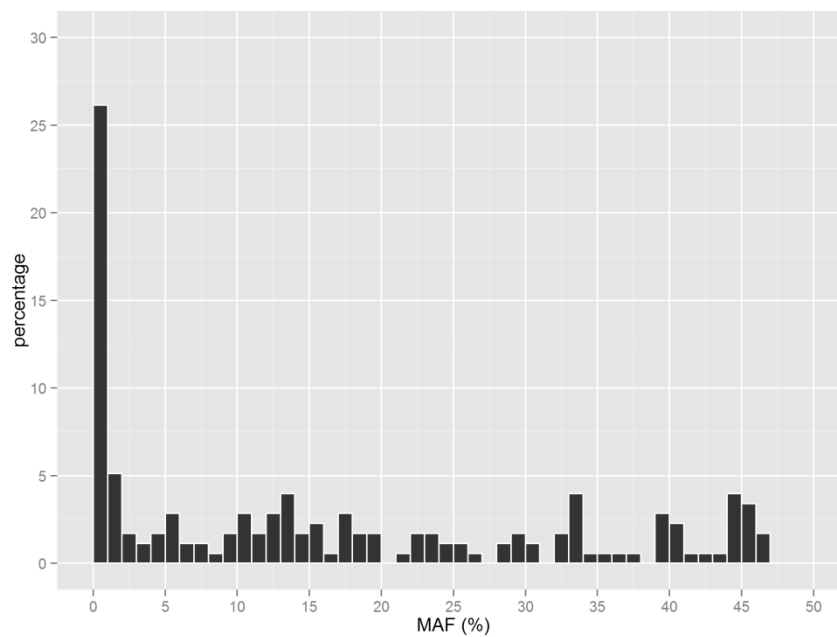
**The supplementary material for the paper covers the following elements:**

- Additional details on the simulation study materials and methods
  - Simulated polymorphisms
  - Pharmacokinetic model used for simulations of concentrations profiles
  - Computation of the correlation coefficient in the stepwise procedure
  - Evaluation
  - Softwares
- Additional results for scenarios presented in the main manuscript
  - Simulated genetic effects
  - Estimation bias
- Additional results for the two other penalised regression methods
- Additional references

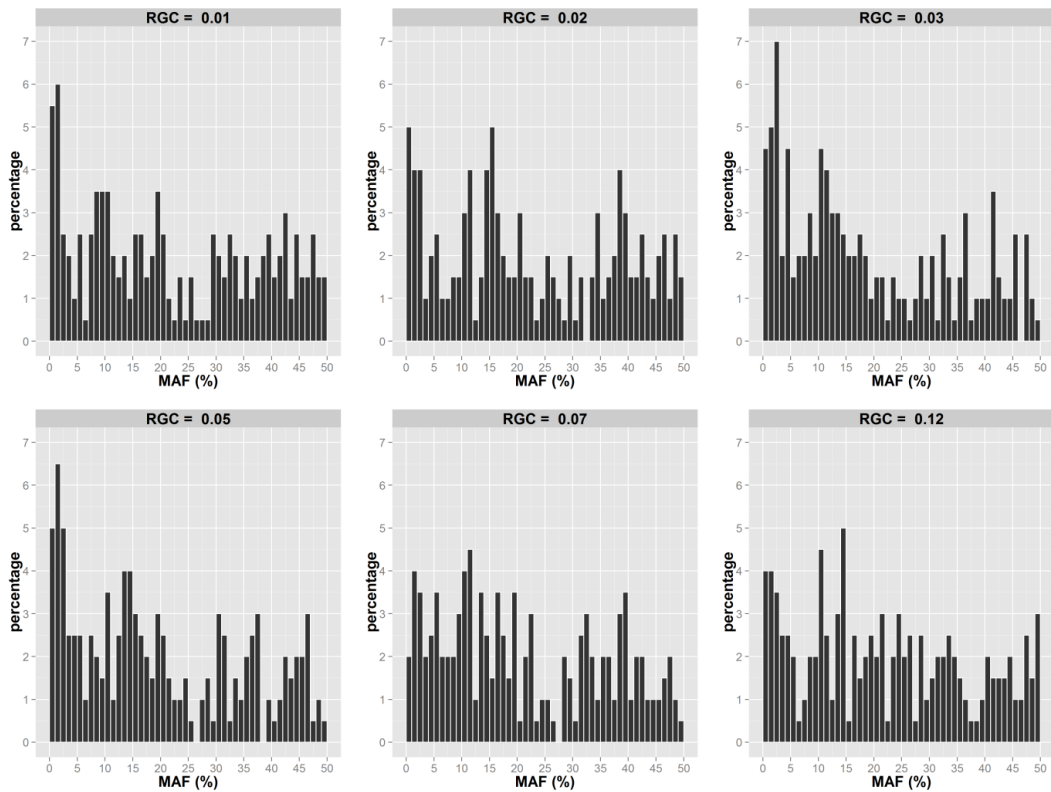
## Additional details on the simulation study materials and methods

### Simulated polymorphisms

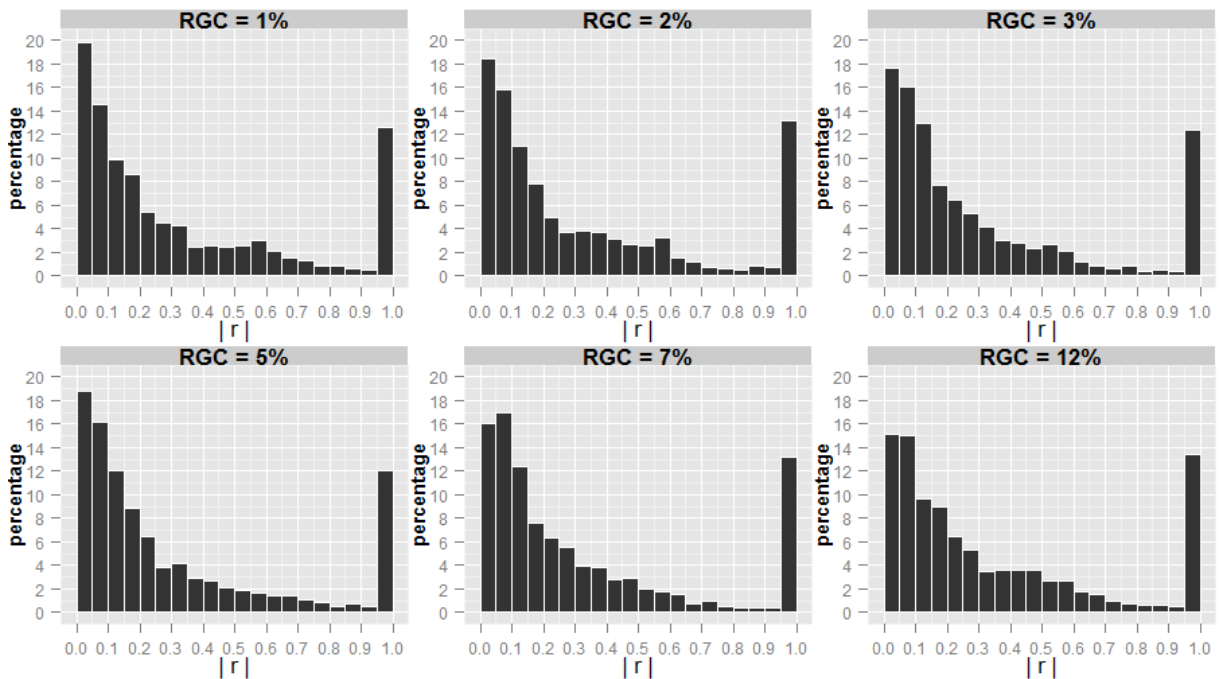
The genotypes for 176 SNPs were simulated using the Hapgen2 software and a reference Hapmap panel (Hapmap 3 release 2) for a Caucasian population. Hapgen2 resamples haplotypes from the reference panel to simulate new haplotypes as an imperfect mosaic of reference haplotypes using a Hidden Markov Model, mimicking the effect of recombination<sup>1</sup>.



**Figure S1.** Distribution of the mean of Minor Allele frequencies (MAF) for the 176 SNPs simulated across the 200 dataset.



**Figure S2.** Distribution of Minor Allele frequencies (MAF) for the 6 causal SNPs simulated across the 200 dataset as a function of their respective  $R_{GC}$ .

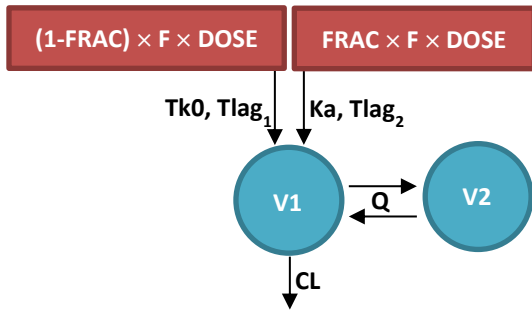


**Figure S3.** Distribution of correlations (absolute value of  $r$ ) between the causal SNP and other SNPs in linkage disequilibrium as a function of the  $R_{GC}$  associated with the causal SNP across the 200 simulated dataset.

Simulated polymorphisms have the same frequencies (**Figure S1- S2**) and correlations patterns than the reference HapMap panel (**Figure S3**). On average, 21% of the correlations between the causal SNP and SNPs in linkage disequilibrium have correlation coefficient  $|r|$  higher than 0.89, regardless to the  $R_{GC}$  associated. This can be attributed to the DNA chip design, where groups of SNPs from the same gene, so physically close, are present.

### Pharmacokinetic model used for simulations of concentrations profiles

PK profiles were simulated using the model and parameters estimates best describing the nonlinear PK of the motivating data example. The structural model includes two-compartment for disposition, a double absorption function nonlinear with dose and a linear elimination (**Figure S4**).



**Figure S4.** Structural pharmacokinetic model of drug S. Double absorption compartments in red and distribution compartments in blue. F: bioavailability ; FRAC: fraction of dose ; Tk0: zero order absorption constant rate ; Tlag<sub>1</sub>: lag time of zero order absorption ; Ka: first order absorption constant rate ; Tlag<sub>2</sub>: lag time of first order absorption ; V1: central compartment volume ; V2: peripheral compartment volume ; Q: intercompartmental clearance ; CL: linear elimination clearance

The simulated data sets were fitted using NLMEM, as in the original analysis. The drug concentration at time  $t_{ij}$  in the  $i^{th}$  subject ( $i = 1, \dots, N, j = 1, \dots, n$ ),  $y_{ij}$ , was assumed to follow a nonlinear function  $f$ , such as:

$$y_{ij} = f(\theta_i ; \xi_{ij}) + \varepsilon_{ij} \quad (1)$$

where  $\theta_i$  is the vector of individual parameters and  $\varepsilon_{ij}$  the residual error. We assume that  $\theta_i$  follow a log-normal distribution,  $\theta_i = \mu e^{\eta_i}$ , with  $\mu$  the average population parameter and  $\eta_i$  the random effect in individual  $i$ .  $\xi_{ij}$  is the elementary design for the subject  $i$ , i.e. a vector of  $n$  sampling times ( $\xi_{ij} = t_{i,1}, \dots, t_{i,n}$ ). We typically assume  $\eta_i \sim N(0, \omega^2)$  and  $\varepsilon_{ij} \sim N(0, \sigma^2)$ .

The SAEM estimation algorithm<sup>2</sup> implemented in the Monolix software ([www.lixoft.eu](http://www.lixoft.eu)) was used to estimate the population parameters ( $\mu$ ,  $\omega^2$  and  $\sigma^2$ ) by maximum likelihood. *Maximum a Posteriori* method (MAP) estimates of the individual parameters were obtain by

a Bayesian approach<sup>3</sup>. The individual parameters are also called Empirical Bayes Estimates (EBE).

### Computation of the correlation coefficient in the stepwise procedure

The Pearson's correlation coefficient  $r$  is a measure of the linkage disequilibrium between two markers. For example for two SNPs with respective alleles A/a and B/b, the correlation coefficient is computed as follows:

$$r = \frac{p_{AB} - p_A \cdot p_B}{\sqrt{p_A(1 - p_A) \cdot p_B(1 - p_B)}} \quad (2)$$

where  $p_{AB}$  is the frequency of the haplotype AB and  $p_A$  and  $p_B$  the frequencies of the alleles A and B respectively<sup>4</sup>. The coefficient  $r$  can take values between -1 and 1. If this value is close to 0, the SNPs are in equilibrium. On the contrary if  $|r|$  tend to 1, the markers are in strong disequilibrium. The Pearson's correlation coefficient was computed using the PLINK software<sup>5</sup>, which also compute the square of the coefficient, *i.e.* the coefficient of determination, to remove the arbitrary sign introduced.

### Evaluation

For each analysis scenario, 200 data sets were simulated under each hypothesis  $H_0$  and  $H_1$ . We first evaluated the ability of the designs to estimate the population and individual parameters under  $H_0$  through diagnostic plots: the population estimates of CL ( $\hat{\mu}_{CL_{est}}$ ) were compared to the population CL value ( $\mu_{CL_{sim}}$ ) used in simulations through the relative estimation errors ( $REE_{pop}$ ) estimates obtained on each of the 200 data sets:

$$REE_{pop}(\%) = \frac{\hat{\mu}_{CL_{est}} - \mu_{CL_{sim}}}{\mu_{CL_{sim}}} \times 100 \quad (3)$$

The individual CL simulated ( $CL_{sim_i}$ ) using parameters distributions from the motivating example to compute PK profiles were also compared with CL estimated by NLMEM on the simulated data ( $EBE_{CL_i}$ ). In each data set, we computed the  $REE_{indiv}$  between estimated and simulated individual clearances in the same way, as:

$$REE_{indiv}(\%) = \frac{EBE_{CL_i} - CL_{sim_i}}{CL_{sim_i}} \times 100 \quad (4)$$

To evaluate the shrinkage on random effects ( $Sh_\eta$ ), Bertrand et al<sup>4</sup> used a metric based on estimated variances:

$$Sh_{\eta}^{var} = 1 - \frac{var(\hat{\eta}_i)}{\hat{\omega}^2} \quad (5)$$

where  $\hat{\eta}_i$  are the EBE and  $\hat{\omega}^2$  the estimate of variance of random effects from the population step. Under  $H_0$ , we computed this metric which reflects the quantity of information brought by each subject to the estimation of individual parameters. When shrinkage is high ( $Sh_{\eta}^{var} \geq 50\%$ ), the individual estimates are shrunk toward  $\hat{\mu}_{CL_{est}}$ , and covariate tests based on individual parameters may be misleading for instance a covariate relationship may be masked<sup>5</sup>.

The number of true positives ( $TP$ ) was the count of causal variants associated to CL (maximum of 1200 over the 200 simulations) and the number of false positives ( $FP$ ) the count of non-causal variants associated to CL and any variants associated to Q and V2. A 95% confidence interval was estimated assuming the number of TP or FP follows a Poisson distribution. The true positive rate ( $TPR$ ) and the false positive rate ( $FPR$ ) were calculated as follow:

$$TPR = \frac{TP}{TP + FN} \quad (6.1)$$

$$FPR = \frac{FP}{FP + TN} \quad (6.2)$$

where  $FN$  and  $TN$  are respectively the count of false and true negatives. The probability to detect a given number  $x$  ( $x = 1, \dots, 6$ ) of the 6 SNPs which were associated to CL by simulation was estimated by the percentage of data sets where  $x$  or more SNPs were selected under  $H_1$ .

To quantify the loss of genetic signal between simulated and estimated individual clearances, we performed linear univariate regressions on  $CL_{sim_i}$  or  $EBE_{CL_i}$  as a function of the different causal SNPs:

$$\log(CL_{sim_i}) = a + b_{CL_{sim_i}} \cdot SNP_{ik} \quad (7)$$

$$\log(EBE_{CL_i}) = a + b_{EBE_{CL_i}} \cdot SNP_{ik}$$

where  $a$  represents an intercept,  $b_{CL_{sim_i}}$  the regression slope on simulated clearance ( $CL_{sim_i}$ ),  $b_{EBE_{CL_i}}$  the regression slope on estimated clearance ( $EBE_{CL_i}$ ), and  $SNP_{ik}$  the

genotype of the  $k^{\text{th}}$  causal variant ( $k = 1, \dots, 6$ ) for subject  $i$ ;  $SNP_{ik} = \{0,1,2\}$ . The slopes associated to the 6 causal variants were computed separately for each simulated dataset, regardless of their significance with the different association methods. They represent the increase in clearance due to mutated alleles (genetic signal), and were compared to quantify the departure of the estimated genetic signal from the simulated value (relative deviation,  $RD_{signal}$ ):

$$RD_{signal}(\%) = \frac{b_{EBE_{CL_i}} - b_{CL_{sim_i}}}{b_{CL_{sim_i}}} \times 100 \quad (8)$$

## Softwares

The C++ software Hapgen version 2.1.2<sup>6</sup> was used to simulate SNPs from reference haplotype panels provided by the Hapmap Project. Among the 176 SNPs present on the DNA microarray, 55 were also present in the Hapmap reference panel. For the others, we chose the closest variant in the Hapmap database. The SNPs selected from the Hapmap database to target the SNPs from the IRIS DNA microarray were very close to the original ones with a median of 1730 bases pair departure. ([mathgen.stats.ox.ac.uk/genetics\\_software/hapgen/hapgen2.html](http://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html))

The software Monolix version 4.2.2 ([www.lixoft.eu](http://www.lixoft.eu)) implements the SAEM algorithm for parameters estimation in nonlinear mixed effects models. It was used here to derive the EBE using the PK model and the simulated PK profiles.

The C++ software PLINK version 1.07<sup>7</sup> ([pngu.mgh.harvard.edu/~purcell/plink/](http://pngu.mgh.harvard.edu/~purcell/plink/)), is a free and open-source whole genome association analysis toolset, which was used to performed the stepwise procedure.

The statistical software R version 2.15.2<sup>8</sup> ([www.R-project.org/](http://www.R-project.org/)) was used to simulate the PK profiles, as well as for all the statistical analyses and creation of graphical material. Some R packages were also used such as “ridge”<sup>9</sup>, a package for ridge regression with automatic selection of the regularization parameter.

The C++ program HyperLasso<sup>10</sup> ([www.ebi.ac.uk/projects/BARGEN](http://www.ebi.ac.uk/projects/BARGEN)) was used for the Lasso as well as the HyperLasso regression methods. Of note, the phenotype is standardized within this program so  $\sigma$  is equal to 1 in the formula to set  $\gamma$  for HyperLasso.

## Additional results for scenarios presented in the main manuscript

### Simulated genetic effects

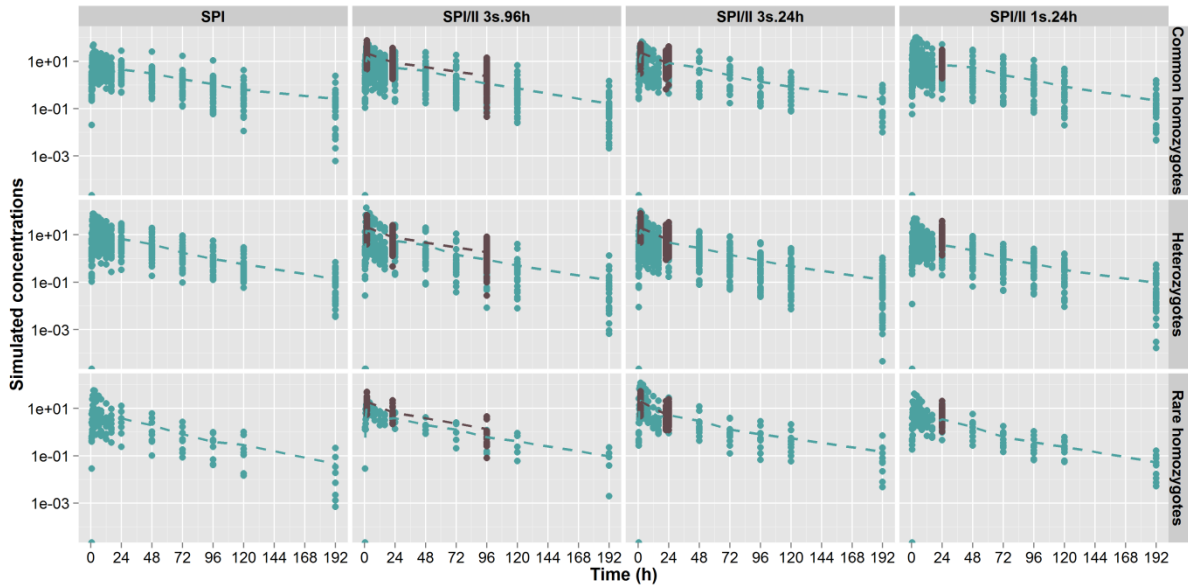
**Table S1.** The 10<sup>th</sup>, 50<sup>th</sup> and 90<sup>th</sup> percentiles of simulated effect sizes and simulated individual clearances as a function of the genotype of the causal SNP explaining 1 or 12% of clearance variability, for scenarios SPI and SPI/II<sub>3s.96h</sub>.

Scenario	RGC	Percentile	$\beta_k$	Genotype-based typical CL <sup>a</sup>		
				SNP = 0	SNP = 1	SNP = 2
SPI	1%	0.1	0.03	94.9	97.8	100.8
		0.5	0.04	94.9	98.4	102.1
		0.9	0.1	94.9	104.8	115.8
	12%	0.1	0.11	94.9	106	118.3
		0.5	0.15	94.9	109.9	127.4
		0.9	0.35	94.9	135.3	193
SPI/II <sub>3s.96h</sub>	1%	0.1	0.03	94.9	97.8	100.8
		0.5	0.04	94.9	98.6	102.4
		0.9	0.11	94.9	106.1	118.6
	12%	0.1	0.11	94.9	106	118.3
		0.5	0.14	94.9	108.8	124.7
		0.9	0.34	94.9	133.1	186.6

a. the typical clearance was computed as  $CL = \mu_{CL} \cdot e^{\beta_k \cdot SNP_k}$ , where  $\mu_{CL}$  is the typical value of CL in the overall population (94.9),  $\beta_k$  the effect size associated to the causal  $SNP_k$  ( $SNP_k = \{0,1,2\}$ ).

**Figure S5** represents the concentration profiles from one simulated data set under the alternative hypothesis  $H_1$  for all scenarios, sorted by genotypes of the SNP explaining 12% of the clearance interindividual variability. As expected because of the moderate variability for CL (25%), the trough concentrations were slightly lower in heterozygote subjects, and decreased a little further in rare homozygotes, for phase I as well as phase II data in all scenarios. The impact of genetics on PK was difficult to observe based on individual profiles only, justifying the use of sophisticated approaches as NLMEM.



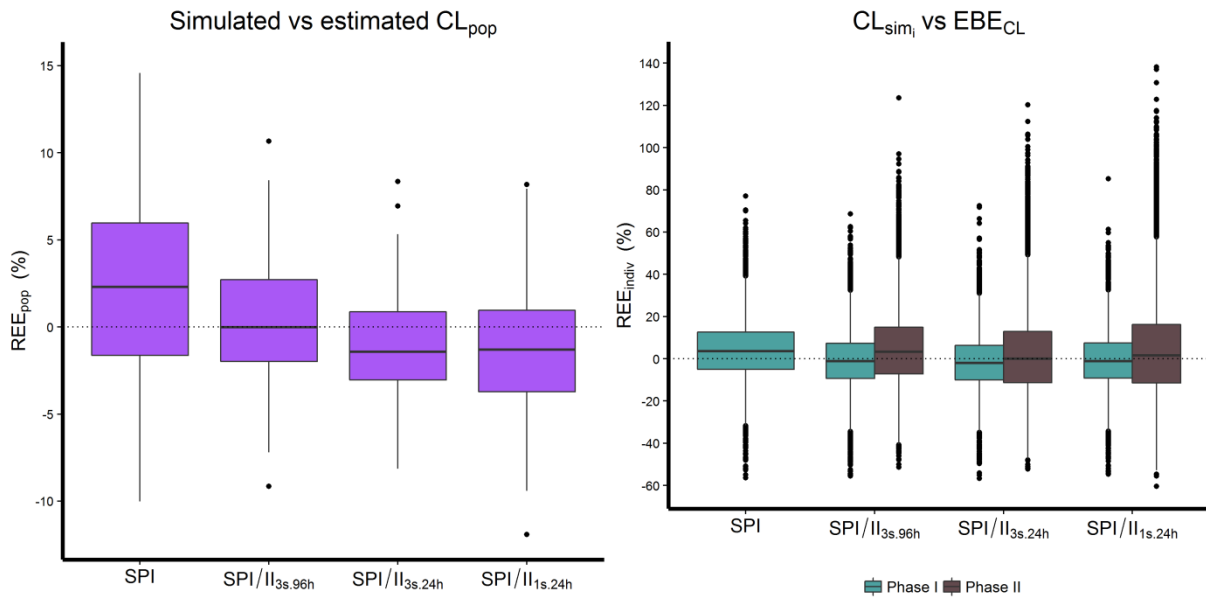


**Figure S5.** Phase I (blue dots) and phase II study (brown dots) concentrations versus time for an example data set for each main scenario simulated with  $IIV_{CL} = 25\%$  under  $H_1$ . Mean profiles (dashed lines) and concentrations are plotted in log-scale on the Y-axis. Concentrations are sorted for the SNP explaining 12% of CL interindividual variability; common homozygotes (*top*), heterozygotes (*middle*) and rare homozygotes (*bottom*).

### Estimation bias

On average, the bias on the population estimate of CL was less than 5% for all scenarios (**Figure S6, left**), showing that this parameter is well estimated in all our settings. Combining phase II data with the phase I data brought a small increase in precision, as shown from the smaller range of REE. This was true even in scenario  $SPI/II_{1s,24h}$ , where the information on CL from the time point at 24h is very limited, but the best compromise was obtained in the scenario with a late sample at 96h ( $SPI/II_{3s,96h}$ ), where there was no bias on the estimate and the precision was less than 10% for most simulated datasets. Overall, adding phase II data had no major impact on the estimation of population CL.

In contrast to population estimates,  $REE_{indiv}$  were much more widely spread out. However, medians for both phase I and phase II data were close to 0 (**Figure S6, right**). The range of  $REE_{indiv}$  was constant across analysis scenarios. For phase I data it ranged from -55% to 75%, and from -60% to 130% for phase II data, due to the differences in number of sampling times per subject. The bias in mean  $REE_{indiv}$  was smaller in phase II data with three samples ( $SPI/II_{3s,24h}$  and  $SPI/II_{3s,96h}$ ) than in phase II data with one sample ( $SPI/II_{1s,24h}$ ).



**Figure S6.** Boxplots of the simulated and estimated population clearance in all scenarios simulated under  $H_0$  ( $REE_{pop}$ , left). Boxplots of the simulated and estimated individual clearance in all scenarios simulated under  $H_0$  ( $REE_{indiv}$ , right) for subjects in the phase I (blue) and the phase II study (brown).

### Detection of genetic effects

**Table S2.** True positives counts under  $H_1$  for Lasso and stepwise procedure methods in all scenarios

Method	TP (CL)			
	SPI	SPI/II <sub>3s.96h</sub>	SPI/II <sub>3s.24h</sub>	SPI/II <sub>1s.24h</sub>
Lasso	44 [32;59]	380 [343;420]	326 [292;363]	277 [245;312]
Stepwise procedure	43 [31;58]	386 [348;426]	327 [293;364]	262 [231;296]

Total number of true positives (TP) with their 95% confidence interval under the alternative hypothesis.

On 200 simulated data sets, overall 1200 SNPs were set to impact clearance (maximum TP number).

**Table S3.** False positives counts on CL, Q and V2 under  $H_1$  for Lasso and stepwise procedure methods in all scenarios

Method	$S_{\text{phase1}}$			SPI/II <sub>3s.96h</sub>			SPI/II <sub>3s.24h</sub>			SPI/II <sub>1s.24h</sub>		
	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)
Lasso	21 [13;32]	35 [24;49]	8 [3;16]	131 [110;155]	40 [29;54]	73 [57;92]	122 [101;146]	112 [92;135]	31 [21;44]	120 [99;43]	68 [53;86]	31 [21;44]
Stepwise procedure	18 [11;28]	35 [24;49]	9 [4;17]	91 [73;112]	47 [35;63]	61 [47;78]	84 [67;104]	107 [88;129]	53 [40;69]	84 [67;104]	74 [58;93]	50 [37;66]

Total number of false positives (FP) with their 95% confidence interval under the alternative hypothesis.

**Table S4.** True positive rates for each causal variant ( $TPR_{R_{GC}}$ ) as a function of the  $R_{GC}$ , and false positive rates (FPR) under  $H_1$  for Lasso and stepwise procedure methods in all scenarios.

Scenario	Method	FPR	$TPR_{R_{GC}}$					
			$R_{GC} = 1\%$	$R_{GC} = 2\%$	$R_{GC} = 3\%$	$R_{GC} = 5\%$	$R_{GC} = 7\%$	$R_{GC} = 12\%$
SPI	Lasso	0.2	1.5	0.5	1.5	5	4	9.5
SPI	Stepwise procedure	0.2	1.5	0.5	1.5	5	4	9
SPI/PII <sub>3s.96h</sub>	Lasso	0.7	2.5	4	13.5	31	52	87
SPI/PII <sub>3s.96h</sub>	Stepwise procedure	0.6	3.5	6	13	36.5	51.5	82.5
SPI/PII <sub>3s.24h</sub>	Lasso	0.8	3	3.5	8.5	25	41	82
SPI/PII <sub>3s.24h</sub>	Stepwise procedure	0.7	4	4	9	26.5	42	78
SPI/PII <sub>1s.24h</sub>	Lasso	0.6	3.5	5	11	21.5	32.5	65
SPI/PII <sub>1s.24h</sub>	Stepwise procedure	0.6	2.5	4	10	21.5	31.5	61.5

FPR: proportion of false positives detected among all potential false associations in the 200 dataset.

$TPR_{R_{GC}}$ : proportion of true positives detected for a causal variant in the 200 dataset (equivalent to the probability to detect this variant).

### Influence of the phenotype variance

**Table S5.** True positives counts under  $H_1$  for Lasso and stepwise procedure methods in all scenarios simulated with  $IIV_{CL} = 60\%$ .

Method	TP (CL)			
	SPI <sub>60%</sub>	SPI/II <sub>3s.96h_60%</sub>	SPI/II <sub>3s.24h_60%</sub>	SPI/II <sub>1s.24h_60%</sub>
Lasso	80 [63;100]	509 [466;555]	505 [462;551]	450 [409;494]
Stepwise procedure	80 [63;100]	541 [496;589]	531 [487;578]	464 [423;508]

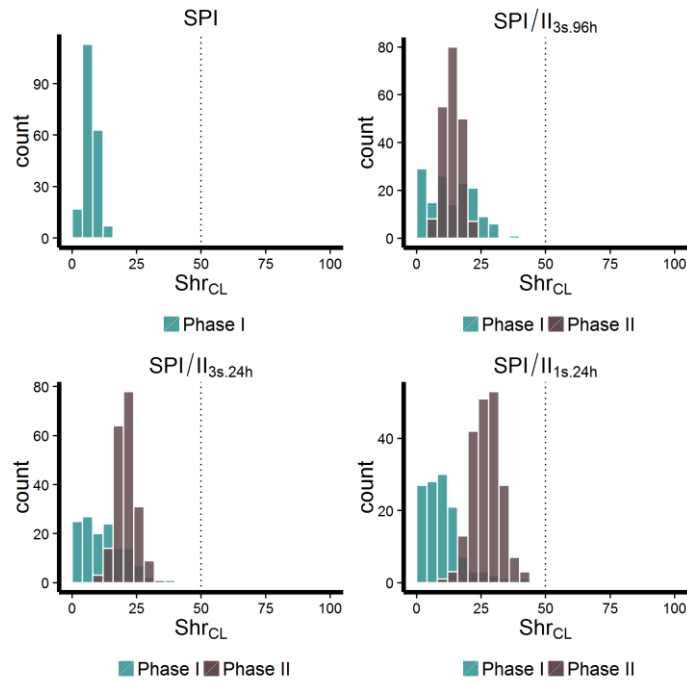
Total number of true positives (TP) with their 95% confidence interval under the alternative hypothesis.

On 200 simulated data sets, overall 1200 SNPs were set to impact clearance (maximum TP number).

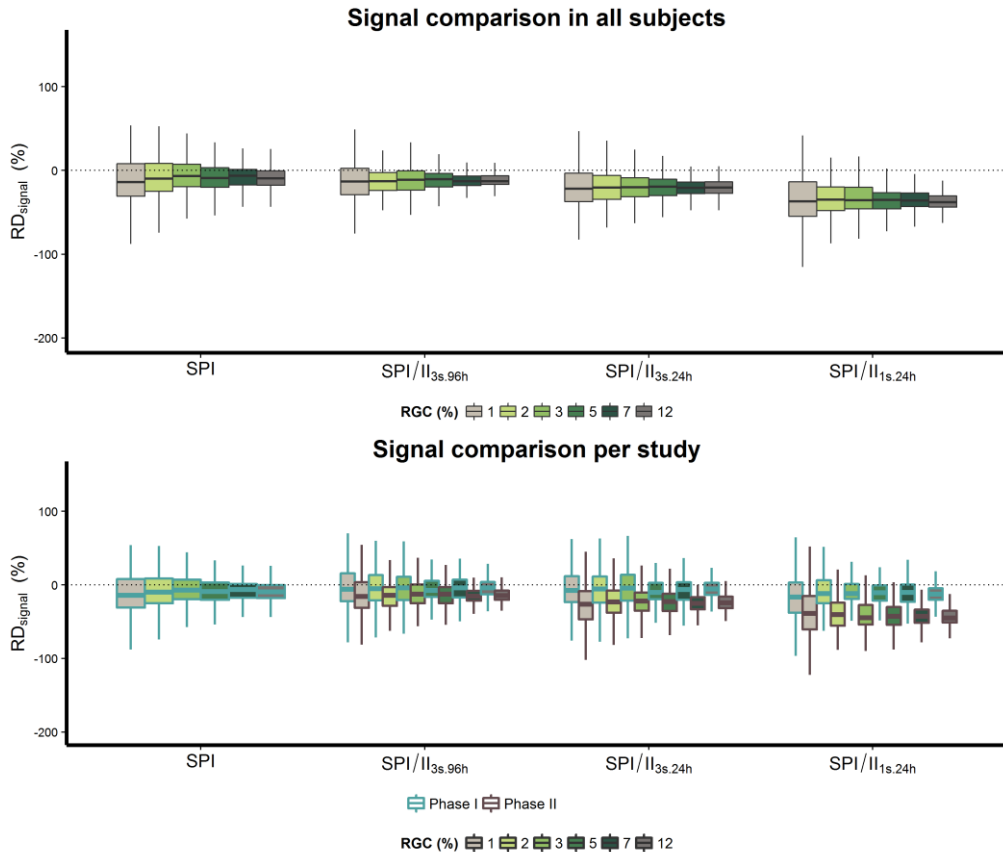
**Table S6.** False positives counts on CL, Q and V2 under  $H_1$  for Lasso and stepwise procedure methods in all scenarios simulated with  $IIV_{CL} = 60\%$ .

Method	SPI <sub>60%</sub>			SPI/II <sub>3s.96h_60%</sub>			SPI/II <sub>3s.24h_60%</sub>			SPI/II <sub>1s.24h_60%</sub>		
	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)
Lasso	34 [24;48]	43 [31;58]	15 [8;25]	171 [146;199]	28 [19;40]	29 [19;42]	161 [137;188]	117 [97;140]	39 [28;53]	150 [127;176]	81 [64;101]	24 [15;36]
Stepwise procedure	28 [19;40]	44 [32;59]	12 [6;21]	99 [80;121]	31 [21;44]	27 [18;39]	97 [79;118]	111 [91;134]	66 [51;84]	106 [87;128]	82 [65;102]	36 [25;50]

Total number of false positives (FP) with their 95% confidence interval under the alternative hypothesis.



**Figure S7.** Distribution of the shrinkages on clearance for subjects in the phase I dataset (blue) and for subjects in the phase II dataset (brown), for each scenario simulated under  $H_0$  with  $IIV_{CL} = 60\%$ .



**Figure S8.** Boxplot showing the loss of the signal for genetic effect in the overall population (*top*), as well as separately for the phase I data (blue borders) and for the phase II data (brown borders) (*bottom*). A boxplot is shown as a function of increasing  $R_{GC}$  (boxplots colour) separately for each scenario simulated under  $H_1$  with  $IIV_{CL} = 60\%$ .

## Additional results for the two other penalised regression methods

Ridge regression<sup>11</sup> imposes a penalty on the size of the  $\beta_k$  to reduce the prediction error without preventing the inclusion of variables in the model, by applying a Gaussian prior of identical variance on the eigenvalues issued from the principal component analysis (PCA) of the data. We used the approach proposed by Cule et al.<sup>9</sup> to semi-automatically set the penalty. The fit was followed by a Wald test on these coefficients and their standard error to perform the variable selection<sup>12</sup>, with a significance threshold equal to  $\alpha$ , the type I error per SNP.

HyperLasso<sup>10</sup> is similar to Lasso but uses a normal-exponential gamma (NEG) distribution as a prior on  $\beta_k$  and depends on two parameters: a shape parameter  $\lambda$  and a scale parameter  $\gamma$ . The sharp peak at zero and the flatter tail of the NEG distribution favour sparse solutions but the estimates of larger effects are shrunken less severely than the Lasso. The smaller the shape parameter the heavier the tails of the distribution and the more peaked at zero, which can result in fewer correlated SNPs being selected. The shape parameter  $\lambda$  was set to 1, which gives realistic effect size distributions<sup>13</sup>. The scale parameter  $\gamma$  was computed, as for the Lasso tuning parameter, depending on  $\alpha$ <sup>10</sup>.

**Table S7.** Empirical estimates of Family-Wise Error Rate under  $H_0$  for ridge regression and HyperLasso methods

Method		FWER (%)			
		SPI	SPI/II <sub>3s,96h</sub>	SPI/II <sub>3s,24h</sub>	SPI/II <sub>1s,24h</sub>
Ridge regression	Without correction <sup>a</sup>	16	15.5	18.5	14
HyperLasso	Without correction <sup>a</sup>	13.5	18	21.5	13.5
Ridge regression	After empirical correction <sup>b</sup>	21	20.5	18.5	19.5
HyperLasso	After empirical correction <sup>b</sup>	20.5	20	21.5	18

a. Set of empirical family wise error rates (FWER) obtain without correction.

b. Set of empirical FWER obtained after correction of thresholds or penalisation parameters. The 95% prediction interval around 20 for 200 simulated data sets is [14.5-25.5].

**Table S8.** Counts of true positives under  $H_1$  for ridge regression and HyperLasso methods in all scenarios

Method	TP (CL)			
	SPI	SPI/II <sub>3s,96h</sub>	SPI/II <sub>3s,24h</sub>	SPI/II <sub>1s,24h</sub>
Ridge regression	57 [43;74]	369 [332;409]	300 [267;336]	240 [211;272]
HyperLasso	41 [29;56]	309 [276;345]	250 [220;283]	225 [197;256]

Total number of true positives (TP) with their 95% confidence interval under the alternative hypothesis.

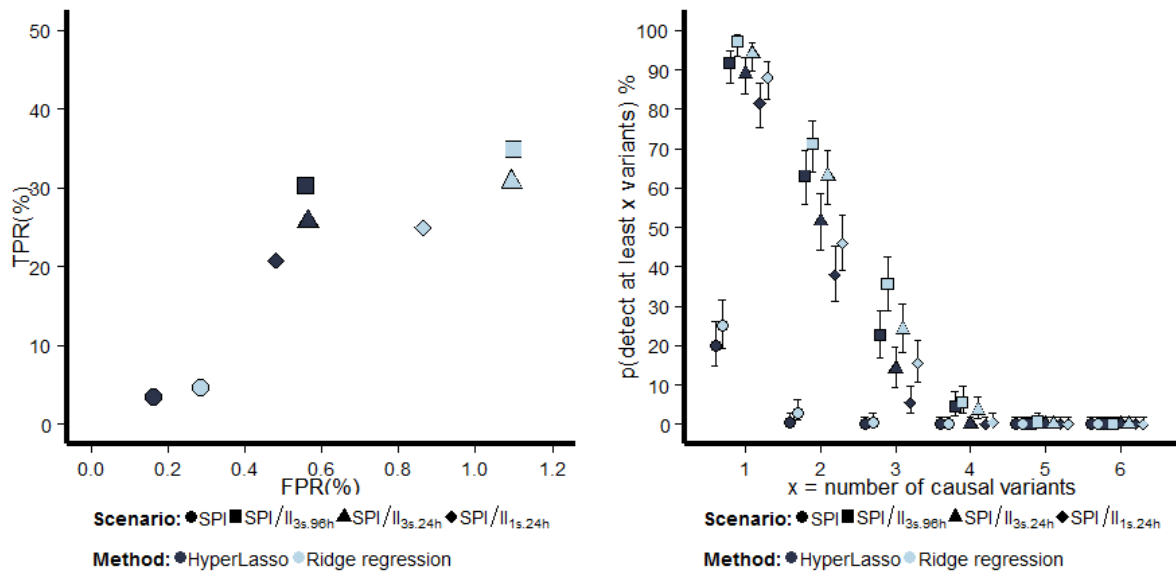
On 200 simulated data sets, overall 1200 SNPs were set to impact clearance (maximum TP number).



**Table S9.** Counts of false positives on CL, Q and V2 under  $H_1$  regression and HyperLasso methods in all scenarios

Method	SPI			SPI/II <sub>3s.96h</sub>			SPI/II <sub>3s.24h</sub>			SPI/II <sub>1s.24h</sub>		
	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)	FP(CL)	FP(Q)	FP(V2)
Ridge regression	45 [33;60]	42 [30;57]	10 [5;18]	247 [217;280]	48 [35;64]	79 [63;98]	201 [174;231]	136 [114;161]	35 [24;49]	179 [154;207]	79 [63;98]	36 [25;50]
HyperLasso	15 [8;25]	33 [23;46]	7 [3;14]	88 [71;108]	38 [27;52]	64 [49;82]	72 [56;91]	93 [75;114]	27 [18;39]	78 [62;97]	59 [45;76]	27 [18;39]

Total number of false positives (FP) with their 95% confidence interval under the alternative hypothesis.



**Figure S9.** True Positive Rate (TPR) versus False Positive Rate (FPR) under  $H_1$  (*left*) and probability estimates (points) and 95% confidence interval (bars) to detect at least  $x$  causal variants explaining the interindividual variability of CL ( $x=1,\dots,6$ ) under  $H_1$  (*right*). Different symbols are used for each scenario, and colours denote the HyperLasso (grey) and the ridge regression (light blue).

## Additional references

1. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
2. Kuhn, E. & Lavielle, M. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat.* **8**, 115–131 (2004).
3. Sheiner, L. B., Rosenberg, B. & Melmon, K. L. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput. Biomed. Res. Int. J.* **5**, 411–459 (1972).
4. Bertrand, J., Comets, E., Laffont, C. M., Chenel, M. & Mentré, F. Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm. *J. Pharmacokinet. Pharmacodyn.* **36**, 317–339 (2009).
5. Savic, R. M. & Karlsson, M. O. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *AAPS J.* **11**, 558–569 (2009).
6. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).
7. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
8. R Development Core Team *A language and environment for statistical computing. R Foundation for Statistical Computing.* (2012).at <<http://www.R-project.org/>>
9. Cule, E. & De Iorio, M. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet. Epidemiol.* **37**, 704–714 (2013).
10. Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
11. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55 (1970).
12. Cule, E., Vineis, P. & De Iorio, M. Significance testing in ridge regression for genetic data. *BMC Bioinformatics* **12**, 372 (2011).
13. Vignal, C. M., Bansal, A. T. & Balding, D. J. Using penalised logistic regression to fine map HLA variants for rheumatoid arthritis. *Ann. Hum. Genet.* **75**, 655–664 (2011).