

Predicting complex diseases: performance, robustness and ..utility..(impact?)

Florian Privé ^{1,*}, Hugues Aschard ^{2,3} and Michael G.B. Blum ^{1,*}

¹Université Grenoble Alpes, CNRS, Laboratoire TIMC-IMAG, UMR 5525, France,

²Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris,
France

³Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts,
USA.

*To whom correspondence should be addressed.

11

Abstract

12

Motivation:

13

Results:

14

Availability:

15

Contact: florian.prive@univ-grenoble-alpes.fr & michael.blum@univ-grenoble-alpes.fr

16

Supplementary information:

1 Introduction

Polygenic Risk Scores (PRSs) consist in combining the information across many single-nucleotide polymorphisms (SNPs) in a score reflecting the genetic risk of developing a disease. PRSs have proven to be useful for different applications, such as finding a common genetic contribution between two diseases (Purcell *et al.* 2009). Personalized medicine is another major application of PRSs. Personalized medicine will use PRSs in screening campaigns in order to identify high-risk individuals for a given disease. As an example of practical application, targeting screening to men at higher polygenic risk could reduce the problem of overdiagnosis and lead to a better benefit-to-harm balance in screening for prostate cancer (Pashayan *et al.* 2015). Moreover, screening based on sequencing seems to make individuals positively change their health behavior while not causing patient anxiety nor depression (Vassy *et al.* 2017). Yet, PRSs would have to show a high discriminative power between cases and controls in order to be used for helping in the diagnosis of diseases. For example, for respectively screening high-risk individuals and for presymptomatic diagnosis of the general population, it is suggested that the AUC must be greater than respectively 75% and 99% (Janssens *et al.* 2007).

Many methods have been developed in order to maximize the predictive power of genotypes for diseases, or more generally for phenotypes. A commonly used technique, called P+T (which stands for “Pruning and Thresholding”) – or genetic profiling ..or even just PRS.. – is used to derive PRSs from results of Genome-Wide Association Studies (GWASs) (Chatterjee *et al.* 2013; Dudbridge 2013; Evans *et al.* 2009; Purcell *et al.* 2009; Wray *et al.* 2007). This technique only use summary statistics which makes it very ..suitable.. and also very fast. Linear Mixed-Models (LMMs) are another widely-used method in fields such as plant and animal breeding or for predicting highly heritable quantitative human phenotypes such as height (Lello *et al.* 2017; Yang *et al.* 2010). Yet, these models are not ..constructed.. for predicting a binary trait such as a disease status and have proven to ..fail.. at such task in another comparative study (Abraham *et al.* 2013). Moreover, these methods and their ..derivations.. are often computationally demanding, both in terms of memory and time ..requirements.., which makes them unlikely to be used for prediction on large cohorts (Golan and Rosset 2014; Maier *et al.* 2015;

Speed and Balding 2014; Zhou *et al.* 2013). Statistical learning methods have also been used to derive PRSs for complex human disease by jointly estimating SNP effects. Such methods include logistic regression, Support Vector Machine (SVM) and random forests (Abraham *et al.* 2012, 2014; Botta 2013; Botta *et al.* 2014; Wei *et al.* 2009).

We recently developed two R packages, *bigstatsr* and *bigsnpr*, for efficient management and analysis of large-scale genome-wide data (Privé *et al.* 2017). Package *bigstatsr* includes efficient functions for computing penalized linear and logistic regressions on huge datasets. In this paper, we present a comprehensive comparative study of the P+T method, our penalized logistic regression and the T-Trees algorithm, which is a derivation of random forests and has shown exceptionally good predictive results in Botta *et al.* (2014). Note that the SVM model is expected to give similar results to logistic regression (Abraham *et al.* 2012) and therefore isn't added to the comparison. For the P+T model, we compare different thresholds of inclusion of SNPs. For the logistic regression, we include two novel approaches. First, we introduce a procedure that we call Cross-Model Selection and Averaging (CMSA) for choosing the amount of regularization used, which directly affects the number of SNPs included in the model. We also show how to use feature engineering in order to capture not only linear effects, but also recessive and dominant effects.

There have already been comparative studies about predicting human binary phenotypes based on genotype data. (Abraham *et al.* 2013) showed that penalized logistic regression and penalized SVM performed better than the standard P+T method and LMMs, for a wide-range of human diseases. Zhou *et al.* (2013) used simulations on real genotypes of Australian individuals to show that a hybrid of linear mixed models and sparse regression models performed better in a range of sparse to polygenic disease architectures. Spiliopoulou *et al.* (2015) investigated the effect of relatedness in prediction performance. Ware *et al.* (2017) evaluated best practices for using the P+T method and found out, for example, that the optimal inclusion threshold is trait-dependent. In this study, we particularly extend the work of Abraham *et al.* (2013) by using a simulation framework similar but ..richer.. than the one used in Zhou *et al.* (2013). In order to make our comparison as comprehensive as possible, we compare different architectures of disease (number, size and location of causal effects and heritability) with

different model of generating liability ..scores..., one with only linear effects, and one which combines linear, dominant and interaction effects. First, we quickly discard T-Trees as being the best method for predicting disease status based on genotypes. Then, we show that penalized logistic regression consistently performs better than the P+T method whereas predictive performance of the P+T method is very sensitive to the threshold of inclusion of SNPs, depending of the architecture of disease, as shown in Ware *et al.* (2017).

2 Methods

2.1 Genotype data

We use real genotypes of European individuals from a case/control celiac disease cohort (Dubois *et al.* 2010). Composition of this dataset is presented in table S1. Details of quality control and imputation for this dataset are available in Privé *et al.* (2017). For simulations, in order to remove the structure induced by the celiac disease status, we first keep only controls from this cohort. Secondly, in order to remove population structure, we keep only people from the UK. Thirdly, we further subset these British people based on deviation of robust Mahalanobis distance on principal components. This 3-step filtering results in 7102 individuals with minimal population structure (see ..SupMat..). We decided to get rid of population structure because it can affect the predictive performance of methods, which we will investigate in another study.

2.2 Simulations of phenotypes

To simulate phenotypes, we use the Liability Threshold Model (LTM) with a prevalence of 30%. We vary different parameters: the number of chromosomes used (either all 22 chromosomes or only the chromosome 6), the size of the training set (from 1000 to 6000 by 1000, and the corresponding test set is composed of the remaining individuals), the number of causal variants and their location (30, 300 or 3000 anywhere on the dataset or 30 only in the HLA region), the heritability (50% or 80%), the distribution of effects associated with causal SNPs (Normal or Laplace), and the model used to generate the genetic ..part.. of the phenotypes (one “simple

model” with only linear effects, and one “fancy model” which combines linear, dominant and interaction effects). For the “simple model”, we compute

$$y_i = \sum_{j \in E_{\text{causal}}} w_j \cdot \widetilde{G_{i,j}}$$

where w_j are weights (generated with a Gaussian or a Laplace distribution), $G_{i,j}$ is the allele count of individual i for SNP j and $\widetilde{G_{i,j}}$ corresponds to its standardized version (zero mean and unit variance for all SNPs). For the “fancy model”, we separate the causal SNPs in three equal sets $E_{\text{causal}}^{(1)}$, $E_{\text{causal}}^{(2)}$ and $E_{\text{causal}}^{(3)}$ ($E_{\text{causal}}^{(3)}$ is further separated in two equal sets, $E_{\text{causal}}^{(3.1)}$ and $E_{\text{causal}}^{(3.2)}$). We then compute

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

where $D_{i,j} = \mathbb{1} \{G_{i,j} \neq 0\}$ and $E_{\text{causal}}^{(q)} = \{e_k^{(q)}, k \in \{1, \dots, |E_{\text{causal}}^{(q)}|\}\}$. Note that for the interaction part of the model, we scale interactions, not the raw allele counts, so that corresponding SNPs still display a marginal effect. For both models, the y_i are then standardized such that they have a variance equal to the desired heritability h^2 and we further add some environmental noise ϵ_i to y_i where $\epsilon \sim N(0, 1 - h^2)$.

We make 5 different simulation procedures, summarized in table 1. ..Simulation.. №1 uses the whole dataset (the 22 chromosomes), a training set of size 6000 and an heritability of 80%. It compares all methods (see section 2.4) on all other combination of parameters in 5 simulations. The goal of the first simulation is to quickly discard the T-Trees method. For simulation №2, we make also vary the heritability (50% or 80%) but we don’t run the T-Trees method anymore. The goal of the second simulation is to compare all remaining methods, especially to see how well the “logit-triple” method works. We run 20 simulations for each combination of parameters. For simulation №3, we continue with 80 more simulations (to reach 100 simulations, i.e. 5 times more than in Zhou *et al.* (2013)), but discarding the “logit-

triple” method and the “fancy” model. Simulation №4 consists in 100 simulations of remaining methods and combinations of parameters on a dataset composed of only the chromosome 6. Reducing the number of SNPs aims at increasing the ..polygenicity.. of our simulated models and at ..virtually.. increasing the sample size [REF? DUDBRIDGE?]. Finally, simulation №5 reuses the whole dataset but varying the size of the training set in order to assess how the number of samples affects predictive performance of the PRS and the “logit-simple” methods. 100 simulations are run using 300 causal SNPs anywhere on the genome.

2.3 Predictive performance measures

In this study, we use two different measures of predictive accuracy. First, we use the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) (Fawcett 2006; Hanley and McNeil 1982; Lusted 1971). In the case of our study, the AUC is the probability that a PRS of a case is greater than the PRS of a control. This measure indicates how well we can distinguish between cases and controls using PRSs. As a second measure, we report the percentage of cases in the 10% and 20% largest PRSs. This measure indicates how well we can identify high-risk individuals from a higher PRS in screening campaigns. [COMMENT ON PREVALENCE?]

Note that we also report the timing of the main computations and the number of SNPs used in the predictions.

2.4 Compared methods

In this study, we compare three different types of methods: the P+T method, the T-Trees method and a penalized logistic regression.

The P+T (Pruning + Thresholding) method directly derives a Polygenic Risk Score (PRS) from the results of Genome-Wide Associations Studies (GWASs), called summary statistics. For the P+T method, a coefficient of regression is learned independently for each SNP along with a corresponding p-value (the GWAS part). The SNPs are first clumped (P) so that there remain only loci that are weakly correlated with each other. Then, thresholding (T) consists in removing SNPs that are within a certain level of significance (p-value threshold to be deter-

mined). Finally, a polygenic risk score is defined as the sum of allele counts of the remaining SNPs weighted by the corresponding effect coefficients:

$$S_i(T) = \sum_{j \in E_{\text{clumping}}} \mathbb{1}\{p_j < T\} \cdot \beta_j \cdot G_{i,j}$$

where β_j (p_j) are the effect sizes (p-values) learned from the GWAS and T is a threshold on p-values to be determined. In this study, we report scores for a clumping threshold at $r^2 > 0.2$ within regions of 500kb.

For the P+T method, we report three different scores of prediction: one including all the SNPs (remaining after clumping), one including only SNPs that have a p-value under the GWAS threshold of significance ($p < 5 \cdot 10^{-8}$), and one that maximizes the AUC for these two thresholds and a sequence of 100 values of thresholds ranging from $10^{-0.1}$ to 10^{-100} and equally spaced on the log-log-scale (Table S2). Note that, normally, the optimal threshold would have to be learned on the training set only. We consider this reported maximum AUC as an upper bound of the AUC for the P+T method. We call these three reported scores respectively “PRS-all”, “PRS-stringent” and “PRS-max” in the results.

T-Trees (*Trees inside Trees*) is an algorithm derived from random forests (Breiman 2001) that takes into account the correlation structure among the genetic markers implied by linkage disequilibrium in GWAS data (Botta *et al.* 2014). Since this method is significantly different from the other tested here and achieved surprisingly results in the past, we decided to include T-Trees in this comparison in order to assess whether or not it could give superior predictive performance. We use the same parameters as reported in Table 4 of Botta *et al.* (2014). Yet, we use only 100 trees (instead of 1000) because it showed only a very subtle increase of AUC while taking way too much time (e.g. AUC of 81.5% instead of 81%, data not shown).

The last method we compare is penalized logistic regression. We solve:

$$\underset{\beta_0, \beta}{\operatorname{argmin}}(x, y, \lambda) \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^n \log \left(1 + e^{-y_i(\beta_0 + x_i^T \beta)} \right)}_{\text{Loss function}} + \underbrace{\lambda \left((1 - \alpha) \frac{1}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right)}_{\text{Penalization}} \right\}$$

where, in this study, x is the genotypes and covariables (e.g. principal components), y is the disease status we want to predict and λ is a regularization parameter that need to be determined. Different regularizations can be used to prevent overfitting, among other benefits: the L2-regularization (ridge, Hoerl and Kennard (1970)) shrinks coefficients and is ideal if there are many predictors drawn from a Gaussian distribution (corresponds to $\alpha = 0$ in the previous equation), the L1-regularization (lasso, Tibshirani (1996)) forces some of the coefficients to be exactly zero and can be used as a means of variable selection, leading to sparse (and therefore more interpretable) models (corresponds to $\alpha = 1$), the L1- and L2-regularization (elastic-net, Zou and Hastie (2005)) is a compromise between the two previous penalties and is particularly useful in the $m \gg n$ situation (m : number of SNPs), or any situation where there are many correlated predictors (corresponds to $0 < \alpha < 1$) (Friedman *et al.* 2010). In this study, we always use the elastic-net regularization with $\alpha = 0.5$, without trying to tune (e.g. by grid-search and cross-validation) the value of this hyper-parameter α .

To fit this penalized logistic regression, we use a very efficient algorithm (Friedman *et al.* 2010; Tibshirani *et al.* 2012; Zeng *et al.* 2017), also implemented in our package bigstatsr (Privé *et al.* 2017). This type of algorithm builds predictions for many values of λ (typically a “regularization path” of 100 values). To get an algorithm free of the choice of this hyper-parameter λ , we developed a procedure that we called Cross-Model Selection and Averaging (CMSA). 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 and the others ($K - 1$) folds form an inner training set, the model is trained on the inner training set and the corresponding predictions (scores) for the inner validation set are computed, the vector of scores which maximizes a criterion (e.g. AUC) is determined and the vector of coefficients corresponding to the previous vector of scores is chosen. Finally, the K resulting vectors of coefficients are combined into one vector (e.g. using the geometric median). Because of L1-regularization, this vector of coefficients is typically very sparse and can be used to make a PRS based on a *linear* combination of allele counts. We call this method “logit-simple” in the results.

In order to capture recessive and dominant effects in addition to linear effects, we use feature engineering: we construct another dataset with, for each SNP variable, two more variables

coding for recessive and dominant effects. This results in a dataset with 3 times more variables than the initial one, on which we can apply the standard logistic regression with the CMSA procedure, described previously. We call this method “logit-triple” in the results.

2.5 Reproducibility

All the code used in this paper along with results such as figures and tables, are available as HTML R notebooks in the Supplementary Materials.

3 Results

[MONTRER UN MANHATTAN PLOT POUR JUSTIFIER 30 IN HLA]

[MONTRER UN GRAPH EN FCT D’INCLUSION]

3.1 Maximum AUCs

[MOVE IN SUPMAT AND METHODS]

Assuming that $y \sim N(0, h^2)$, we can estimate the theoretical value of the AUC that can be achieved given the heritability h^2 (see ..SupMat..). We report AUCs of 0.8410 and 0.9406 for respectively an heritability of 50% and 80%.

3.2 Overview and naming

3.3 T-Trees

First, we ran only 5 simulations for each combination of parameters), excepted for the heritability that we fixed at 0.8 (see section 2.2). The goal was to quickly discard the T-Trees method as being competitive with the other methods. For example, compared to our penalized logistic regression, T-Trees perform worse for both predictive measures, the AUC and the percentage of cases in highest scores (Figure ??). Moreover, T-Trees takes longer to run and makes more complex predictive models because it uses more SNPs in the models and has non-linear effects (Figure ??). So, we decided to discard this method for the rest of the simulations.

3.4 Scenario №1

In scenario №1, we compared our penalized logistic regression (“logit-simple”) with the P+T method (“PRS”) on the whole genome (Table 1). When there are only 30 causal SNPs, our penalized logistic regression almost reach the maximum achievable AUC (Figures 1 & S1). Still in terms of AUC, our penalized regression consistently performs better or equally to the “PRS-max”. In the case where there are correlation between predictors (i.e. when we simulate 30 SNPs in the HLA region), the “logit-simple” method particularly performs better than the “PRS-max” one, with a mean AUC of 92.5% versus 83%, for a maximum achievable AUC of 94% (Figure 1). [NUL] Moreover, results of PRS can dramatically drop given the threshold of inclusion of SNPs in the model (Figures 2 & S2).

3.5 Main simulations

First, we show that AUC is highly correlated ($r^2 > 95\%$) with the percentage of cases in either the 10% or 20% largest PRSs (Figure ??). Therefore, AUC can be a good measure at determining whether a score can identify high-risk individuals or not, so we report only predictive performance in terms of AUC in the following results.

Using the penalized logistic regressions, we report AUCs greater than 95%, which is not obtained by the upper bound of the P+T method in any scenario (Figure ??). Our penalized regression methods consistently provides better predictive performance than the P+T method for all disease architectures tested [ou pas, FIGURE]. Our method is parallelized and requires only less than 10 minutes to run (with 6 physical cores) [TABLE ?FIGURE?]. Moreover, it results in relatively sparse models with no more than 14,000 predictors entering the models even for the linear model of 3000 causal SNPs [TABLE ?FIGURE?]. P+T runs in only a few seconds because we don’t consider the GWAS computation as part of the execution time of the method because this is generally executed by others (e.g. consortia) and then only the resulting summary statistics are used to construct a PRS.

As for the comparison between our two penalized logistic regression, the classic one (called “simple”) and the one using additional features coding for recessive and dominant effects

(called “triple”), we report that the “triple” method is nearly as good as the “simple” one when there are only linear effects and can lead to significantly greater results when there are also dominant and interactions effects (Figure ??). Yet, the “triple” solution takes approximately 3 more times to run.

As for general results, we report a decrease of AUC when the number of causal variants M is increasing [BATEAU? quoi d’autre?].

[TROUVER DES AUTRES: % of decrease for h^2 0.8 \rightarrow 0.5? Pas bon d’utiliser tous les SNPs]

3.6 Simulations varying training size and number of SNPs

4 Discussion

In this comparative study, we proposed other methods as a replacement to the widely used P+T method. We show that jointly estimating the SNP effects using efficient penalized logistic regressions enable to achieve really good performance, as reported in a previous study (Abraham *et al.* 2013), higher than with the P+T method. We are well aware that the P+T method is widely used because it is simple and can use summary statistics already learned from large genotype data, which is not the case for our penalized logistic regressions. Yet, in the case of prediction of disease to be used for clinical interventions, best possible methods should be used, even if it has to be used by large consortia only [EVEN LARGE CONSORTIA DON’T HAVE ALL DATA?].

One does not simply identify high-risk individuals (Figure ??).

[PARLER DU PROBLEME DE POP? + future paper]

Table 1: Summary of all simulations. When there is symbol ‘-’ in a box, it means that the parameters are the same as the ones in the upper box.

Numero of simulation	Dataset	Size of training set	Causal SNPs (number and location)	Distribution of effects	Heritability	Simulation model	Number of simulations	Methods
1	All 22 chromosomes	6000	30 in HLA 30 in all 300 in all 3000 in all	Gaussian Laplace	0.8	simple fancy	5	PRS logit-simple logit-triple T-Trees
2	-	-	-	-	0.5 0.8	-	20	PRS logit-simple logit-triple
3	-	-	-	-	-	simple	80	PRS logit-simple
4	Chromosome 6 only	-	-	-	-	-	100	-
5	All 22 chromosomes	1000 2000 3000 4000 5000	300 in all	-	-	-	100	-

Acknowledgements

Authors acknowledge Grenoble Alpes Data Institute, supported by the French National Research Agency under the “Investissements d’avenir” program (ANR-15-IDEX-02) and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01). We are also grateful to Félix Balazard for useful discussions about T-Trees.

References

- Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2012). Sparsnp: Fast and memory-efficient analysis of all snps for phenotype prediction. *BMC bioinformatics*, **13**(1), 88.
- Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology*, **37**(2), 184–195.
- Abraham, G., Tye-Din, J. A., Bhalala, O. G., Kowalczyk, A., Zobel, J., and Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS genetics*, **10**(2), e1004137.
- Botta, V. (2013). *A walk into random forests: adaptation and application to Genome-Wide Association Studies*. Ph.D. thesis, Université de Liège, Liège, Belgique.
- Botta, V., Louppe, G., Geurts, P., and Wehenkel, L. (2014). Exploiting snp correlations within random forest for genome-wide association studies. *PloS one*, **9**(4), e93379.
- Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.
- Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics*, **45**(4), 400–405.
- Dubois, P. C., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zernakova, A., Heap, G. A., Ádány, R., Aromaa, A., *et al.* (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, **42**(4), 295–302.
- Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, **9**(3), e1003348.
- Evans, D. M., Visscher, P. M., and Wray, N. R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human molecular genetics*, **18**(18), 3525–3531.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, **27**(8), 861–874.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**(1), 1.
- Golan, D. and Rosset, S. (2014). Effective genetic-risk prediction using mixed models. *The American Journal of Human Genetics*, **95**(4), 383–393.
- Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143**(1), 29–36.

- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.
- Janssens, A. C. J., Moonesinghe, R., Yang, Q., Steyerberg, E. W., van Duijn, C. M., and Khoury, M. J. (2007). The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genetics in Medicine*, **9**(8), 528–535.
- Lello, L., Avery, S. G., Tellier, L., Vazquez, A., Campos, G. d. l., and Hsu, S. D. (2017). Accurate genomic prediction of human height. *arXiv preprint arXiv:1709.06489*.
- Lusted, L. B. (1971). Signal detectability and medical decision-making. *Science*, **171**(3977), 1217–1219.
- Maier, R., Moser, G., Chen, G.-B., Ripke, S., Absher, D., Agartz, I., Akil, H., Amin, F., Andreassen, O. A., Anjorin, A., *et al.* (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *The American Journal of Human Genetics*, **96**(2), 283–294.
- Pashayan, N., Duffy, S. W., Neal, D. E., Hamdy, F. C., Donovan, J. L., Martin, R. M., Harrington, P., Benlloch, S., Al Olama, A. A., Shah, M., *et al.* (2015). Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genetics in Medicine*, **17**(10), 789–795.
- Privé, F., Aschard, H., and Blum, M. G. (2017). Efficient management and analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr. *bioRxiv*, page 190926.
- Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O’donovan, M. C., Sullivan, P. F., Sklar, P., Ruderfer, D. M., McQuillin, A., Morris, D. W., *et al.* (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**(7256), 748–752.
- Speed, D. and Balding, D. J. (2014). Multiblup: improved snp-based prediction for complex traits. *Genome research*, **24**(9), 1550–1557.
- Spiliopoulou, A., Nagy, R., Bermingham, M. L., Huffman, J. E., Hayward, C., Vitart, V., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., *et al.* (2015). Genomic prediction of complex human traits: relatedness, trait architecture and predictive meta-models. *Human molecular genetics*, **24**(14), 4167–4182.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(2), 245–266.
- Vassy, J. L., Christensen, K. D., Schonman, E. F., Blout, C. L., Robinson, J. O., Krier, J. B., Diamond, P. M., Lebo, M., Machini, K., Azzariti, D. R., *et al.* (2017). The impact of whole-genome sequencing on the primary care and outcomes of healthy adult patients: A pilot randomized trial. *Annals of internal medicine*, **167**(3), 159–169.
- Ware, E. B., Schmitz, L. L., Faul, J. D., Gard, A., Mitchell, C., Smith, J. A., Zhao, W., Weir, D., and Kardia, S. L. (2017). Heterogeneity in polygenic scores for common human traits. *bioRxiv*, page 106062.
- Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J. T., Chiavacci, R., *et al.* (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS genetics*, **5**(10), e1000678.
- Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research*, **17**(10), 1520–1528.
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., *et al.* (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, **42**(7), 565–569.

309 Zeng, Y., Breheny, P., and Yang, T. (2017). Efficient feature screening for lasso-type problems via hybrid safe-strong rules. *arXiv preprint*
310 *arXiv:1704.08742*.

311 Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, **9**(2), e1003264.

312 Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical*
313 *Methodology)*, **67**(2), 301–320.

Supplementary Data

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 for the celiac disease cohort (after quality control, genotyped on 281,122 ..mutual.. 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 in the P+T method for this study.

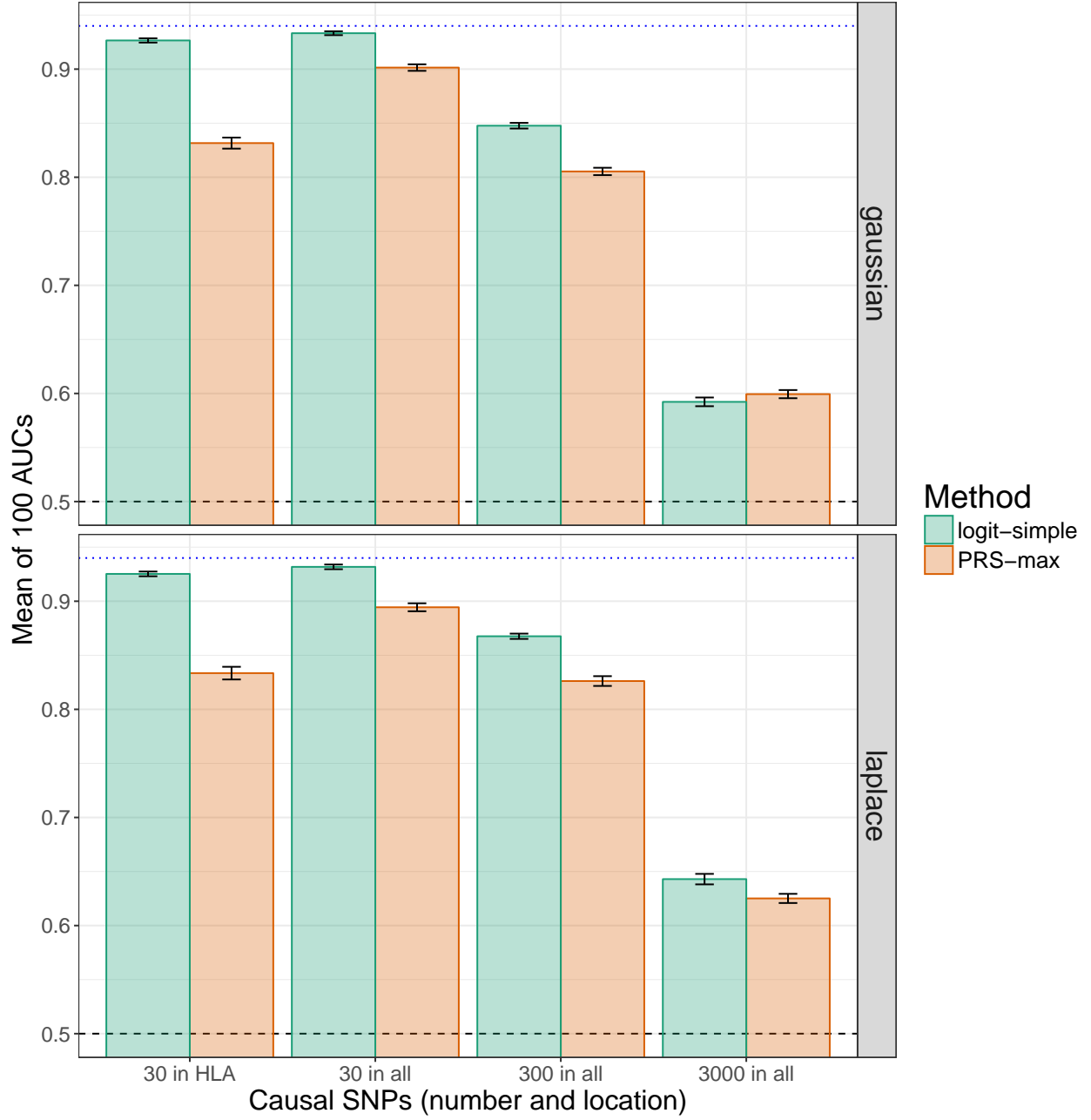


Figure 1: Mean of AUC over 100 simulations for our penalized logistic regression (“logit-simple”) and the maximum AUC reported with the P+T method (“PRS-max”). Upper (lower) panel is presenting results for effects following a Gaussian (Laplace) distribution. Error bars are representing $\pm 2SD$ of 10^5 non-parametric bootstrap of the mean of AUC. The blue dotted line represents the maximum achievable AUC. [MODEL AND H2]

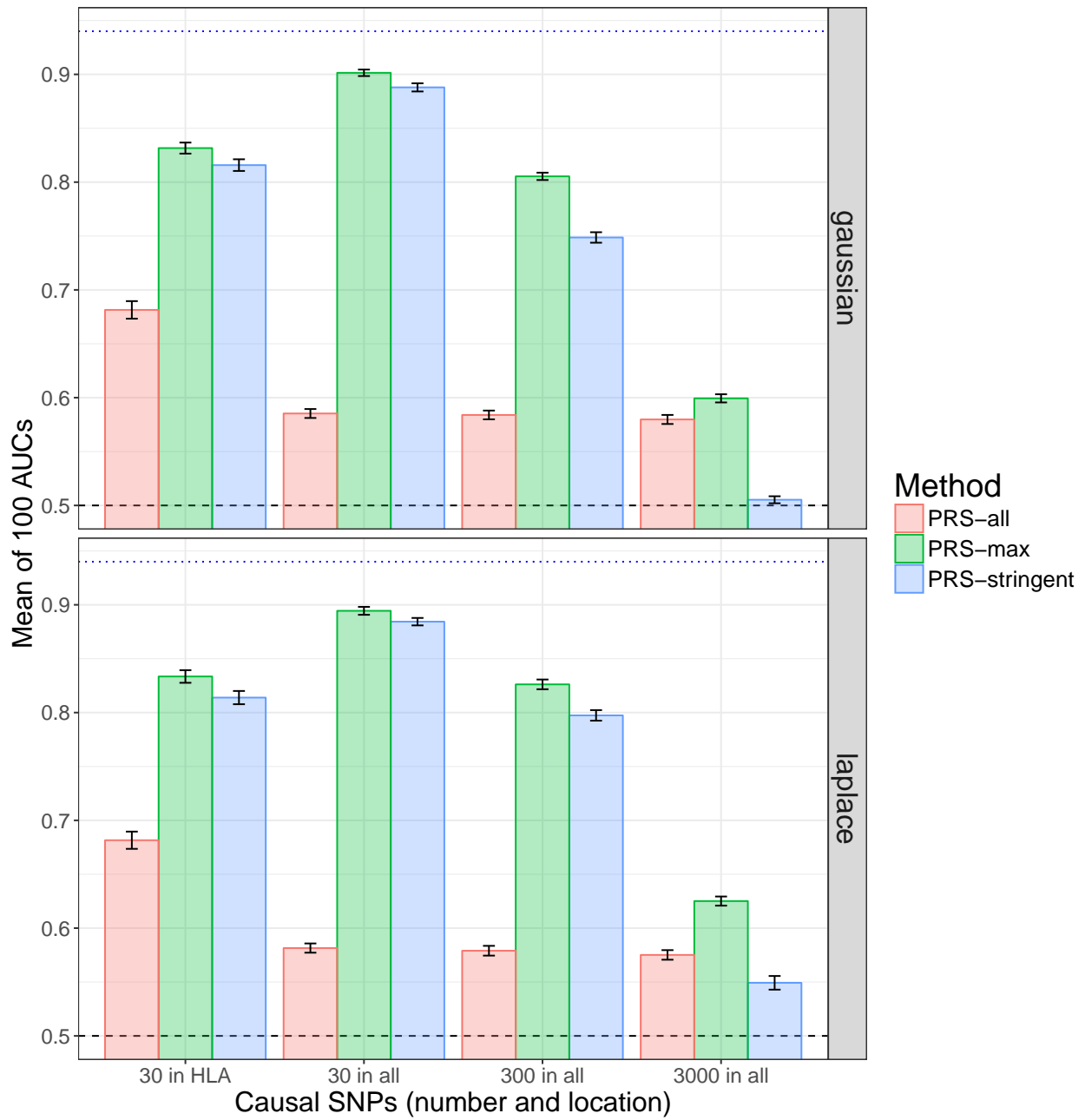


Figure 2: Mean of AUC over 100 simulations for ..the three reported PRSs... Upper (lower) panel is presenting results for effects following a Gaussian (Laplace) distribution. Error bars are representing $\pm 2SD$ of 10^5 non-parametric bootstrap of the mean of AUC. The blue dotted line represents the maximum achievable AUC. [MODEL AND H2]

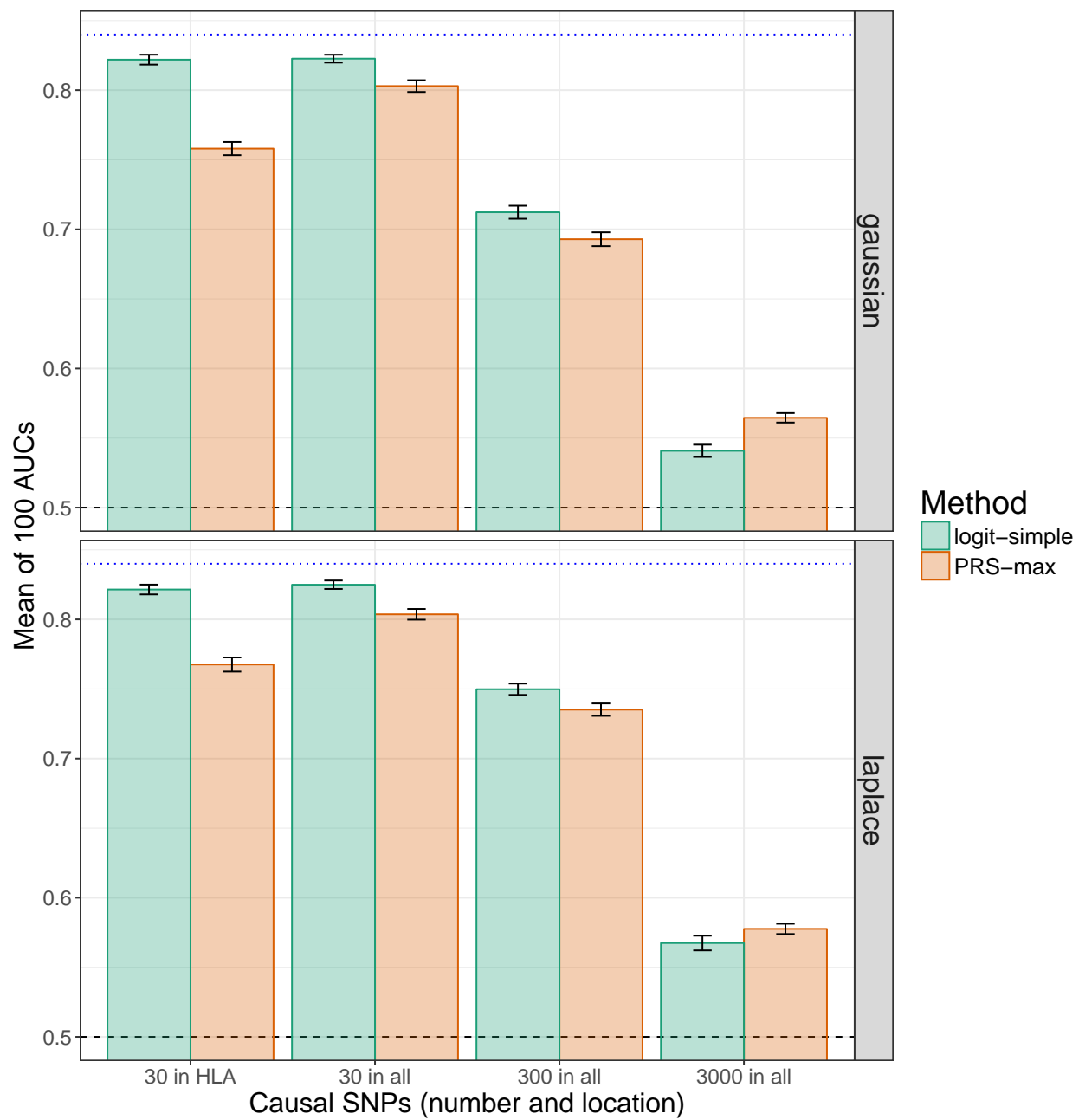


Figure S1: [COPY FOR H2=0.5]

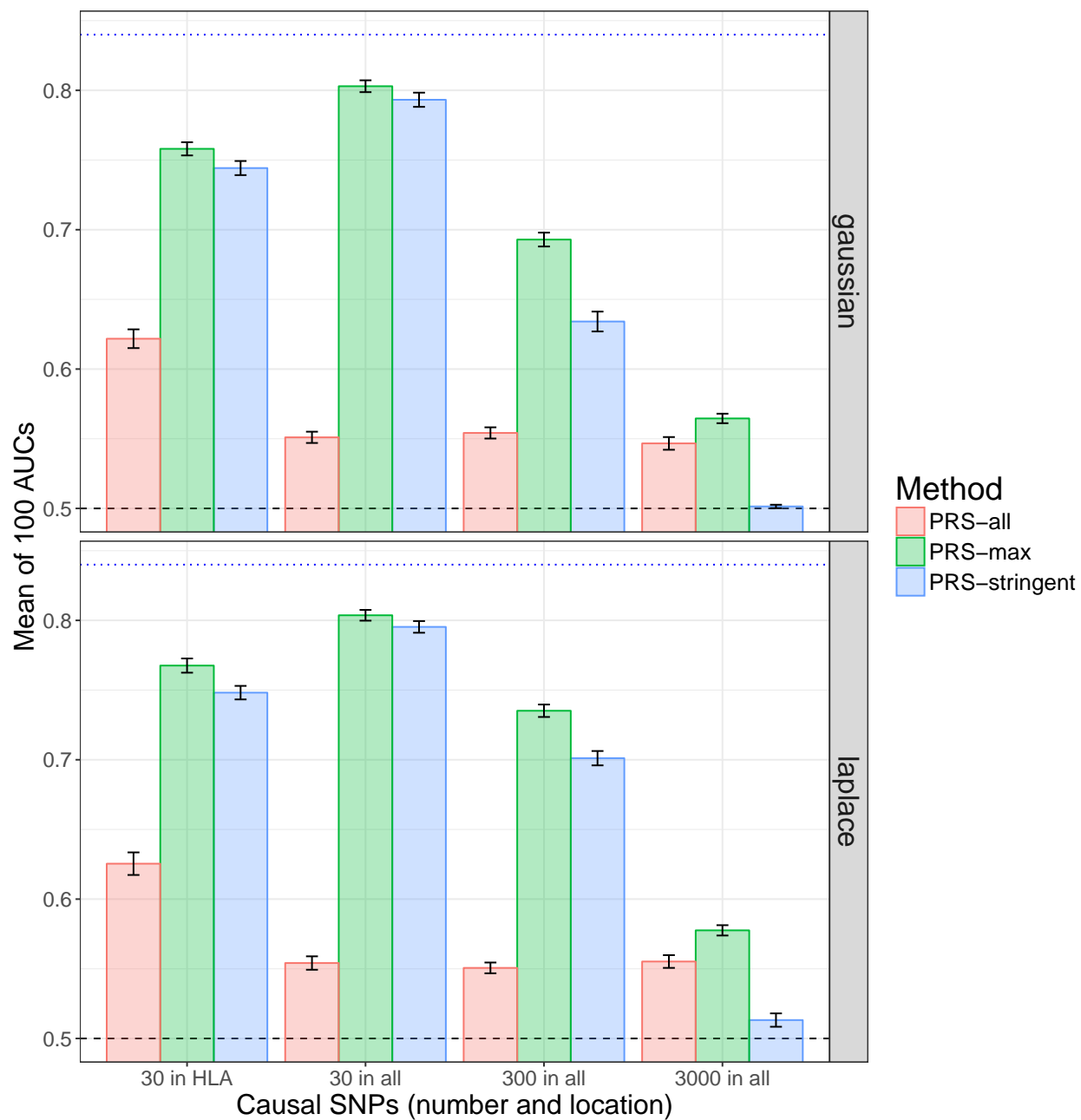


Figure S2: [COPY FOR H2=0.5]