Supplementary Material and Methods

Model "COMP"

For model "COMP", we separate the causal SNPs in three equal sets $S_{\text{causal}}^{(1)}$, $S_{\text{causal}}^{(2)}$ and $S_{\text{causal}}^{(3)}$; $S_{\text{causal}}^{(3)}$ is further separated in two equal sets, $S_{\text{causal}}^{(3.1)}$ and $S_{\text{causal}}^{(3.2)}$. We then compute

$$y_i = \underbrace{\sum_{j \in S_{\text{causal}}^{(1)}} w_j \cdot \widetilde{G_{i,j}}}_{\text{linear}} + \underbrace{\sum_{j \in S_{\text{causal}}^{(2)}} w_j \cdot \widetilde{D_{i,j}}}_{\text{dominant}} + \underbrace{\sum_{\substack{k=1 \\ j_1 = e_k^{(3.1)} \\ j_2 = e_k^{(3.2)}}}^{k=\left|S_{\text{causal}}^{(3.1)}\right|} w_{j_1} \cdot \widetilde{G_{i,j_1} G_{i,j_2}}}_{\text{interaction}} \; + \; \epsilon_i \; ,$$

where w_j are weights generated from a Gaussian or a Laplace distribution, $G_{i,j}$ is the allele count of individual i for SNP j, $\widetilde{G_{i,j}}$ corresponds to its standardized version (zero mean and unit variance for all SNPs), $D_{i,j} = \mathbb{1}\left\{G_{i,j} \neq 0\right\}$, ϵ follows a Gaussian distribution $N(0, 1-h^2)$ and $S_{\text{causal}}^{(q)} = \left\{e_k^{(q)},\ k \in \left\{1, \ldots, \left|S_{\text{causal}}^{(q)}\right|\right\}\right\}$.

Maximum AUCs

We use three different ways to estimate the maximum achievable AUC for simulations (see supplementary notebook "oracle").

First, we use the estimation from equation (3) of Wray *et al.* (2010). For a prevalence fixed at 30% and an heritability of 50% (respectively 80%), the approximated theoretical values of AUC are 84.1% (respectively 93.0%). Note that this approximation is reported to be less accurate for high heritabilities.

Secondly, if we assume that the genetic part of the liabilities follows a Gaussian distribution $N(0, h^2)$ and that the environmental part follows a Gaussian distribution $N(0, 1-h^2)$, we can estimate the theoretical value of the AUC that can be achieved given the disease heritability h^2 (and prevalence K) through Monte Carlo simulations. We report AUCs of 84.1% and 94.1% for heritabilities of 50% and 80%, respectively.

Thirdly, we reproduce the exact same procedure of simulations and, for each combination of parameters (Table 2), we estimate the AUC of the "oracle", i.e. the true simulated genetic part of the liabilities, through 100 replicates. For every combination of parameters, AUC values of oracles vary between 83.2% and 84.2% for an heritability of 50% and between 93.2% and 94.1% for an heritability of 80%.

Given all these estimates of maximal achievable AUC and for the sake of simplicity, we report maximum AUCs of 84% (94%) for heritabilities of 50% (80%) whatever are the simulation parameters.

References

Sachs, M. C. *et al.* (2017). plotroc: A tool for plotting roc curves. *Journal of Statistical Software*, **79**(c02).

Wray, N. R., Yang, J., Goddard, M. E., and Visscher, P. M. (2010). The genetic interpretation of area under the roc curve in genomic profiling. *PLoS genetics*, **6**(2), e1000864.

Zeng, Y. and Breheny, P. (2017). The biglasso package: A memory-and computation-efficient solver for lasso model fitting with big data in R. *arXiv preprint arXiv:1701.05936*.

Population	UK	Finland	Netherlands	Italy	Total
Cases	2569	637	795	495	4496
Controls	7492	1799	828	540	10659
Total	10061	2436	1623	1035	15155

Table S1: Number of individuals by population and disease status in the celiac disease case-control study (after quality control, genotyped on 281,122 SNPs).

1.00e+00	7.22e-01	5.87e-01	4.20e-01	2.43e-01	1.00e-01	2.35e-02	2.21e-03	4.69e-05	8.81e-08	3.18e-12	1.83e-19	2.89e-31	1.70e-50	7.71e-82
5.00e-08	7.05e-01	5.65e-01	3.95e-01	2.20e-01	8.47e-02	1.79e-02	1.42e-03	2.28e-05	2.73e-08	4.69e-13	8.08e-21	1.80e-33	4.30e-54	1.06e-87
7.94e-01	6.87e-01	5.42e-01	3.69e-01	1.97e-01	7.08e-02	1.34e-02	8.83e-04	1.05e-05	7.74e-09	6.03e-14	2.86e-22	7.73e-36	5.97e-58	5.49e-94
7.81e-01	6.69e-01	5.19e-01	3.43e-01	1.75e-01	5.85e-02	9.79e-03	5.31e-04	4.61e-06	2.01e-09	6.69e-15	7.92e-24	2.24e-38	4.37e-62	1.00e-100
7.67e-01	6.50e-01	4.95e-01	3.18e-01	1.54e-01	4.76e-02	7.01e-03	3.08e-04	1.90e-06	4.72e-10	6.32e-16	1.70e-25	4.26e-41	1.61e-66	
7.53e-01	6.30e-01	4.70e-01	2.93e-01	1.35e-01	3.82e-02	4.90e-03	1.72e-04	7.31e-07	1.00e-10	5.04e-17	2.75e-27	5.16e-44	2.83e-71	
7.38e-01	6.09e-01	4.46e-01	2.68e-01	1.17e-01	3.02e-02	3.33e-03	9.18e-05	2.63e-07	1.89e-11	3.35e-18	3.31e-29	3.84e-47	2.26e-76	

Table S2: The 102 thresholds used for the C+T method for this study.

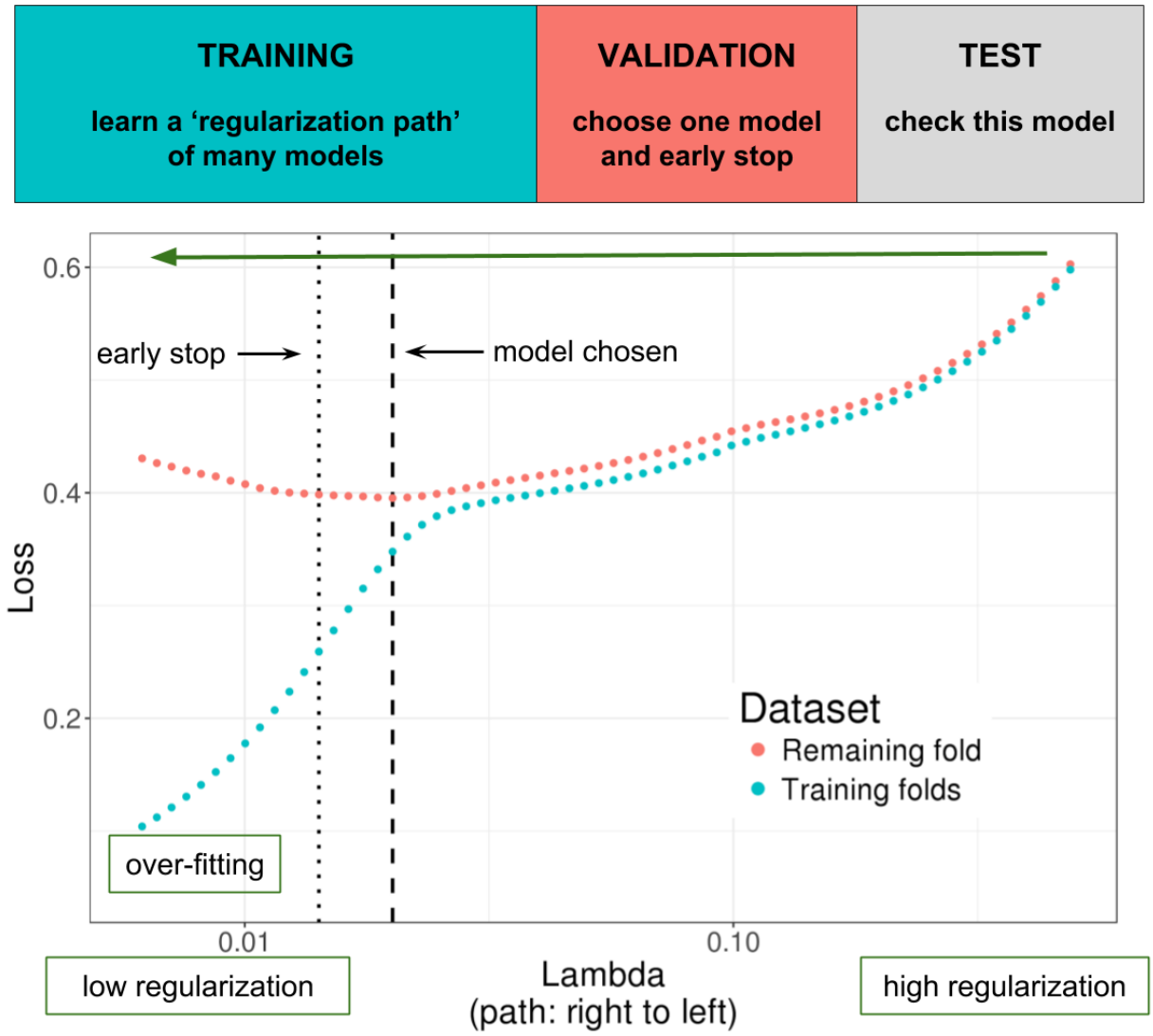

Figure S1: Illustration of one turn of the Cross-Model Selection and Averaging (CMSA) procedure. First, this procedure separates the training set in K folds (e.g. 10 folds). Secondly, in turn, each fold is considered as an inner validation set (red) and the other $(K-1)$ folds form an inner training set (blue). A "regularization path" of models is trained on the inner training set and the corresponding predictions (scores) for the inner validation set are computed. The model that minimizes the loss on the inner validation set is selected. Finally, the K resulting models are averaged. We also use this procedure to derive an early stopping criterion so that the algorithm does not need to evaluate the whole regularization paths, making this procedure much faster.

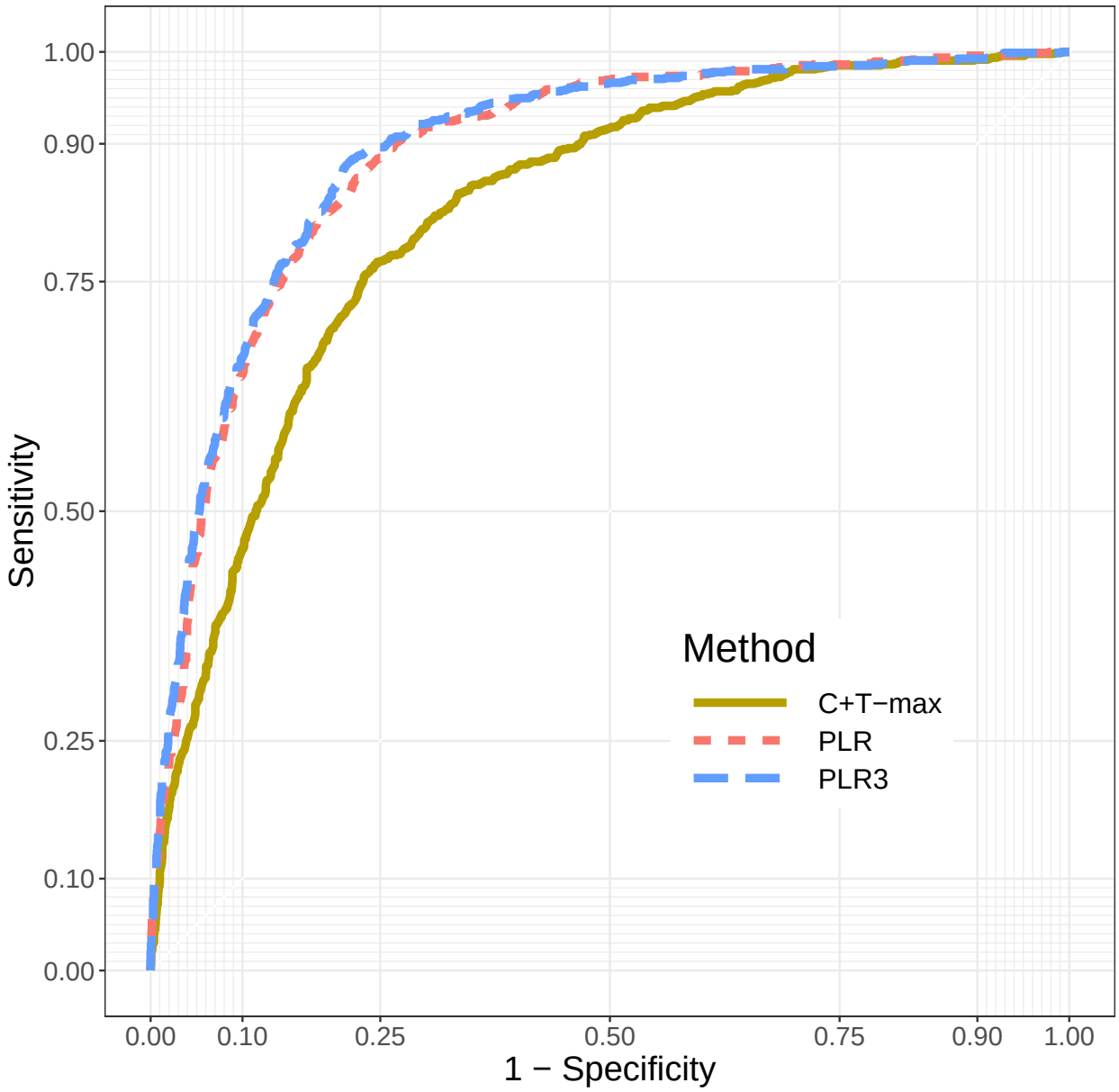

Figure S2: ROC Curves for C+T-max, PLR and PLR3 for the celiac disease dataset. Models were trained using 12,000 individuals. These are results projecting these models on the remaining 3155 individuals. The figure is plotted using R package plotROC (Sachs *et al.* 2017).

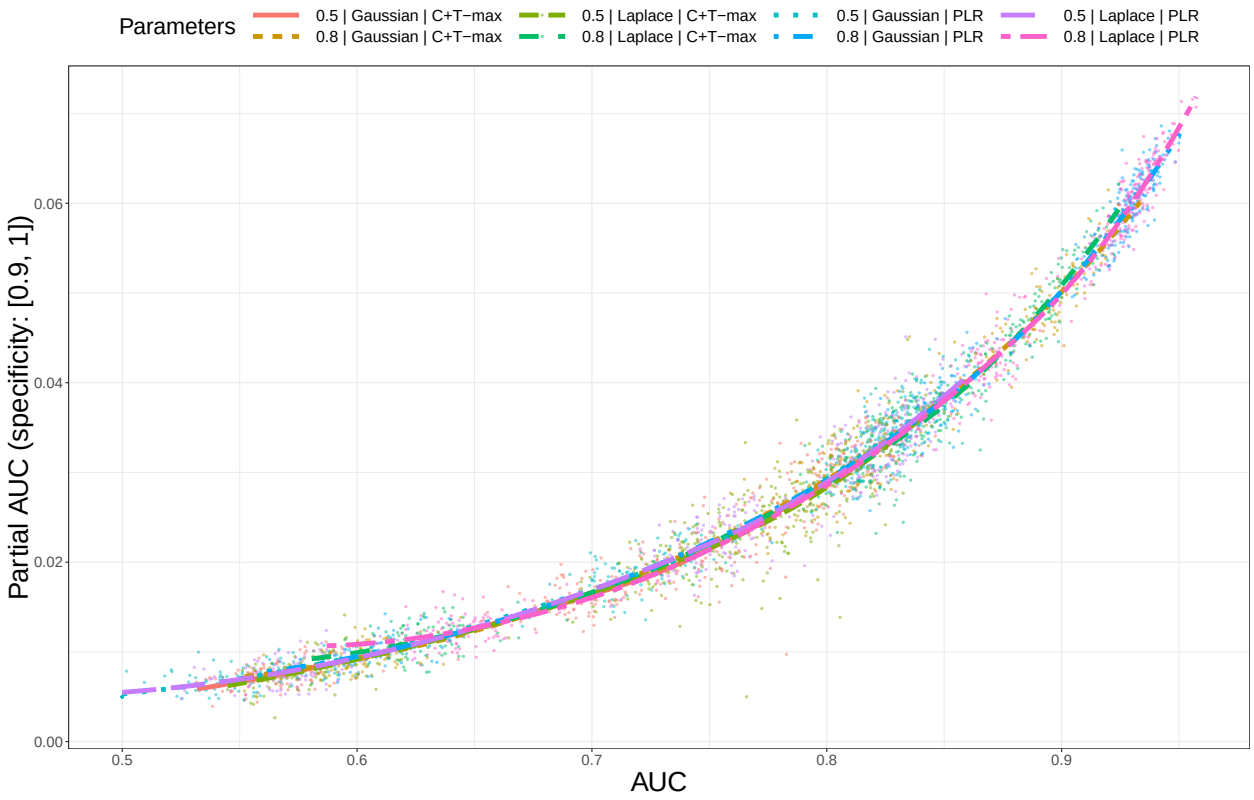

Figure S3: Correlation between AUC and partial AUC values in scenario №1. There is a Spearman correlation of 98% between values of AUC and partial AUC. The relation between the two values are the same whatever are the disease heritability, distribution of effects and method used.

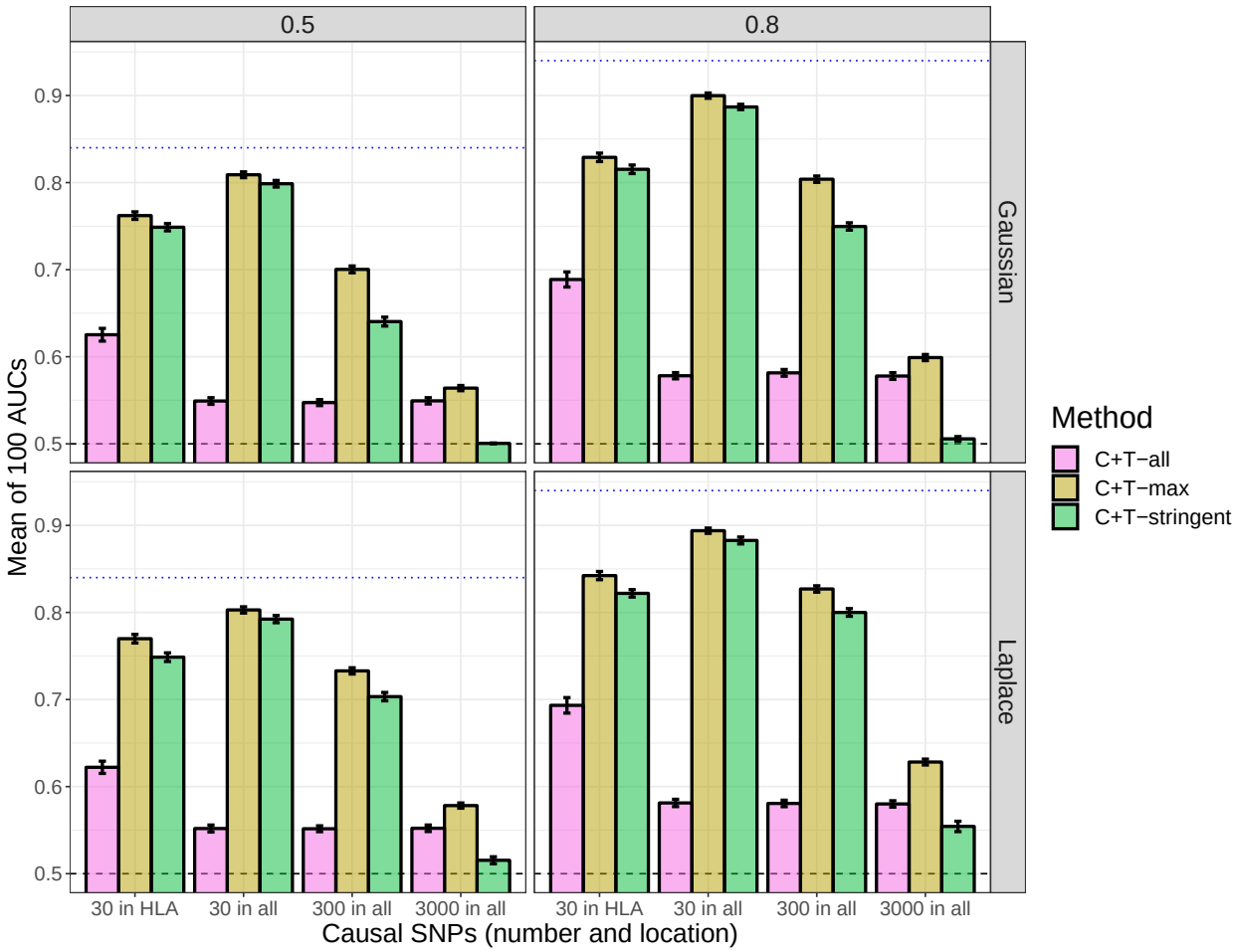

Figure S4: Comparison of three different p-value thresholds used in the C+T method in scenario №1 for model "ADD". Mean AUC over 100 simulations. Upper (lower) panels are presenting results for effets following a Gaussian (Laplace) distribution and left (right) panels are presenting results for an heritability of 0.5 (0.8). Error bars are representing ± 2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC.

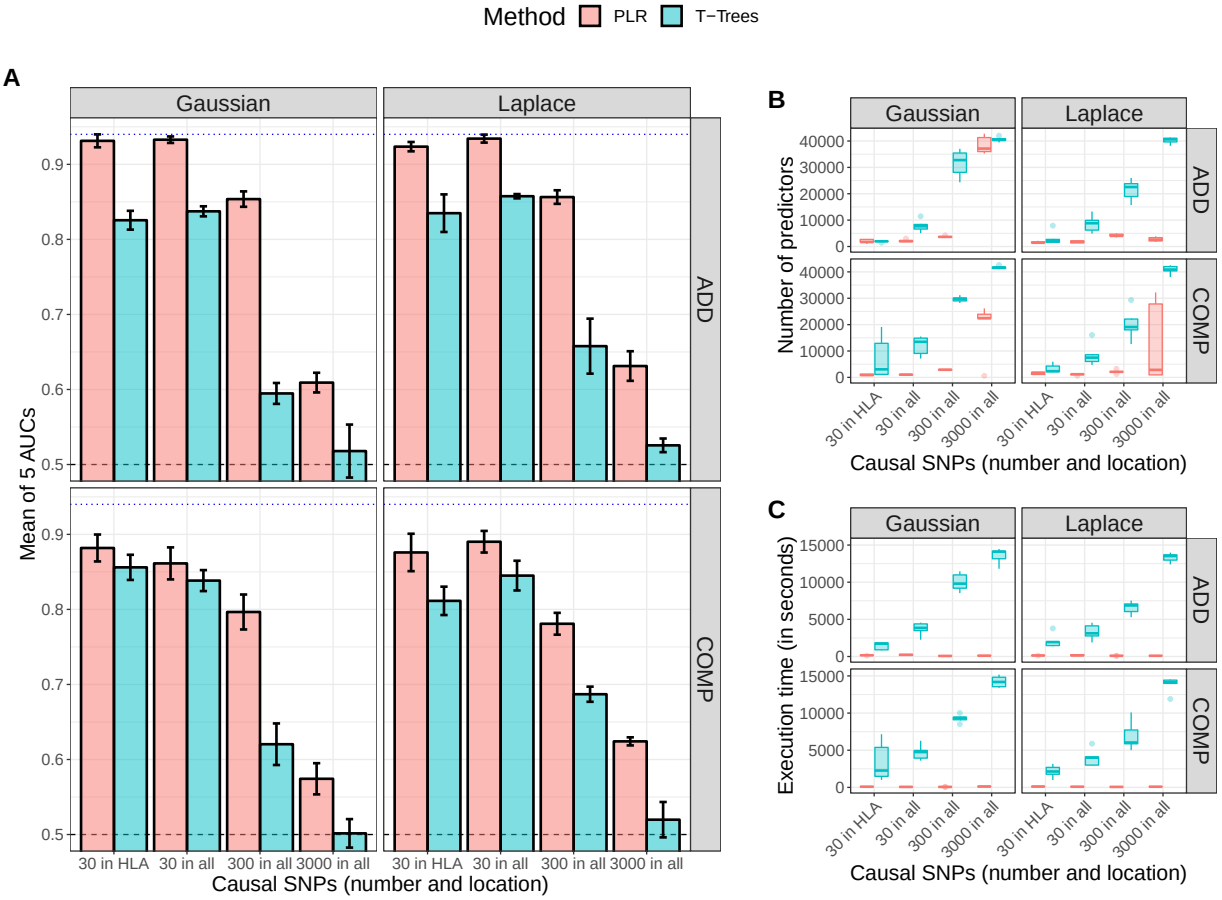

Figure S5: Comparison of T-Trees and PLR in scenario №1 for an heritability of 80%. Vertical panels are presenting results for effects following a Gaussian or Laplace distribution. Horizontal panels are presenting results for models "ADD" and "COMP" for simulating phenotypes. **A:** Mean AUC over 5 simulations. Error bars are representing ± 2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC. **B:** Boxplots of numbers of predictors used by the methods for 5 simulations. **C:** Boxplots of execution times for 5 simulations.

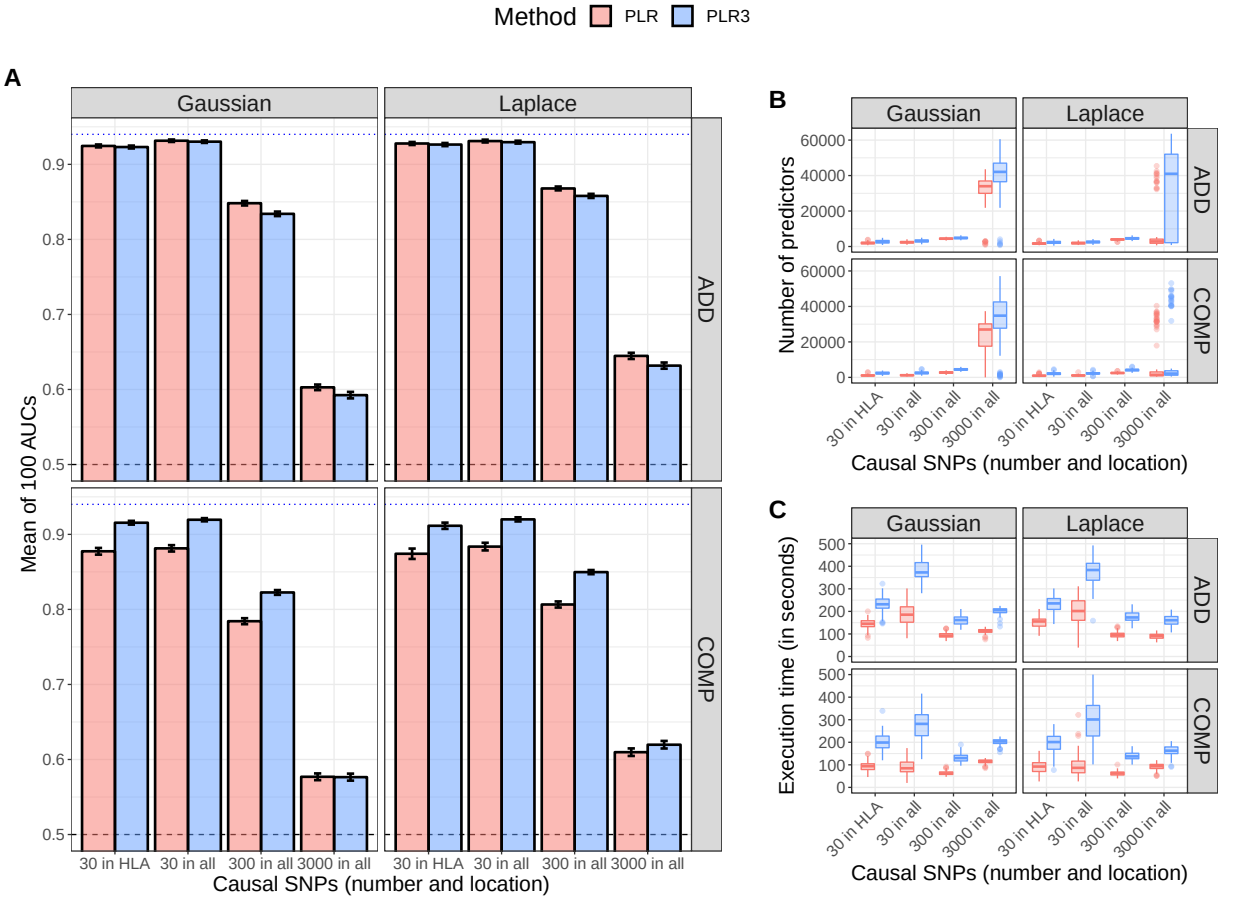

Figure S6: Comparison of PLR3 and PLR in scenario №1 for an heritability of 80%. Vertical panels are presenting results for effects following a Gaussian or Laplace distribution. Horizontal panels are presenting results for models "ADD" and "COMP" for simulating phenotypes. **A:** Mean AUC over 100 simulations. Error bars are representing ± 2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC. **B:** Boxplots of numbers of predictors used by the methods for 100 simulations. **C:** Boxplots of execution times for 100 simulations.

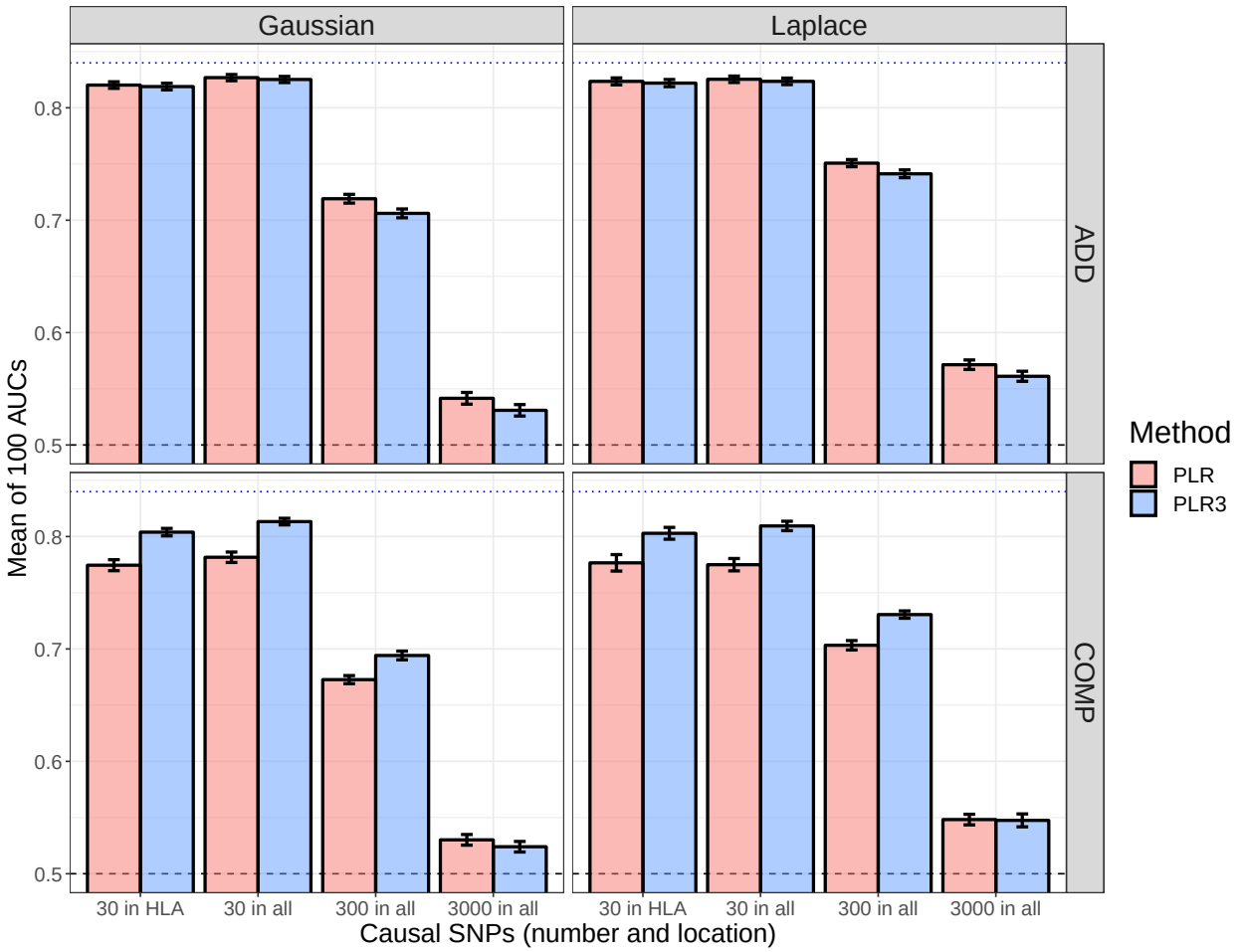

Figure S7: Comparison of PLR3 and PLR in scenario №1 for an heritability of 50%. Vertical panels are presenting results for effects following a Gaussian or Laplace distribution. Horizontal panels are presenting results for models "ADD" and "COMP" for simulating phenotypes. **A:** Mean AUC over 100 simulations. Error bars are representing $\pm$2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC.

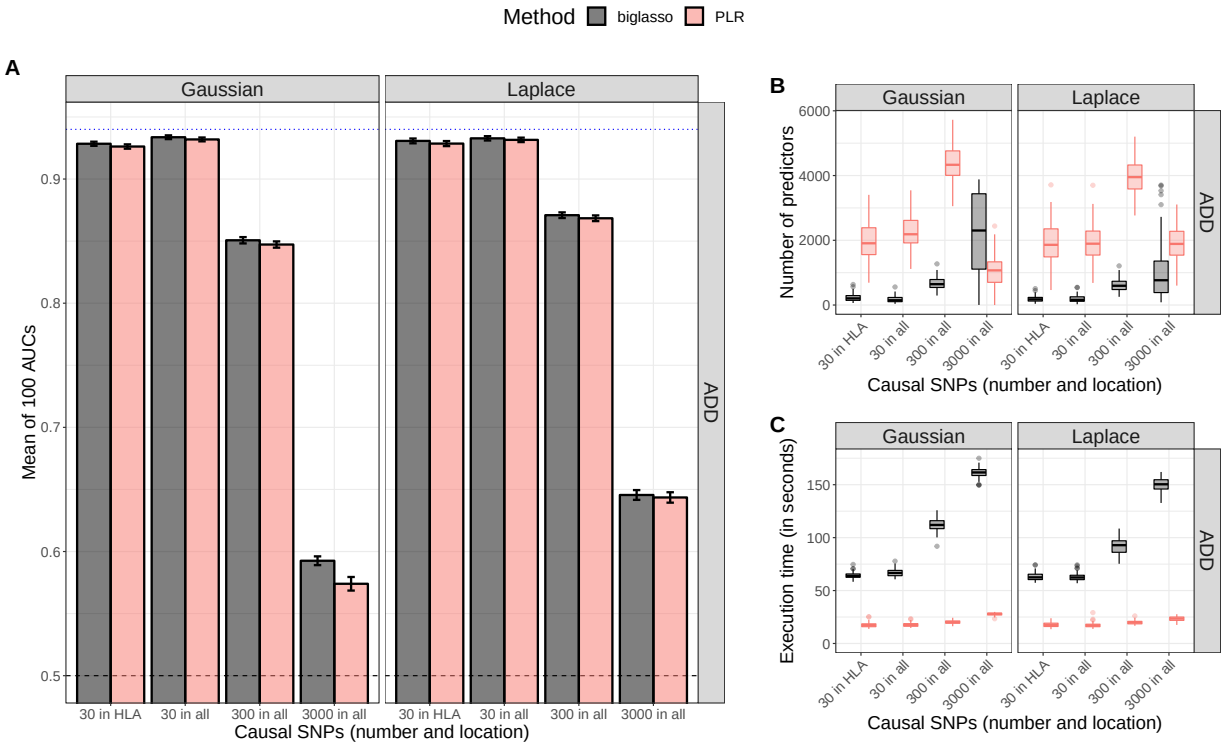

Figure S8: Comparison of PLR and the best prediction (among 100 tested λ values) for "biglasso" (another implementation of penalized logistic regression – Zeng and Breheny (2017)) in scenario №1. Simulations use model "ADD", an heritability of 80% and $\alpha = 1$. Vertical panels are presenting results for effects following a Gaussian or Laplace distribution. **A:** Mean AUC over 100 simulations. Error bars are representing $\pm$2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC. **B:** Boxplots of numbers of predictors used by the methods for 100 simulations. **C:** Boxplots of execution times for 100 simulations.