Bibliographic report

Florian Privé

September 21, 2017

Introduction

The main objective of my thesis is to make Polygenic Risk Scores (PRS) that can differentiate an healthy person (control) from a diseased person (case) to be used for precision medecine. These PRSs consists in a combination of information on DNA mutations at multiple loci of the genome, typically from hundreds of thousands to dozens of millions. PRSs have been used for other goals such as finding a common genetic contribution between two diseases (S. M. Purcell et al. 2009). Yet, only the prediction aspect will be treated in my thesis because I find this question more useful.

There are three main concerns when constructing a PRS. The first one is that the scores have to be constructed while taking care of confounding effects. The second one, which is partially related to the first one and which has been on particular interest recently, is that these PRS can be used on a global basis, i.e. not only for the population they were train on. Finally, the third concern that need to be overcome is the size of the datasets. They can require several gigabytes of memory or even terabytes for the largest datasets, e.g. the UK Biobank (Bycroft et al. 2017).

State-of-the-art for Polygenic Risk Scores

Since 2007, genome-wide association studies (GWAS) have multiplied. The goal of these studies is to find loci which variation is associated with a trait of interest, e.g. a status of disease. In a GWAS, each locus is tested independently. Then, researchers tried to find a way to combine all the GWAS results, i.e. the size and significance of the effects of all loci, in a predictive score.

For computing PRSs for human diseases, what is widely used is the Pruning + Thresholding (P+T) model (Chatterjee et al. 2013; Dudbridge 2013; Golan and Rosset 2014). Under the P+T model, a coefficient of regression is learned independently for each locus along with a corresponding p-value (the GWAS part). The loci are first clumped (P) so that there remains only loci that are weakly correlated with each other. Thresholding (T) consists in removing loci that are under a certain level of significance (P-value threshold to be determined). A polygenic risk score is defined as the sum of allele counts of the remaining SNPs weighted by the corresponding regression coefficients. As the weights are learned independently, this model can be applied to very large dataset, which is also why this is widely used.

All steps of the P+T model

The steps required in the P+T model are described in figure 1.

For animal and vegetal species, knowing a “breeding value” is often of interest, which is a continous trait. Moreover, in these studies, the number of samples usually don’t exceed a few thousands, which are manageable datasets. The preferred model for predicting a “breeding value” is the genome Best Linear Unbiased Prediction (gBLUP).

State-of-the-art for GWAS

For computing a GWAS, the PLINK software (Chang et al. 2015; S. Purcell et al. 2007) is often used because it is fast and memory-efficient. It also provides some tools for conversion and quality control. Yet, before

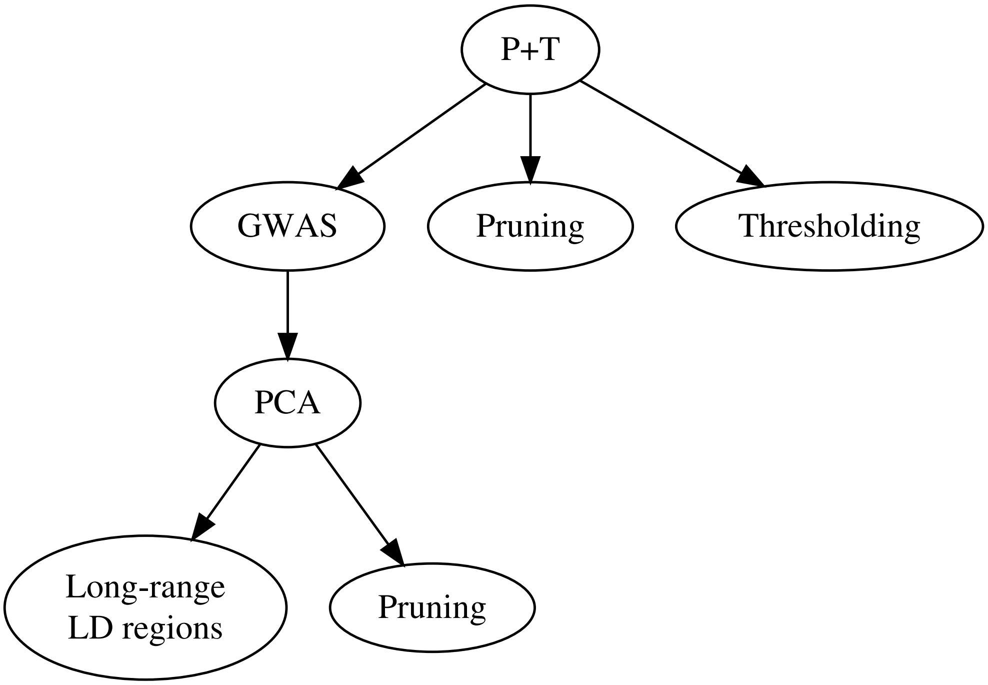

Figure 1: All steps required in the P+T model.

References

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, et al. 2017. "Genome-wide genetic data on ~500,000 UK Biobank participants." *Doi.org*, July. Cold Spring Harbor Laboratory, 166298. doi:10.1101/166298.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-generation PLINK: rising to the challenge of larger and richer datasets." *GigaScience* 4 (1): 7. doi:10.1186/s13742-015-0047-8.

Chatterjee, Nilanjan, Bill Wheeler, Joshua Sampson, Patricia Hartge, Stephen J Chanock, and Ju-Hyun Park. 2013. "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies." *Nature Genetics* 45 (4): 400–405, 405e1–3. doi:10.1038/ng.2579.

Dudbridge, Frank. 2013. "Power and Predictive Accuracy of Polygenic Risk Scores." *PLoS Genetics* 9 (3). doi:10.1371/journal.pgen.1003348.

Golan, David, and Saharon Rosset. 2014. "Effective genetic-risk prediction using mixed models." *American Journal of Human Genetics* 95 (4). The American Society of Human Genetics: 383–93. doi:10.1016/j.ajhg.2014.09.007.

Purcell, Shaun M, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, Pamela Sklar, et al. 2009. "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." *Nature* 10 (AuGuST). Nature Publishing Group: 8192. doi:10.1038/nature08185.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *American Journal of Human Genetics* 81 (3). Elsevier: 559–75. doi:10.1086/519795.