Haplotype based genetic risk estimation for complex diseases

Félix Balazard

Genome-wide association studies (GWAS) have uncovered thousands of associations between genetic variants and diseases. Using the same datasets, prediction of disease risk can be attempted. Phase information is an important biological structure that has seldom been used in that setting. We propose here a multi-step machine learning method that aims at using this information. Our method captures local interactions in short haplotypes and combines the results linearly. We show that it outperforms standard linear models on some GWAS datasets. However, a variation of our method that does not use phase information obtains similar performance. Regarding the missing heritability problem, we remark that interactions in short haplotypes contribute to additive heritability. Source code is available on github at https://github.com/FelBalazard/Prediction-with-Haplotypes.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

Haplotype Based Genetic Risk Estimation for Complex Diseases

Félix Balazard[1,2]

[1]**Sorbonne Universités, UPMC Univ Paris 06, CNRS, Paris, France**
[2]**INSERM U1169, Hôpital Bicêtre, Université Paris-Sud, Kremlin-Bicêtre, France**

ABSTRACT

Genome-wide association studies (GWAS) have uncovered thousands of associations between genetic variants and diseases. Using the same datasets, prediction of disease risk can be attempted. Phase information is an important biological structure that has seldom been used in that setting. We propose here a multi-step machine learning method that aims at using this information. Our method captures local interactions in short haplotypes and combines the results linearly. We show that it outperforms standard linear models on some GWAS datasets. However, a variation of our method that does not use phase information obtains similar performance. Regarding the missing heritability problem, we remark that interactions in short haplotypes contribute to additive heritability. Source code is available on github at https://github.com/FelBalazard/Prediction-with-Haplotypes.

Keywords: GWAS, Genetic Risk Estimation, Machine Learning

INTRODUCTION

Genome-wide association studies (GWAS) have used micro-array technology to genotype hundreds of thousands of single nucleotide polymorphisms (SNPs) in thousands of patients and controls. The main goal of those studies has been to identify associations between SNPs and diseases that could help understand the genetic component of the disease. The methodology used for this purpose relies on univariate hypothesis tests with corrections for multiple testing. GWAS have unravelled over one thousand new SNP disease associations (Welter et al., 2014).

A potential clinical utility of GWAS is implementation of personalized medicine. For example, genetic risk prediction could be useful for prevention of complex diseases. Unfortunately, the SNPs found significant in GWAS are not sufficient in aggregate to be used for prediction of disease status (Manolio et al., 2009). However, it is not necessary that each variable passes a stringent p-value threshold to be useful in a multivariate setting. In order to increase predictive power, one approach has been to use a lenient significance threshold to preselect SNPs before applying a machine-learning algorithm such as support vector machine or lasso regression in type 1 diabetes (Wei et al., 2009) and Crohn's disease (Wei et al., 2013). The preselection step allowed for manageable computation time. Optimization efforts were made to allow L^1-penalized linear regression with square-hinge loss to be run over the whole dataset (Abraham et al., 2012). This was applied to celiac disease (Abraham et al., 2014). All those approaches led to significant improvement of predictive power compared to including only GWAS significant SNPs.

The methodology used in those articles is to apply general purpose machine learning algorithm to GWAS datasets. The biological structure of genetic data is therefore not taken into account. A first example of such a structure is distance inside chromosomes measured in base pairs. In (Botta et al., 2014), the T-trees method was introduced to capture interactions inside small blocks of nearby SNPs as well as between blocks. The rationale is that SNPs that are next to each other are more likely to impact the same function and therefore to interact (in the statistical sense) together. The T-Trees method is a variation on the random forests (RF) (Breiman, 2001) algorithm tailored to focus on local interaction between SNPs. It can also assess the importance of individual SNPs as well as the importance of blocks of SNPs. It is very successful in increasing predictive performance compared to RF or linear methods. It also identifies new associations between loci and disease.

SNPs can take three possible values – 0, 1 or 2 – coding the number of copies of the mutant allele

present in the pair of chromosomes. This coding is appropriate to perform univariate tests. In a multivariate setting, this coding does not allow to know if mutant alleles of two heterozygous SNPs on the same chromosome pair are on a single chromosome or on the two distinct chromosomes of the chromosome pair. The sequences on the two homologous chromosomes are called haplotypes and phase information is knowledge of haplotypes instead of genotype. It is an important biological information (Tewhey et al., 2011). For example, if the two alleles of a gene have a distinct non-sense mutation, a condition called compound heterozigosity, there will be no expression of the gene. In contrast, if the two mutations were on the same chromosome, there would still be a functioning allele. This is a simple example where phase information is crucial.

Phase information is the second structure that we will use in our design of a machine learning algorithm tailored to GWAS data. It complements chromosomal distance. It is reasonable to expect that two SNPs that are physically on the same chromosome and not too distant are more likely to interact than if not. Interaction, here and throughout this article, is understood in the statistical sense as a departure from linear effects. Previous work has used haplotypes of two contiguous SNPs with a simple methodology for prediction of Crohn's disease (Kang et al., 2011). Their results are suggestive of the interest of haplotypes in a predictive context. Phase information is not available using micro-array technology. However, computational methods have been developed to allow phase imputation (Delaneau et al., 2013). They have limited accuracy which means that only short haplotypes should be used.

Up to this point, we only discussed the potential interest of haplotypes regarding prediction accuracy but haplotypes are also interesting for heritability. Heritability quantifies the proportion of phenotypic variance explained by genetic factors. It is estimated through family studies. It can give upper bounds for prediction accuracy (Wray et al., 2010). A distinction exists between broad-sense heritability and narrow-sense heritability (Visscher et al., 2008). The latter only includes additive effects while the former also includes interaction terms such as dominance and epistasis. The rationale behind this distinction is that when estimating heritability with pedigrees, one only estimates narrow-sense heritability. Dominance and epistasis are lost due to genetic mixing. However, interactions in haplotypes are actually part of narrow-sense heritability as they are shared among all members of the same family. Haplotypes are the support of heredity and not single SNPs. Narrow-sense heritability includes additive effects of *haplotypes*. Of course, long haplotypes are broken by recombination but short haplotypes are seldom concerned. This may sound counter-intuitive but the interactions inside haplotypes are part of additive heritability. This is the main theoretical contribution of our work.

Considering interaction in haplotypes is more general than the idea that for each association signal at a locus there is a causal variant responsible for it. If there is a causal variant that is not part of the typed SNPs but that is associated with a particular haplotype, capturing interaction in haplotype should recover this variant's effect better than relying on unphased data. If the variant is only in a subset of the haplotype, the effect will be diluted but will still be captured more precisely. Moreover, it is possible that there exists haplotypic effects not linked to a single variant.

The contribution of this paper is to introduce a multi-step machine-learning method –noted PH for Prediction with Haplotypes– that captures interactions in short haplotypes centered around association signal, then combines the results using Lasso regression. This can be seen as logistic regression by blocks. In order to know what phase information adds to the analysis, we also applied a similar method on genotypes and not haplotypes. We also adapt our method to capture dominance effect between the two haplotypes at a same loci.

We compare our method and its two variations to lasso regression with preselection on GWAS datasets made available by the Wellcome Trust Case Control Consortium (WTCCC) (Burton et al., 2007).

MATERIALS AND METHODS

In this section, we first briefly describe the machine learning methods used in PH: lasso logistic regression and random forests. A more detailed presentation of those techniques is available in the monograph (Friedman et al., 2001). We then present and motivate our PH algorithm. We conclude with a description of the experimental protocol and quality control filters used in this study.

We introduce a few notations: we have n observations (in our case patients) of p variables (for example, SNPs) that we can summarize in an n by p matrix $X = (x_{ij})$. The value of variable j for observation i is x_{ij}. We also have a binary response variable $Y = (y_i)$ that we want to predict using the other variables. In our case, it is disease status and $y_i = 1$ if individual i is diseased and 0 otherwise.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

Lasso logistic regression

In logistic regression, the posterior probability of being a case or a control is modelled by a linear combination of the variables :

$$\log \frac{P(y=1|X=x)}{1-P(y=1|X=x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

The vector $\beta = (\beta_0, \beta_1, ..., \beta_p)$ of weights is chosen to maximize the likelihood of the training data. When the dimension p becomes large compared to n, the maximum likelihood estimate will closely fit to the training data but have no predictive power on test data. This problem is called overfitting. To adress this issue, several penalization procedures have been proposed. They force the model to be simpler and keep predictive power (Friedman et al., 2001, p61-73). We will use L^1 penalization also known as Lasso (Tibshirani, 1996). It has the nice additional property of sparsity: some variable's coefficients will be assigned to 0 which makes the model more interpretable.

We will refer to the function $x \in (0,1) \mapsto \log(\frac{x}{1-x})$ as evidence following the terminology of Jaynes (Jaynes, 2003, p116-117). It is the link function in logistic regression. It is sometimes refered to as log-odds or logit. It maps $(0,1)$ to $(-\infty, \infty)$. For independent variables, odd ratios multiply and hence evidence is additive.

Decision trees and random forests

Decision trees are non-linear machine learning algorithms. They can capture interactions between variables and are easily interpretable. They are represented by a binary tree with a binary test of the form $x_j > c$ on each node (Friedman et al., 2001, p305-316).

However, single decision trees are often poor classifiers. They also are very unstable as alls splits are conditioned by the first split. To take advantage of this unstability, the random forests (RF) algorithm was designed (Breiman, 2001). It consists in randomizing the growth process of the tree, growing many independent such random trees and aggregating the results of all the trees (Friedman et al., 2001, p587-604). It is very effective in increasing predictive accuracy. An importance score can be attributed to variables. However, the classifier is less interpretable than a single decision tree. RF's use in computational biology and its challenges are reviewed in (Boulesteix et al., 2012).

One of the ways that the trees are randomized is that they use bootstrap versions (subsets obtained by sampling with replacement) of the training data to train different trees. This means that a specific observation will not be used to train all the trees. For the trees that did not use the observation, we say the observation is out of bag. For each observation, we can look at all the trees for which it is out-of-bag and aggregate the predictions of those trees for the observation. This allows to have predictions on the training set that should behave similarly to predictions on the test set i.e. without overfitting. This is critical in our setting.

Prediction with Haplotypes

Motivation The diploid nature of the genome is an important structure left mostly unused in earlier attempts at genetic risk estimation. It is challenging to use this information from a machine learning point of view. Indeed, once phasing is performed, we have the same set of variables twice but with different values and a metric structure. Interactions inside short haplotypes are what we aim to exploit thanks to phase information.

Algorithm A preliminary step is to use Shapeit 2 (Delaneau et al., 2013) to obtain estimates of haplotypes.

The first step of the algorithm is doing univariate test of association of SNPs to disease on the unphased training set. This is done using PLINK (Purcell et al., 2007). This is to work with a computationally manageable number of variables. We define blocks around the most associated SNPs. Those blocks consist of all the SNPs (not only highly associated one) under a fixed distance in kb (thousand of base pair) from the associated (or central) SNP as shown in Fig.1. The window size L_w is an important hyper-parameter with biological signification. The order of magnitude we used for L_w is 10 kb. Blocks are allowed to overlap but the central SNP of a block must be outside of the other blocks. Therefore, a highly associated SNP will not be used to define a block if its distance with a more highly associated SNP is smaller than the half window size i.e. the SNP is already included in a block. Besides reduced computation, the motivation of centering the blocks on associated SNPs compared to using a fixed grid

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

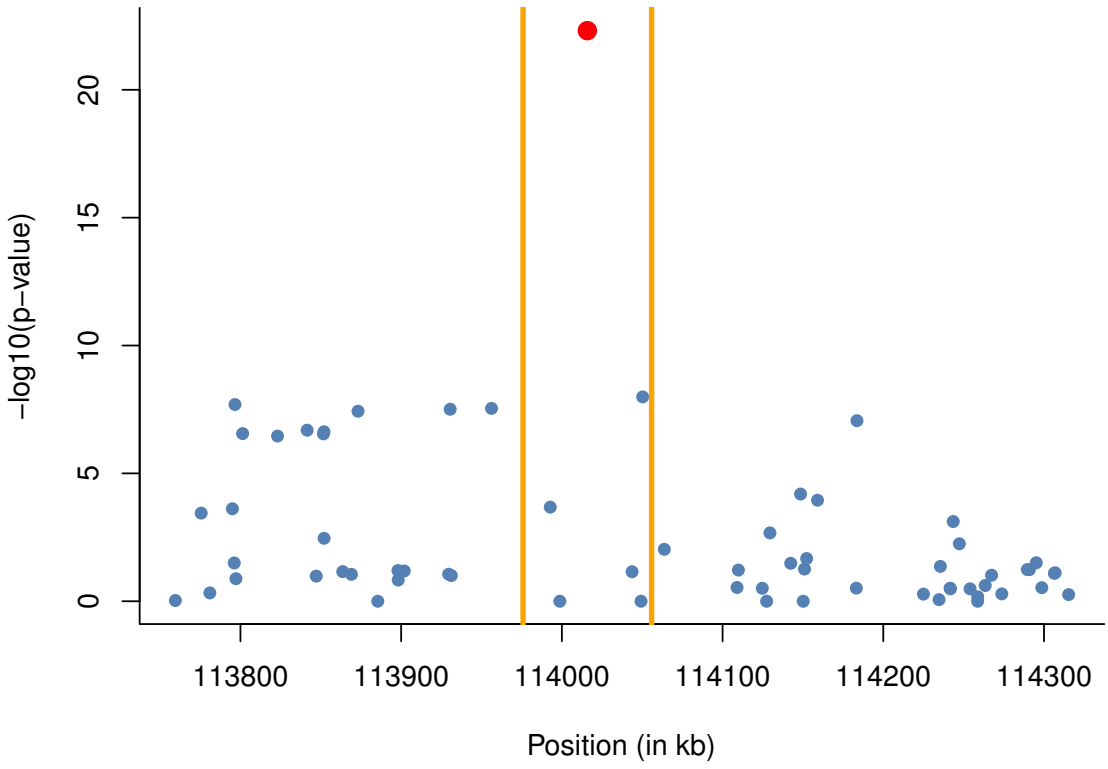

Figure 1. Block definition around associated SNP Blocks include all SNPs at distance smaller than L_w of the central SNP.

like in (Botta et al., 2014) is to be able to capture the important interactions. If two interacting nearby SNPs fall by misfortune on both sides of the border between two blocks, their interaction will not be captured. It seems reasonable to assume that the locally highest associated SNP will be part of the local interaction if there is one. The number of blocks N_b is another hyper-parameter.

Inside a block, we want to capture interactions inside haplotypes. For each observation, there are two haplotypes and therefore, we have two times the same set of variables with different values. We treat each haplotype as a distinct observation and attribute it the response variable of the individual it belongs to. We train random forests on the haplotypes of the training set and this gives us an estimated probability that the haplotype belongs to a diseased person. This estimated probability is the out-of-bag estimate for haplotypes belonging to the training set and the prediction using the full forest for the test set. The Gini impurity was used as node-splitting criterion. The default value for classification was used for the m_{try} parameter. We tried different values for N_{leaf} the minimum number of data points in a leaf. At the level of the haplotype, we are interested in estimating probabilities and not in classification, therefore the mean (over the forest) predicted probability was used as the method of aggregation of results. As the computation for one block are independent from the computation in the other blocks, computation was parallelized.

Every individual has two estimated probability of being sick that come from the two independent haplotypes. We combine those by adding the evidence of being sick given by the two haplotypes i.e. we take the evidence or log-odds of each probability and add them. This gives us a new variable that is the evidence of being sick given the haplotypes in the block. There is one such variable for each block. This step combines the results for the two haplotypes in a principled way and it builds a variable that is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

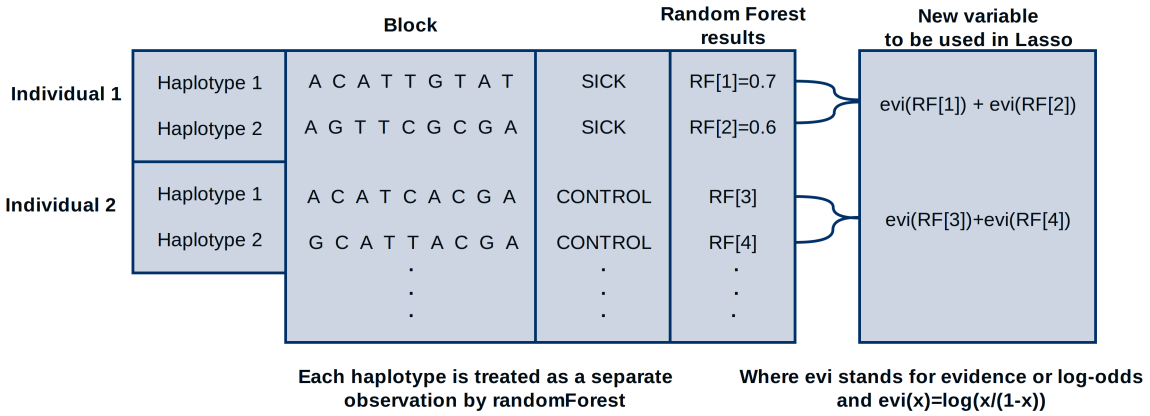

Figure 2. Interactions inside haplotypes captured with Random Forests Inside a block, each haplotype is treated as a distinct observation. The SNPs in the block are the input variables used to train a RF predictor. The results are then combined into a variable that summarizes the information contained in the block.

homogeneous to logistic regression. These two steps are illustrated in Fig. 2.

For each block, we obtain one variable that summarizes the information we obtained from it. We use those N_b variables as predictors of disease using Lasso regression. We train the Lasso regression on the block variables obtained for the training set. Using the trained regression model, we predict on the block variables obtained for the validation set. The full procedure with emphasis on training set and test set separation is summarized in Fig. 3.

Variation of the method The two variations of the method we considered differ from PH only in the computation inside blocks illustrated in Fig 2.

The first variation of PH is designed to look at whether phase information increases predictive accuracy or if the same information can be captured using SNPs. It is the closest variation of PH not using haplotypes. Block definition stays the same but inside the block, we train random forests on SNPs instead of using haplotypes. We only have one result and we compute its evidence to create a new variable in the same way as before. This variation is no longer capturing only additive heritability as it can potentially capture dominance effect. We call it PwoH for Prediction without Haplotypes.

The second variation we consider aims at capturing dominance effect. Dominance is understood in a broad sense as interaction between the two haplotypes of the same loci. Inside blocks, we train random forests on pairs of haplotypes instead of single haplotypes by concatenating the haplotype of the homologous chromosome. Each individual is thus still represented by two observations varying only by the order in which the two haplotypes appear. There are therefore twice the number of variables inside each block compared to PH. We call it PHd for Prediction with Haplotypes and dominance.

Comparison point We compared the three variations of the method to lasso regression with pre-selection.First, the N most associated SNPs in the training set were selected. Lasso regression is then fitted to the training set. The penalization parameter is selected through cross-validation in the training set. The resulting regression model is then used to predict on the validation set. The number N of preselected variables is a hyper-parameter.

Implementation details The source code is a mix of bash, R and python scripts, uses Plink (Purcell et al., 2007) and Shapeit 2 (Delaneau et al., 2013) and is available on github at https://github.com/FelBalazard/Prediction-with-Haplotypes. The glmnet R package was used for lasso regression (Friedman et al., 2010). The python machine learning package scikit-learn was used for random forests (Pedregosa et al., 2011).

Datasets and protocol

We tested our method on the GWAS datasets made available by the WTCCC and first described in (Burton et al., 2007). The WTCCC data collection contains 17000 genotypes, composed of 3000 shared controls

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

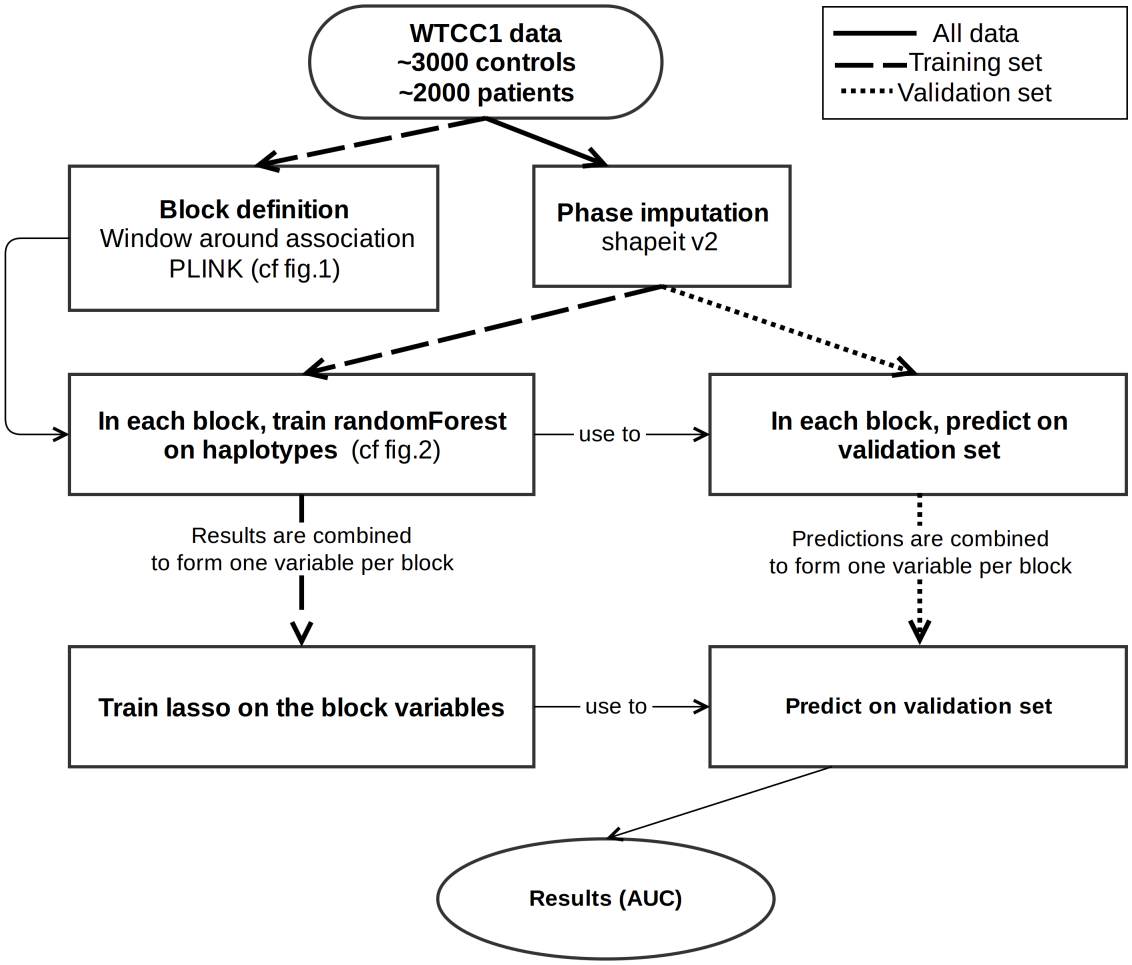

Figure 3. Pipeline of the method The different kinds of line indicate the separation between training set and validation set.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

and 2000 cases for each of 7 complex diseases: bipolar disorder (BD), Crohn's disease (CD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). Individuals were genotyped with the Affymetrix GeneChip 500K Mapping Array Set and are described by about 500,000 SNPs (before the application of quality control filters).

Quality control (QC) is important for GWAS datasets. Corrupt variables can allow for almost perfect discrimination while not respecting Hardy-Weinberg Equilibrium (HWE) (Botta et al., 2014). We first excluded the exclusion lists for individuals and SNPs used in (Burton et al., 2007) and provided with the data. Then, for each disease, an exclusion list was defined for SNPs that were missing in more than 5% of the individuals (patients and controls), that had a minor allele frequency smaller than 0.1% or that had a p-value for HWE smaller than 10^{-6} for controls or smaller than 10^{-10} for patients.

With Shapeit 2, phasing accuracy increases with sample size (Delaneau et al., 2013). To achieve maximum accuracy, we phased all the 17000 patients and controls together excluding only the intersection of all disease specific exclusion lists for SNPs. We then used the disease specific exclusion list to obtain each phased disease dataset with proper exclusions.

The predictive performance of all methods were assessed by the area under the ROC curve (AUC). We performed 10-fold cross-validation and averaged the AUCs over the 10 folds. The same 10 folds are used for the different methods to limit variability.

RESULTS

In this section, we present our results on the seven WTCCC datasets. We first investigate the importance of two hyperparameters on the CD dataset. We then use parameters that obtained good performance on the CD dataset to evaluate predictive performance and influence of window size on the 7 datasets.

Influence of the hyperparameters

Concerning lasso regression with preselection, we had one hyperparameter to select: the number N of pre-selected SNPs. We tried the values 500, 1000 and 1500 on all diseases. The best result for all diseases except BD was obtained for $N = 500$. We use the values obtained for $N = 500$ in the following. The results are available in the supplementary material.

On the CD dataset, we studied the influence of two hyperparameters of PH and its variants: the minimum number of data points in a leaf N_{leaf} and the number of blocks N_b.

For N_{leaf}, the values 1, 5, 10, 15, 25, 50, 100 were assessed. The results (available in the supplementary material) imply that the choice of $N_{leaf} = 5$, the standard value for regression, could not be improved upon notably by another choice of value for this parameter. We chose $N_{leaf} = 5$ for subsequent analysis.

For N_b, we tried the values 300, 500, 700. The results (available in the supplementary material) were similar for all three values. We chose $N_b = 500$ for subsequent analysis.

Predictive performance and influence of the window size

Given the biological significance of window size L_w, we studied its influence on all diseases. The values 10kb, 20kb, 30kb, 40kb, 60kb, 80kb, 100kb and 150 kb were tried. Results are shown together with the result for pre-selection and lasso in Fig.4 and are also available in the supplementary material. When the window is too large, prediction accuracy is impaired. This is true except for T1D for which performance seems stable.

For CAD, T2D and to a lesser extent HT and RA, the best performance is obtained for intermediate values of the window size. The optimal value is 60kb for CAD, 40kb for T2D and 20kb for HT and RA.

Lasso slightly outperforms our methods on three out of seven datasets: BD, CAD and HT. This shows that our methods do not always recover all the information contained in the central SNPs. For RA and T1D, performances are very similar for all methods. However, for CD and T2D, our methods outperform lasso for most values of window size considered.

The three variants obtained similar performances. PHd does not outperform PH, it fails to capture any dominance effect. PH does not consistently outperforms PwoH. The decrease in performance with increasing window size is true for all three variants. However, for large window sizes, PH outperforms or equals the other variants.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

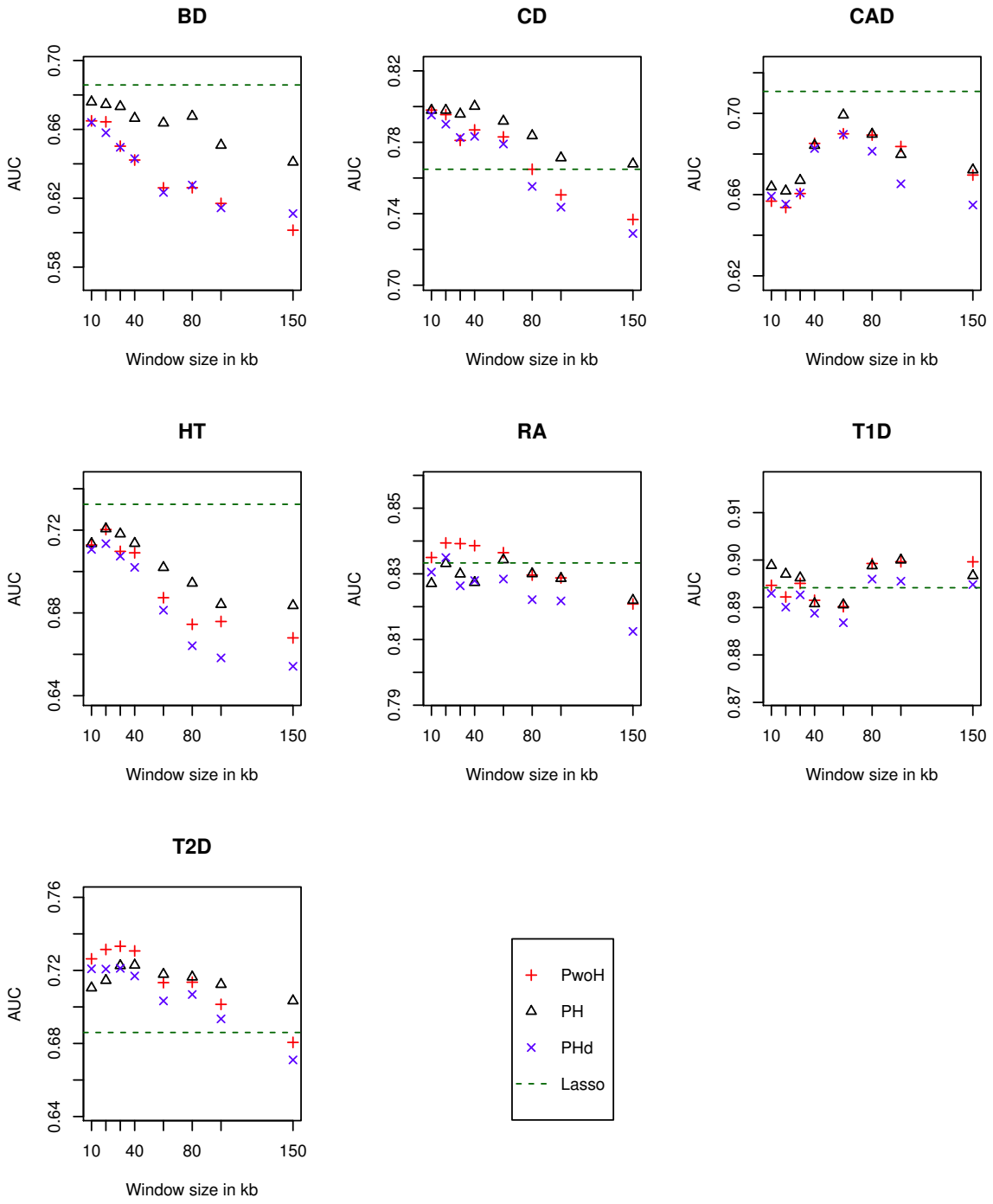

Figure 4. Predictive performance and influence of the window size Results for PwoH, PH and PHd on all diseases are displayed depending on the window size. The green line shows the performance obtained by preselection with lasso. All AUCs are obtained on the same folds.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

DISCUSSION

In this work, we developed a method to try and capture interactions inside haplotypes. This implies a different setting than is customary in machine learning. Variables have two values for each observation and there is a metric structure to take into account. The design of PH allows it to take that structure into account.

PH outperformed standard lasso regression on two datasets but not on all of them. This is suggestive of haplotypic effects in Crohn's disease and type 2 diabetes. The result for CD is reminiscent of the results of (Kang et al., 2011) as well as the two new loci discovered in the CD dataset by (Botta et al., 2014). On the other hand, in three datasets, PH was slightly less accurate than lasso regression. This might be due to the multi-step design which is not a standard approach and that results in some loss of information.

Our more theoretical contribution is to note that interactions inside haplotypes are a part of additive heritability. Our results therefore show that some of the missing heritability can be explained by the lack of consideration for interaction inside haplotypes. For type 2 diabetes, the proportion of genetic variance explained went from 40% to 66% (for prevalence $K = 20\%$ and heritability $h^2 = 0.30$) (Wray et al., 2010). For Crohn's disease, this proportion went from 16% to 22% (for $K = 1\%$ and $h^2 = 0.8$).

These estimates of explained heritability and all of the above AUCs are optimistic due to various study specific quality problems that result in overestimation of predictive performance as shown in the drop in out-of-study performance in (Wei et al., 2009). The limited availability of comparable datasets is therefore a hindrance to progress in this area of research.

PwoH obtains similar performance than PH on all datasets. This means that even if there are haplotypic effects, it is not necessary to perform phase imputation to capture them. It can be sufficient to capture local interactions using genotype to recover haplotypic effects. PHd did not outperform PH. This suggests that dominance effects are not an important part of genetic risk for complex diseases.

Small window sizes obtained the best performances while larger window sizes led to decreased performance. This is consistent with the results in (Botta et al., 2014). This shows that the information that we recover is very local.

Further work on the importance of phase information for the prediction of complex diseases could try and adapt the method in (Botta et al., 2014) so that it can use phase information.

ACKNOWLEDGMENTS

The author thanks Gérard Biau and Pierre Bougnères for their supervision and proteiform help in this work. He also thanks Melanie Hayr for the idea of PwoH at a conference. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under awards 076113 and 085475. The author thanks the WTCCC for its commitment to provide data to researchers.

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., Tye-Din, J., Bhalala, O., 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., Louppe, G., Geurts, P., and Wehenkel, L. (2014). Exploiting snp correlations within random forest for genome-wide association studies. *PloS one*, 9(4).

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Burton, P., Clayton, D., Cardon, L., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D., McCarthy, M., Ouwehand, W., Samani, N., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10(1):5–6.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge university press.

Kang, J., Kugathasan, S., Georges, M., Zhao, H., and Cho, J. (2011). Improved risk prediction for crohn's disease with a multi-locus approach. *Human molecular genetics*, 20(12):2435–2442.

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., McCarthy, M. I., Ramos, E., Cardon, L., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., De Bakker, P., Daly, M., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E., and Schork, N. (2011). The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

Wei, Z., Wang, K., Qu, H., 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.

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P., et al. (2013). Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *The American Journal of Human Genetics*, 92(6):1008–1012.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2074v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016