UVIC
UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL
DE CATALUNYA

Master of Science in Omics Data Analysis

Master Thesis

Polygenic Risk Score in complex diseases

by

Laureano Tomás Daza

Supervisor: Juan R González, Barcelona Institute for Global Health, ISGlobal

Academic Tutor: M. Luz Calle, University of Vic, UVic-UCC

Department of Systems Biology

University of Vic – Central University of Catalonia

September 2018

Table of Contents

Polygenic Risk Score in complex diseases

Laureano Tomás-Daza

Barcelona Institute for Global Health, ISGlobal and University of Vic, UVic-UCC

Received on September 17, 2018; revised on September 18-27, 2018; accepted on September 28, 2018

ABSTRACT

Motivation: Plenty genome-wide datasets are produced from complex diseases by traditional GWAS studies, but they are limited. A new approach has emerged in the last decade, the Polygenic Risk Scores (PRS), to combine several SNP into a single predictor to try to explain the complex genetic behind diseases like Asthma or Autism Spectrum Disorders.

Results: Here we analyse genome-wide data from these two diseases a compute PRS with three different approaches, PLINK's method, a machine learning approach (biglasso) and a targeted-based method using SFARI database. We find that this kind of analysis are quite complex like the diseases they try to predict, and PRS only explain a very low percentage of the variance of the disease. The validation analysis we performed show us that the parameters used to compute the PRS have to be optimize using bigger datasets. We also used a machine learning approach (XGBoost) to impute the data in certain analysis.

Reproducibility: github.com/isglobal-brge/master_thesis/tree/master/genetic_score

Contact: laureano.tomas@uvic.cat

1. INTRODUCTION

A polygenic risk score (PRS) is an estimate of the cumulative contribution of genetic factors to a specific outcome of interest in an individual that takes into account the reported risk alleles[1]. Another more detailed definition could be a weighted sum of the number of risk alleles carried by an individual, where the risk alleles and their weights are defined by the loci and their measured effects as detected by genome wide association studies[2]. The idea behind of this concept came up as an attempt of extract hidden information from GWAS data which have become routine over the past decade. With this idea scientists tried to explain a considerable proportion of phenotypic variation by assembling markers not achieving significance[3].

The main two purposes of PRS are: (1) predict the probability of an individual developing a disease based on some amount of available information, usually genetic and (2) to estimate the level of predictive power or variability explained that is captures by associated variants. The goal is to be able to predict if an outcome of interest, i.e. a disease, a reaction to a drug, etc., could appear in a person based on genetic data[1].

PRS were developed mainly to investigated deeper the genetic component of complex traits, that could not be explained by a few single nucleotide polymorphisms (SNP). Some of these complex traits are diseases as asthma and autism spectrum disorder (ASD). These diseases are highly heritable and recent GWAS have found that a portion of this heritability is attributable to common genetic variants[4].

Asthma a is a chronic disease of the airways defined by its symptoms, which include reversible airflow obstruction, inflammation, and bronchial hyperresponsiveness, which make it clinically heterogeneous. Also, asthma has a strong evidence of heritability as we mention above, but progress in defining its genetics however has been slow and hampered by issues of inconsistency. Recent advances in tools available for analysis, such as PRS, have substantially altered the landscape[5].

ASD is characterized by impairments in social interaction and stereotyped behaviours. For most individuals with ASD, the causes of the disorder remain unknown; however, in up to 25% of cases, a genetic cause can be identified. For those complex diseases whose genetic cause remains unknown, PRS could be a promising approach to decipher their genetic causes[6].

Due to all this unknown genetic information of complex diseases calculating PRS is now a common approach, but their potential complications and pitfalls are also emphasized[1]. There are several approaches to select SNPs used to create PRS. One possibility is to select SNPs that are significant at single level (i.e. from GWAS). However, this approach does not consider possible interactions among them. Another possibility is to build a multivariate model (i.e. using machine learning algorithms) and consider those SNPs

that are selected in that model. Finally, *a priori* knowledge can also be used to filter those genes with an impact in the disease of interest. In this study, we aim to evaluate those different approaches used to build PRS by analysing asthma and ASD cases.

2. MATERIAL AND METHODS

For a better understanding of this section we recommend following visually the workflow represented in **Supplementary Figure 1**.

2.1. Data

In this study we used four datasets (**Table 1**), two from Asthma and two from ASD. For each disease one dataset was used to compute association and PRS analysis, and the other one to validate the analysis. All datasets came from public data of different consortiums. And only autosomal chromosomes were selected.

2.2. Quality Control

The five quality-control (QC) steps[7,8] consist of filtering out of SNPs and individuals based on the following: (1) **individual and SNP missingness**, (2) **minor allele frequency (MAF)**, (3) **deviations from Hardy–Weinberg equilibrium (HWE)**, (4) **heterozygosity rate** and (5) **cryptic relatedness**. This steps are performed using the free software PLINK[9].

2.2.1. Step I: individual and SNP missingness

Missingness can lead to false associations if it is non-random with respect to phenotypes or genotypes. SNP missingness is the complement to individual missingness and is correlated with SNP quality from the original genotyping assay. Missingness is investigated using PLINK '*--missing*'. We removed markers and individuals by using the parameters '*--geno*' and '--

Table 1. Description of the datasets used in the study. * 10 ambiguous sex individuals

	Controls	*Cases*	*Males*	*Females*	*Array*	*Database*
*Asthma_I**	1480	1000	1296	1174	Illumina Human610 -Quadv1	dbGAP (phs000490.ecv1.p1)
Asthma_II	1587	725	1097	1215	Affymetrix Human Mapping 500K GeneChip	ECRHS[7,8]
ASD_I	5228	2652	4901	2979	ILLUMINA_Human_ 1M	AGP (phs000267.v1.p1)
*ASD_II**	1480	515	965	1020	Illumina Human610 -Quadv1	eMERGE (phs000360.v3.p1)

mind' respectively. This step is performed twice with two different thresholds 0.2 and 0.02 to increase the accuracy of this step.

2.2.2. Step II: minor allele frequency (MAF)

Minor allele frequency filtering is important because rare genotypes will not show up as often and thus will have less evidence in a GWAS and the calls will be less certain, and it is also difficult to detect associations with them. In this study we used a conventional threshold of 0.05.

2.2.3. Step III: deviations from Hardy–Weinberg equilibrium (HWE)

Markers out of HWE can indicate that there were genotyping errors. However, a strong association signal can also result in deviations from HWE. Normally, only variants from control samples are checked for deviations from HWE with a p-value of $1x10^{-6}$, but we also checked cases samples but with a less stringent threshold ($1x10^{-10}$).

2.2.4. Step IV: heterozygosity rate

Individuals resulting from random mating within a population should have predictable heterozygosity (H) values. H is a measure of the number of loci in an individual that are heterozygous. Departure from expected H values can signify DNA quality issues (high H) or samples from a different population (low H). We computed the heterozygosity rate as: (Number of non-missing autosomal genotypes)-(Observed number of homozygotes)/(Number of non-missing autosomal genotypes) and we remove those individuals deviated more than 3 standard deviations from the mean.

2.2.5. Step V: cryptic relatedness

Cryptic relatedness (CR) is when pairs of individuals are closely related and can lead to false positive or negative correlations when subjects are treated as independent.

The PLINK '*--genome*' command can estimate relatedness but is quite slow when there are a large number of markers in a dataset. Therefore, markers in high linkage disequilibrium (LD) are removed first to thin the data. This is done using PLINK '*--indep-pairwise*'.

PLINK '*--genome*' estimates relatedness of all pairs of samples and reports identify by decent in the PI_HAT column of the result file. A PI_HAT value close to 1 would indicate a

duplicate sample. The threshold 0.2 represents the half-way point between 2nd and 3rd degree relatives and is a common cut-off to use. Of each pair of related individuals, the one with the greater proportion of missing SNPs is dropped from the final dataset. This analysis for cryptic relatedness is done only on founder samples, so in the case of ASD_I dataset the step that filters by founders is skipped due to this step removed all cases, because individuals were in trio data and the children (non-founders) were the cases.

2.3. Imputation

Once the datasets have passed the QC analysis the workflow splits in two main branches, one will be imputed and will compute the PRS using a R package; and the other one will continue using PLINK to compute the association and PRS without imputation.

Imputation of SNP data could be done by several methods (IMPUTE2, imputation servers, etc.), in our study we used the method proposed and implemented in the R package '*bigsnpr*'[10]. This method is based on a machine learning approach that computes local XGBoost models and does not use phasing, allowing to reduce dramatically the computational time of the imputation process. The algorithm basically for each SNP divide the individuals in test set (with missing genotype) and train set (non-missing genotype). The train set is divided into training set and validation set. The training set is used to build the XGBoost model for predicting missing data and the validation set is used to evaluate this model providing an estimator of the accuracy of the imputation. Then this model is used on the test set to impute the missing values.

2.4. Association Analysis

The association between the genotypes and the phenotype, a binary trait, were performed by two methods implemented in PLINK and other method implemented in the R[11] package '*bigstatsr*'.

2.4.1. Logistic association – Population Stratification

All datasets were associated correcting by population stratification. This association were performed by using PLINK's parameters '*--cluster --mds-plot*' to compute the first 10 principal components and using them as covariates to compute a logistic regression between the SNPs and the case/control status. We obtained for

each SNP a p-value from t-statistic and an odd ratio that will be used as weight of the SNP.

2.4.2. Family-based association

Since ASD_I dataset was formed by parent-offspring data we had to take that into account to correct for these relations. PLINK has implemented a method for transmission disequilibrium test (TDT) to detect association using family trio design, this method used by the parameter '*--tdt*' and the output is similar from the output of the logistic regression, for each SNP a p-value from a chi-square test and an odd ratio to be used as weight.

2.4.3. Column-Wise Logistic Regression

This method is used on the imputed dataset in order to compute a PRS with the R package '*bigsnpr*'. This method is similar to the logistic method of PLINK because firstly we have to calculate the first 10 principal components by a singular value decomposition (SVD) to use them as covariates. Then the association is performed using the function '*big_univLogReg*' that estimate beta values to be used as weight for each SNP[10].

2.5. PRS methods

2.5.1. PLINK

The allelic scoring method implemented in PLINK is used by the parameter '*--score*'. This method basically computes the score as a sum across SNPs of the number of reference alleles (0,1 or 2) at that SNP multiplied by the score for that SNP. We used as score, or weight, for the SNP the logarithm of the odd ratio computed previously in the association analysis. We computed several PRS using different p-value thresholds from more stringent to less stringent.

2.5.2. Machine learning

The approach used by '*bigsnpr*' to compute the PRS need some previous data processing. First, as we mention above data must be imputed, because some functions do not work with missing values. Then, we made a first pruning of those SNP in long-range linkage disequilibrium regions. This pruned data is used to compute the SVD to estimate the covariates for the association explained in Section 2.4.3, and finally before computing the PRS itself we performed a clumping of the SNP. With all these step genotype data is prepared to be used to compute the PRS.

In this package the PRS is computed using another machine learning approach, called biglasso. This method does not use univariate summary statistics but instead train a multivariate model on all the SNPs and covariables at once, these models are very fast sparse linear and logistic regressions and they include lasso and elastic-net regularizations, which reduce the number of predictor (SNP) included in the predictive models[10]. Similar to the PLINK's method we used different p-values thresholds to compute several PRS.

2.5.3. Targeted-based

The two first methods compute the PRS without *a priori* information, only using the genotype data provided, but this last method used SNP data external from the genotype SNP, i.e. information from some database, so we do not need to compute the association. In this study we used the database SFARI to obtain autism-related genes reported in literature, then we retrieved the SNP of those genes from Ensembl and then we compute the PRS with this SNPs. In this case the PRS is computed with a method implemented in the R package '*SNPassoc*'[12] using the function '*getScore*' which calculate the PRS using the MAF as weight for each SNP. We used this method to estimate the PRS for each category of SFARI.

2.6. Predictive Models

To test if the computed PRS are predictive or not of the disease or phenotype, we calculated generalized linear models (glm) of the phenotype against the PRS, and we obtained the Nagelkerke's r^2 as measure of the goodness of fit of the model[13]. We could have chosen another estimator of the goodness of fit but there is no standard or consensus measure[14].

2.7. Validation

To validate the previously obtained models we used another dataset both of Asthma and ASD, i.e. Asthma_II and ASD_II. These datasets passed the QC and were not used in association, instead we extract the SNPs used in each PRS and the weight of each SNP from the previous association analysis in order to replicate them in these new datasets. Once we computed the PRS with the SNPs lists used in the previous analysis we calculated again the Nagelkerke's r^2 of each model.

2.8. Reproducibility

All code used in this study for each step of the quality control, the association,

the imputation and the PRS computation are available in the next link of GitHub. (github.com/isglobal-brge/master_thesis/tree/master/genetic_score)

3. RESULTS

3.1. Quality Control

Before any analysis we performed a quality control on the different datasets, a summary of the SNP and individuals removed is shown in **Table 2**. We can see that most SNP are filtered out in due to missingness and HWE deviation, except in ASD_I that any SNP was removed due to HWE deviation. According to individuals most of the removed ones were due to heterozygosity rate, and in Asthma_II and ASD_I a considerable number of individuals were removed due to missingness.

Thanks to this quality control in all datasets we removed the majority of missing phenotypes and other important considerations like individuals related to each other which would alter the results. In average after these five steps we increased the genotyping rate from 0.90-0.95 to 0.99 in all datasets facilitating the imputation step.

3.2. Association

After the quality control we performed the typical association analysis for the Asthma_I and ASD_I datasets. For the Asthma_I dataset we corrected for population stratification computing the first 10 principal components as we

Table 2. Summary of quality control removed SNP and individuals.

	Asthma_I		*Asthma_II*		*ASD_I*		*ASD_II*	
Missingness SNP	31777		16317		136318		31416	
Missingness ind.	5		136		207		0	
MAF	34853		56406		0		33569	
HWE	20138		541		0		47315	
Het. Rate	63		28		246		49	
Cryptic Related	45		17		78		36	
	SNP	**Ind.**	**SNP**	**Ind.**	**SNP**	**Ind.**	**SNP**	**Ind.**
Total Removed	116768	113	73264	181	136318	531	112300	85

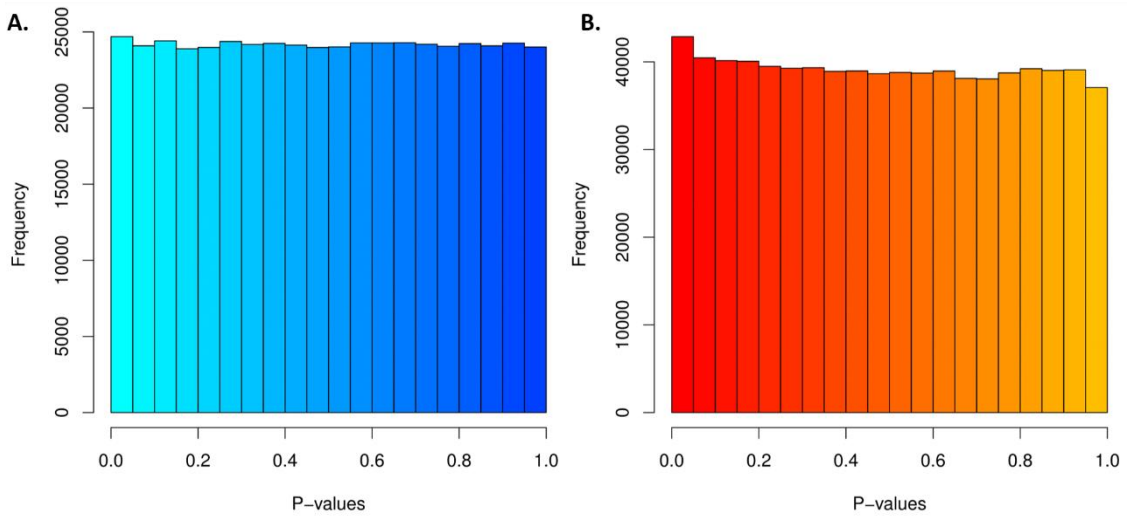


Figure 1. Histogram of P-values distribution from association analysis. A: Distribution of P-values from Asthma_I dataset. B: Distribution of P-values from ASD_I dataset.

described in Section 2.4.1, and we can see in **Figure 1A** how P-values from association are uniformly distributed, indicating that no many SNPs are significantly associated with the disease after multiple testing correction.

For ASD_I dataset we performed the association analysis taken into account that individuals were in trio-data, so we did a family-based association and the P-values distribution is shown in **Figure 1B** and it follows a uniform distribution as expected. However, in that case, we observe a little deviation in the P-values close to 0, indicating that there are some SNP which do not follow the null hypothesis, and hence, could be associated with the disease.

3.3. PRS

In this section we explained the results from the different methods of computing the PRS.

3.3.1. PLINK method

Once we had the association analysis for both diseases we took the P-values to be used as thresholds for group the SNP and we used the odd-ratio to compute its logarithm to be used as betas for each SNP. We can see in both barplots **Figure 2** the Nagelkerke's R^2 from the generalized linear models of each PRS, in both datasets we observed the same behaviour, the softer the P-value threshold is the greater the R^2 value is, i.e. the more SNP are taken into account to compute the PRS the be tter is the R^2 value. We could think this behaviour was due to the fact that when we include more predictors in a

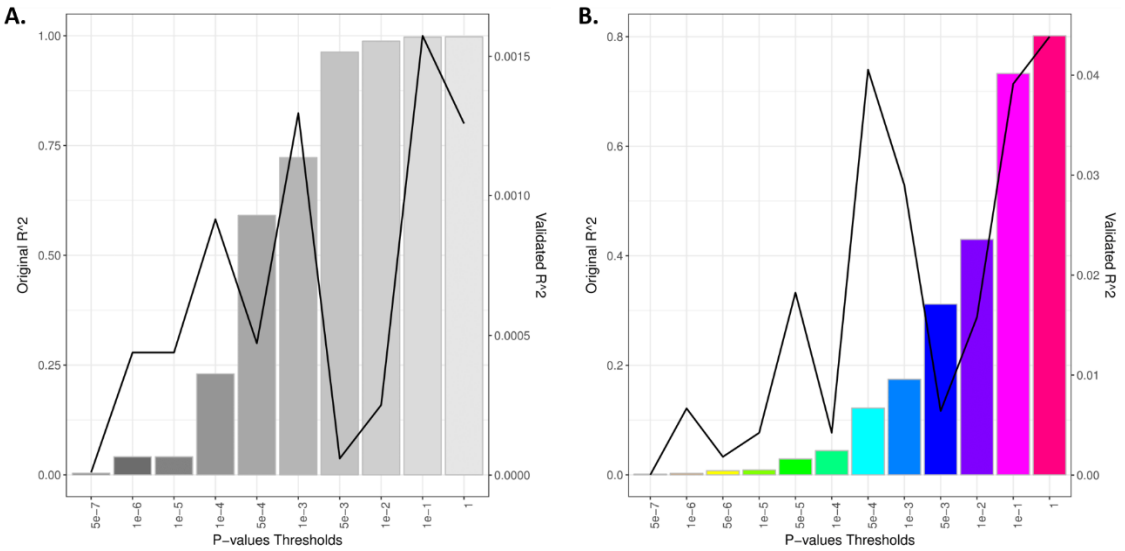


Figure 2. Barplots of Nagelkerke's R^2 and validation line computed by PLINK method. Both barplots represent the R^2 (y-axis) obtained from the glm of the PRS computed with different p-values thresholds (x-axis) by PLINK and the R^2 from the validation dataset represented as a black line over the barplot (secondary y-axis). A: Asthma datasets; B: ASD datasets.

model the R^2 increased, but in this case we did not do that, we took more predictors to compute the PRS not the generalized linear model.

In order to prove if this R^2 values where really that good for both diseases we computed the same PRS from a validation dataset (Asthma_II and ASD_II). We selected the SNP used to compute each PRS from the original datasets and used these lists of SNP and the betas computed in the association analysis of Asthma_I and ASD_II and we computed the PRS in the validation datasets. The R^2 from the validation is represented as a black line over the barplots in **Figure 2**. The first notorious fact is the drastic change in the scale, the validation R^2 only reached 0.0015 and 0.04 in Asthma and ASD respectively against the 1 and 0.8 values from original datasets. The second notorious difference is the shape of the validation line, is not the same that the shape of the barplots, in both cases the R^2 values tend to increase but at some thresholds the value decreases drastically, and it reached its maximum when almost all SNP are grouped to compute the PRS.

3.3.2. Machine learning method

The second method used was a machine learning approach using the R package *bigsnpr* on the imputed data. In this case the association is performed by the package and both P-values and betas are computed by itself to be used in the PRS.

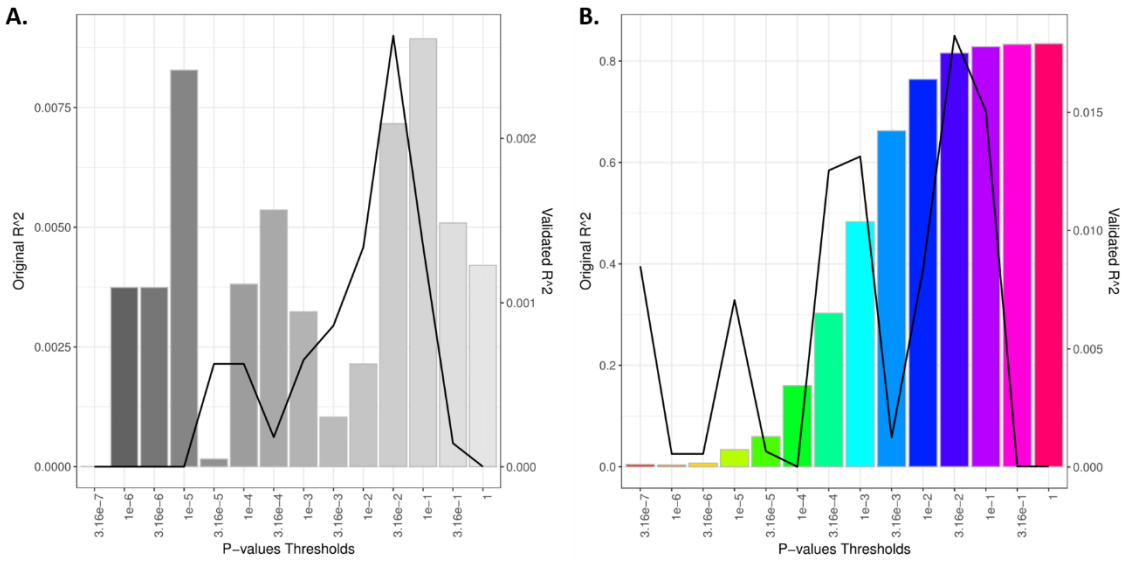


Figure 3. Barplots of Nagelkerke's R^2 and validation line computed by *Bigsnpr* method. Both barplots represent the R^2 (y-axis) obtained from the glm of the PRS computed with different p-values thresholds (x-axis) by R package *bigsnpr* and the R^2 from the validation dataset represented as a black line over the barplot (secondary y-axis). A: Asthma datasets; B: ASD datasets.

We established different P-values thresholds, the same for both datasets, to compute several PRS from less predictors to more predictors. As we explained in the first method the barplots shown in **Figure 3** represent the Nagelkerke's R^2 value from the predictive models. With this method the barplots are not similar between both datasets, in ASD_I it is quite similar to the one generated by PLINK's method because it tends to increase the R^2 values as the threshold is softer and the maximum is 0.8. But in Asthma_I the barplot is quite irregular, it shows some increases and decreases over the barplot. According to the validation lines in ASD the difference from the original data is huge as it was in PLINK's method, but in Asthma the difference is not so much. The shape of the line in both datasets is quite different from the original data, we could say that in both cases the line tends to increase the value of R^2 although there are some decreases. In Asthma the maximum of the barplot and the validation lines is reached almost in the same P-values threshold, whereas in ASD the maximum of both barplot and line is near but the maximum of the barplot corresponds to the minimum of the validation line.

3.3.3. Targeted-based method

The last method that we tried was based on *a priori* selection of SNP from the SFARI database. The first step in this method was to select a pruning threshold to apply to the dataset. In

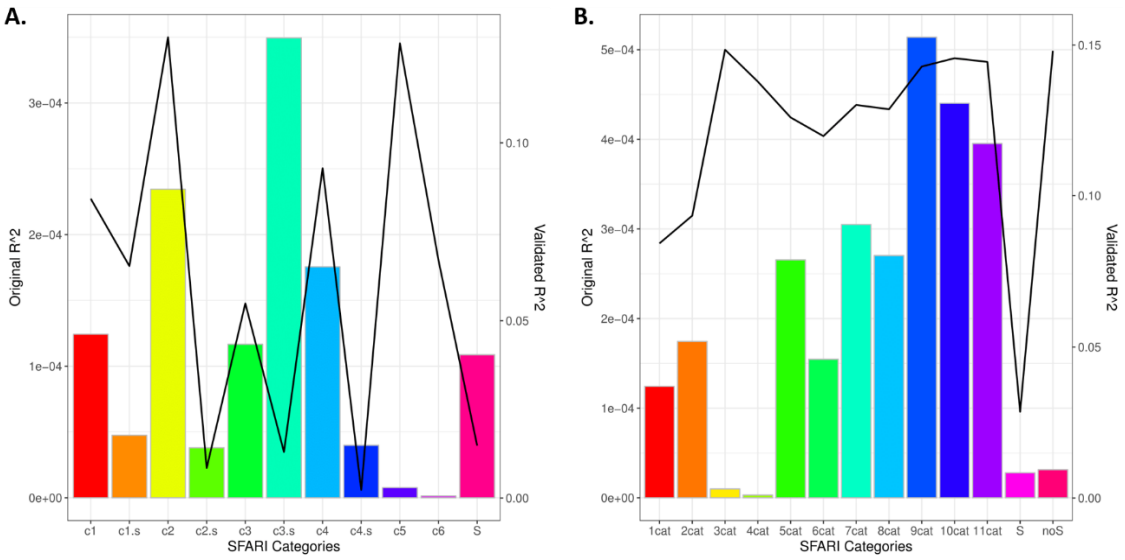


Figure 4. Barplots of Nagelkerke's R^2 and validation line computed by SNPassoc method based on SFARI categories of ASD dataset. A: Barplot of R^2 (y-axis) obtained from the glm of the PRS computed for the different SFARI categories (x-axis) by R package *SNPassoc* and the R^2 from the validation dataset represented as a black line over the barplot (secondary y-axis). B: Barplot of R^2 (y-axis) obtained from the glm of the PRS computed for increasing groups of SFARI categories and syndromic categories (S) together and non-syndromic categories (noS) (x-axis) by R package *SNPassoc* and the R^2 from the validation dataset represented as a black line over the barplot (secondary y-axis).

Supplementary Figure 2 it is shown the R^2 for each SFARI category at different pruning thresholds for almost all categories the best threshold was none, so we selected 0 as pruning threshold for further analysis.

We conducted the PRS analysis in two ways, first we computed the PRS using the SNP in each SFARI categories (**Figure 4A**), and then we tried to see if grouping the categories (**Figure 4B**) increased the R^2 of the PRS. In the barplot from each SFARI category we can see that the best category was C3.s, however in the validation line the best categories were C2 and C5, even with a better R^2 value than the original data. We would like to mention that the shape of the line in the first five categories follows the same behaviour as in the barplot.

On the other hand, we expected that when we added more categories the R^2 should be higher and that happens almost every time we add a new category, except in the case of adding C2 and C2.s mainly. Furthermore, the group of syndromic, and non-syndromic categories had a very low R^2 value. According to the validation line, in this case, like the validation of each category, the R^2 values are better than the original ones. It is curious how it tends to increase every time we added a new category and how when we added the category C2 now it reached a

maximum and not a minimum like it did in the original data. It is also interesting to point out that the non-syndromic categories have the maximum R^2 value.

4. DISCUSION

This study shows us how different the three methods of computing PRS are, and how heterogeneous are the R^2 values from the predictive model estimated from the PRS at different thresholds. Maybe we could say that the most similar prediction between original data and validation data is the one obtained from PRS computed by PLINK, but these results have to be taken into account carefully because the high values of R^2 obtained in the original datasets (0.8-1) are caused because this values in Asthma_I and ASD_I were computed using the same data that was used to estimate betas in association analysis, so these values are inflated. Another more efficient approach could be to estimate these betas from bigger databases that contains more SNP and more individual in order to compute proper betas for each SNP.

Besides, the fact that ASD_I was a family-based dataset could interfere with the computation of betas and could not be the best option to use these betas in non-family-based dataset as it were ASD_II. And in the case of Asthma, the original dataset came from paediatric cases while the validation one came from adult patients.

The machine learning approach is the one with the results more similar between the original data and the validation data, in the asthma datasets, that could tell us that this method is quite accurate. But in the ASD datasets the different in obvious but as we mention above the reason could be the family nature of the data.

According to the SFARI results we could see that there is almost no relation between the category and the PRS, because the categories goes from those genes (SNPs in our case) with high evidence in the disease to those with less to non-evidence, and we could see that some categories with less evidence have a greater R^2 which does not make sense. On the other hand, when we computed the cumulative PRS of the categories the R^2 increases as expected, this may be caused because when we take into account more SNPs to compute the PRS the predictability in

higher. These discordant results could highlight the necessity of a better classification of the genes related with the diseases, because SFARI databases used its own criteria to group the related genes.

In general, to improve these low rates of predictive power, i.e. R^2 values, we could consider adding more information to the predictive models and the computations of PRS like environmental and sociodemographic data, as well as clinical history of patients. Some studies[15] used these approached and they get better results in predicting their outputs. Another consideration should be to estimate the heritability of the disease in order to estimate how much of this heritability is explained by the PRS, i.e. by grouping some SNPs[3].

And finally, as an interesting future study to explore the genetic nature of these complex diseases the GWAS data from these patients could be analysis using a pathway-networks[16] approach to try to find which biochemical pathways are the ones enriched in these diseases, and maybe with the genes significantly present in those pathways a new PRS could be computed.

5. ACKNOWLEDGEMENTS

I would like to express thanks to Juan R Gonzalez and Malu Calle for their availability and supervision of this project.

I would like to mention Natalia and Mercedes for the infinite support during all this process.

Thanks to David for everything.

And last but not least I would like to remark the moral support from friends and family. Thank you very much.

6. REFERENCES

1. Cooke Bailey, J. N. & Igo, R. P. Genetic Risk Scores. *Curr. Protoc. Hum. Genet.* **2016,** 1.29.1-1.29.9 (2016).

2. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 1–10 (2018). doi:10.1038/s41576-018-0018-x

3. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9,** e1003348 (2013).

4. Clarke, T.-K. *et al.* Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. *Mol. Psychiatry* **21,** 1–7 (2015).

5. Willis-Owen, S. A. G., Cookson, W. O. C. & Moffatt, M. F. The Genetics and Genomics of Asthma. *Annu.*

Rev. Genomics Hum. Genet. **19,** 223–246 (2018).

6. Huguet, G., Ey, E. & Bourgeron, T. The Genetic Landscapes of Autism Spectrum Disorders. *Annu. Rev. Genomics Hum. Genet.* **14,** 191–213 (2013).

7. Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27,** e1608 (2018).

8. Ellingson, S. R. & Fardo, D. W. Automated quality control for genome wide association studies. *F1000Research* **5,** 1889 (2016).

9. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–75 (2007).

10. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* (2017). doi:10.1093/bioinformatics/bty185

11. R Core Team. R: A Language and Environment for Statistical Computing. (2017).

12. González, J. R. *et al.* SNPassoc: An R package to perform whole genome association studies. *Bioinformatics* **23,** 644–645 (2007).

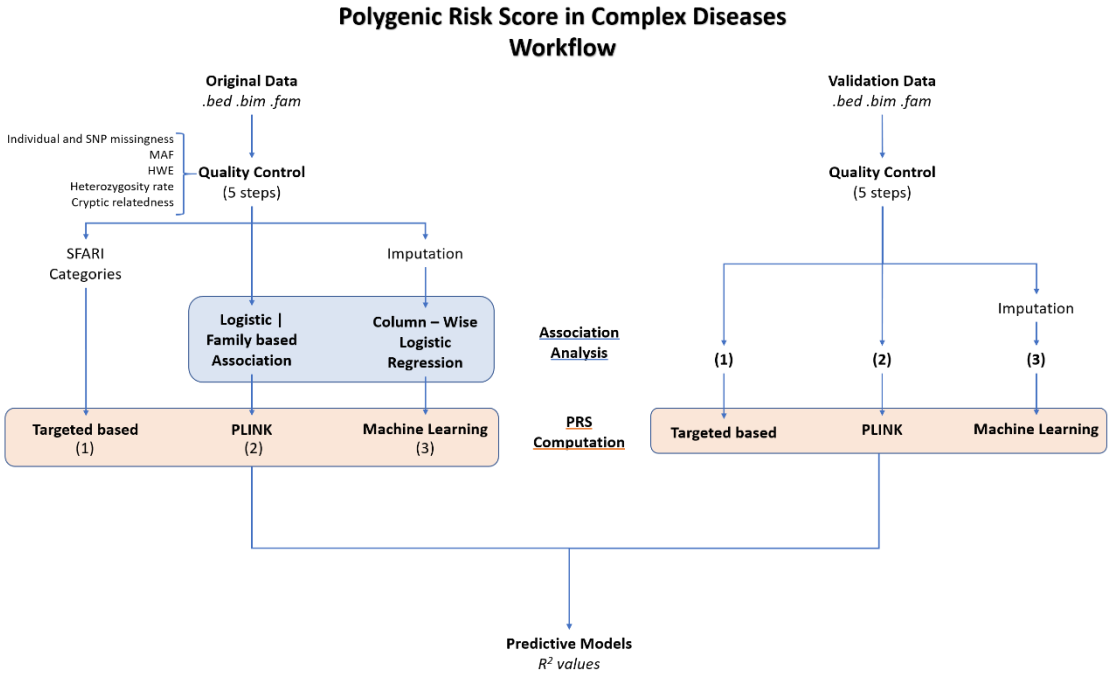
13. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31,** 1466–1468 (2015).

14. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A Better Coefficient of Determination for Genetic Profile Analysis. *Genet. Epidemiol.* **36,** 214–224 (2012).

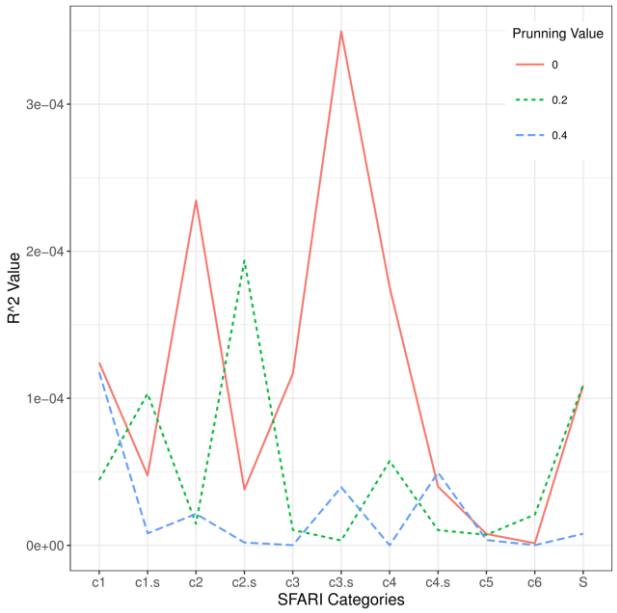
15. Pereira, A. C. On the use of genetic risk scores to predict cardiovascular disease in the general population. *Heart* **102,** 1612–1613 (2016).

16. Bakir-Gungor, B., Egemen, E. & Sezerman, O. U. PANOGA: a web server for identification of SNP-targeted pathways from genome-wide association study data. *Bioinformatics* **30,** 1287–1289 (2014).

7. SUPPLEMENTARY FIGURES



Supplementary Figure 1. Overview of the workflow used in this study. (1) It corresponds to the list of SNPs from SFARI categories. (2) It corresponds to betas from logistic/family-based association and SNP lists from each PRS computed with PLINK. (3) It corresponds to betas from column-wise logistic regression and the SNP lists from each PRS computed by machine learning.



Supplementary Figure 2. Comparison of Nagelkerke's R^2 between different thresholds of pruning by *SNPassoc* computed for the different SFARI categories.