# Haplotype-based genetic risk estimation of complex diseases

Félix Balazard[1] [2], Gérard Biau[1], Pierre Bougnères[2]

[1] : Laboratoire de Statistique Théorique et Appliquée, UPMC [2] : Unité 1169 INSERM

## Motivation

Personalized prediction should use all the genetic data available to predict each person's risk of developing a complex disease.

Only one[1] earlier attempts to do this took into account phase information: the information that neighboring mutations are on the same chromosome or on the homologous chromosome. This information is biologically important. Here, we propose a method PH (Prediction with haplotypes) to use it to improve genetic risk estimation.

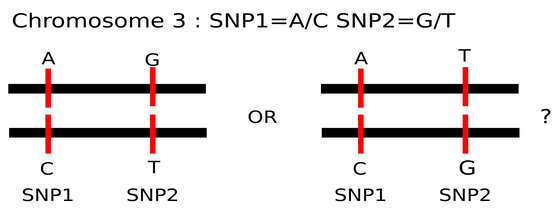



Figure 1 : Definition of phase information

## Method

Phasing of genomic data can be done probabilistically using Shapeit[2]. This method has reasonable accuracy but it means that long distance haplotypes are not reliable. It is also reasonable to think that short distance haplotypes are more likely to have functional consequences. We therefore limited our approach to use short haplotypes.

We took the strongest association signals and defined blocks of a given length around them as shown in fig.2.

Inside each block, we trained randomForest on the haplotypes. Each haplotype is taken as a different observation even if it comes from the same individual. The results are then combined to define a new variable, one for each block. Cf fig.3. This allows for parallel computation.

These new variables are then used to train a lasso regression.

This method captures interactions in *cis*. It keeps the interpretability of a sparse method like lasso. It is summarized in the pipeline of figure 4.

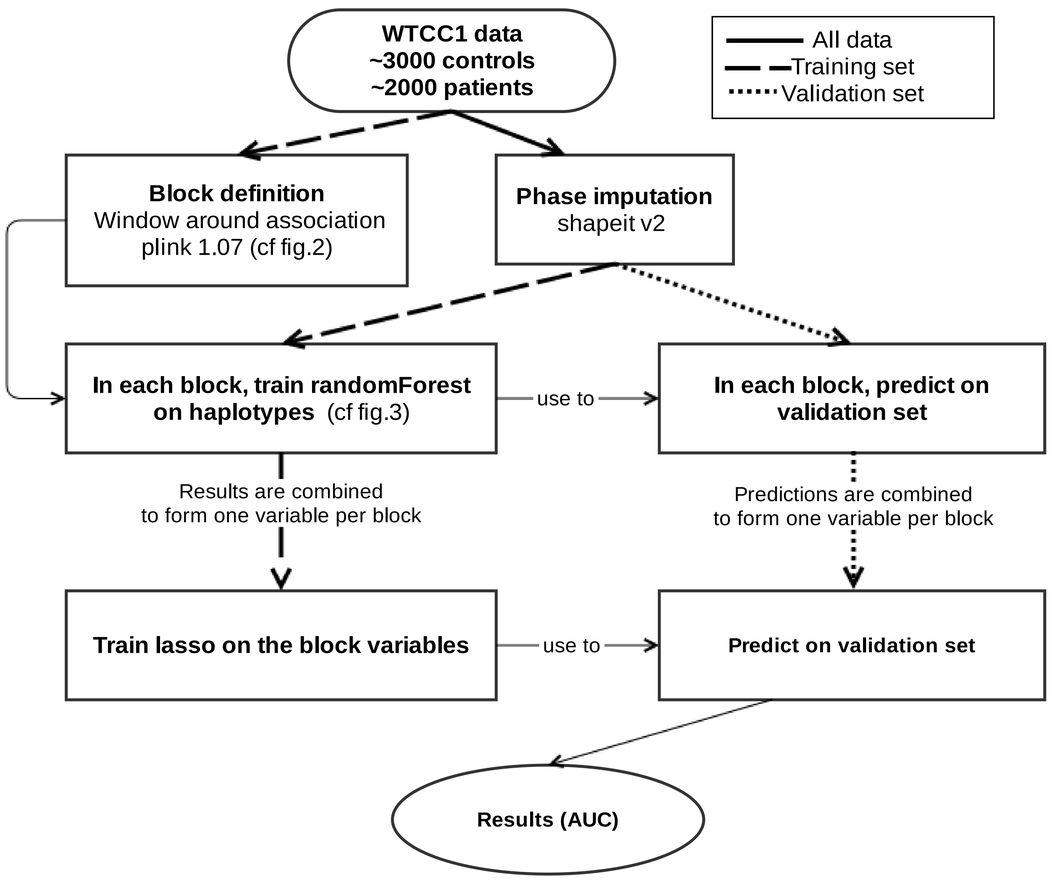



Figure 4: Bioinformatics pipeline of the method

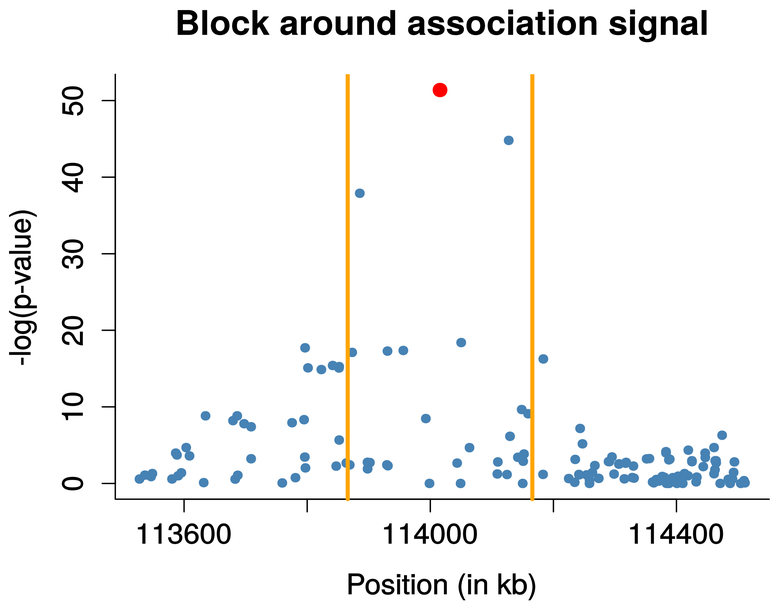



Figure 2: Block around association signal
Here, a block of 300 kb around a peak for T1D on chromosome 1

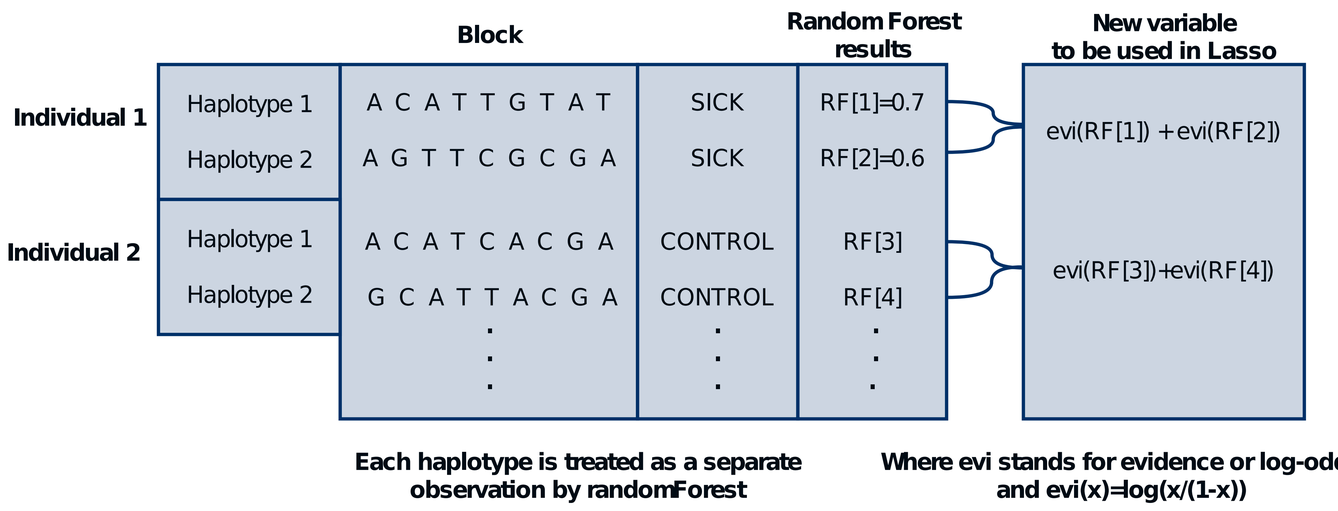
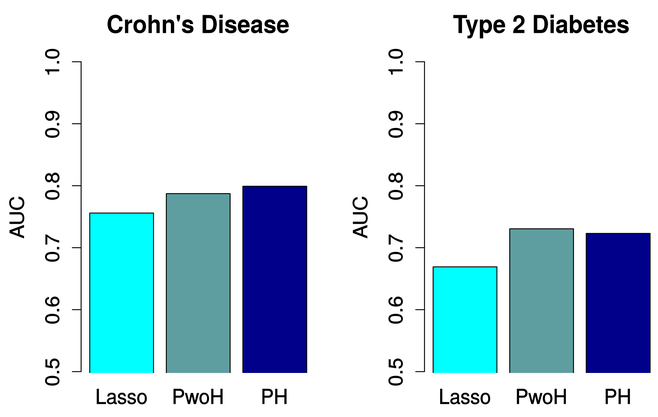



Figure 3 : Data analysis in each block using randomForest

## Results

We used the WTCCC1 data on seven diseases to test our method. We compared our method with standard preselection and lasso regression as well as a variant of our method where we do not use haplotypes (PwoH). Results were similar for all three methods on 5 diseases. For Crohn's disease and type 2 diabetes, we saw an increase in performance for PH and PwoH. This shows there are interactions in cis but that they can be captured without phase information.




## References

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

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

[3] : Wei, Z., Wang, K., Qu, H. Q., Zhang, H., Bradfield, J., Kim, C., ... & Hakonarson, H. (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS genetics, 5(10), e1000678.

Blog : felixbalazard.wordpress.com