Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement

Arrêté ministériel : 25 mai 2016

Présentée par

Florian PRIVÉ

Thèse dirigée par **Michael BLUM**, Directeur de recherche CNRS, Université Grenoble Alpes,
et co-encadrée par **Hugues ASCHARD**, chercheur Institut Pasteur

préparée au sein du **Laboratoire Techniques de L'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications.**
dans **l'École Doctorale Ingénierie pour la santé la Cognition et l'Environnement**

Score de risque génétique utilisant de l'apprentissage statistique. Genetic risk score based on statistical learning.

Thèse soutenue publiquement le **05/09/2019**,
devant le jury composé de :

Florence DEMENAIS
Directeur de recherche INSERM, Rapporteur
Julien CHIQUET
Chargé de recherche INRA, Rapporteur
Benoit LIQUET
Professeur, Université de Pau et des pays de l'Adour, Président du jury
Laurent JACOB
Chargé de recherche CNRS, Examinateur

et des membres invités suivant:

Michael BLUM
Directeur de recherche CNRS, directeur de thèse
Hugues ASCHARD
Chercheur Institut Pasteur, co-encadrant de thèse

Abstract

Genotyping is becoming cheaper, making genotype data available for millions of individuals. Moreover, imputation enables to get genotype information at millions of loci capturing most of the genetic variation in the human genome. Given such large data and the fact that many traits and diseases are heritable (e.g. 80% of the variation of height in the population can be explained by genetics), it is envisioned that predictive models based on genetic information will be part of a personalized medicine.

In my thesis work, I focused on improving predictive ability of polygenic models. Because prediction modeling is part of a larger statistical analysis of datasets, I developed tools to allow flexible exploratory analyses of large datasets, which consist in two R/C++ packages described in the first part of my thesis. Then, I developed some efficient implementation of penalized regression to build polygenic models based on hundreds of thousands of genotyped individuals. Finally, I improved the "clumping and thresholding" method, which is the most widely used polygenic method and is based on summary statistics that are widely available as compared to individual-level data.

Overall, I applied many concepts of statistical learning to genetic data. I used extreme gradient boosting for imputing genotyped variants, feature engineering to capture recessive and dominant effects in penalized regression, and parameter tuning and stacked regressions to improve polygenic prediction. Statistical learning is not widely used in human genetics and my thesis is an attempt to change that.

Résumé

Le génotypage devient de moins en moins cher, rendant les données de génotypes disponibles pour des millions d'individus. Par ailleurs, l'imputation permet d'obtenir l'information génotypique pour des millions de positions de l'ADN, capturant l'essentiel de la variation génétique du génome humain. Compte tenu de la richesse des données et du fait que de nombreux traits et maladies sont héréditaires (par exemple, la génétique peut expliquer 80% de la variation de la taille dans la population), il est envisagé d'utiliser des modèles prédictifs basés sur l'information génétique dans le cadre d'une médecine personnalisée.

Au cours de ma thèse, je me suis concentré sur l'amélioration de la capacité prédictive des modèles polygéniques. Les modèles prédictifs faisant partie d'une analyse statistique plus large des jeux de données, j'ai développé des outils permettant l'analyse exploratoire de grands jeux de données, constitués de deux packages R/C++ décrits dans la première partie de ma thèse. Ensuite, j'ai développé une implémentation efficace de la régression pénalisée pour construire des modèles polygéniques basés sur des centaines de milliers d'individus génotypés. Enfin, j'ai amélioré la méthode appelée "clumping and thresholding", qui est la méthode polygénique la plus largement utilisée et qui est basée sur des statistiques résumées plus largement accessibles par rapport aux données individuelles.

Dans l'ensemble, j'ai appliqué de nombreux concepts d'apprentissage statistique aux données génétiques. J'ai utilisé du "extreme gradient boosting" pour imputer des variants génotypés, du "feature engineering" pour capturer des effets récessifs et dominants dans une régression pénalisée, et du "parameter tuning" et des "stacked regressions" pour améliorer les modèles polygéniques prédictifs. L'apprentissage statistique n'est pour l'instant pas très utilisé en génétique humaine et ma thèse est une tentative pour changer cela.

Remerciements

Je voudrais d'abord remercier le LabEx PERSYVAL-Lab, l'université Grenoble Alpes, l'école doctorale EDISCE et le laboratoire TIMC-IMAG pour m'avoir permis de faire cette thèse. J'aimerais ensuite remercier les membres du jury, Florence, Julien, Benoit et Laurent pour avoir accepté de faire partie du jury de thèse, s'être déplacé pour ma soutenance et s'être intéressé à mon travail de thèse. J'aimerais aussi remercier les membres de mon comité de suivi de thèse, Thomas et Julien, pour avoir suivi mes travaux de thèse et assuré que je gardais un bon cap.

J'aimerais remercier également mes encadrants de thèse, Michael et Hugues, pour m'avoir accompagné lors de cette thèse. J'ai parfois été dur en négociations. Michael m'a beaucoup apporté au point de vue communication, comment écrire un papier, comment articuler une présentation, comment faire un beau tweet. Quant à Hugues, on ne s'est pas beaucoup vus mais il a pu apporter un regard neuf souvent utile. Je le remercie aussi de m'avoir mis en contact avec Bjarni avec qui j'ai passé quelques mois à Aarhus au Danemark en tant que doctorant visiteur, et où je vais continuer un postdoc pour les deux prochaines années.

J'aimerais aussi remercier tous les membres de l'équipe et du laboratoire pour leur bon humeur. Cela va de même pour l'équipe de Hugues à l'Institut Pasteur, ainsi que les chercheurs à Aarhus au Danemark. Particulièrement, Keurcien pour, entre autres, les soirées "bière, saucisson, macdo, chartreuse". J'aimerais remercier Nicolas pour sa mise à disposition et sa gestion des serveurs de l'équipe, que j'ai beaucoup utilisés pendant la dernière année de ma thèse. J'aimerais aussi remercier Magali et Michael pour m'avoir aidé à démarrer le groupe R de Grenoble. Cela fait déjà maintenant deux ans qu'il tourne, merci à tous ceux qui ont présenté et participé, et merci à Matthieu d'avoir repris le flambeau quand je suis parti au Danemark.

Enfin, merci à ma chérie, Sylvie, de m'avoir supporté tout au long de cette thèse. J'ai eu un gros coup de mou à la moitié. Aussi, les “bon j'en ai marre de travailler, je vais faire la sieste” à 15h alors qu'elle travaillait jusqu'à 18h30, ce n'était pas forcément très sympa. Et merci d'avoir accepté de partir au Danemark avec moi. Un dernier mot pour mes parents, pas forcément pour la thèse, mais pour tout mon parcours scolaire. Ils ont toujours fait en sorte que je ne manque rien, ce qui m'a permis de faire mes études dans les meilleures conditions possibles. Merci.

Contents

Chapter 1

Introduction

In my thesis work, we have been focusing on assessing someone's risk of disease based on DNA data. Except for somatic mutations, DNA data do not change over lifetime so that we could, in theory, assess someone's genetic risk of disease at birth. Thus, this could have potentially large implications in disease prevention (Mavaddat *et al.*, 2015; Pashayan *et al.*, 2015). As an example, about 12% of women in the general population will develop breast cancer sometime during their lives (DeSantis *et al.*, 2016). By contrast, a recent large study estimated that about 72% (95% CI: 65%-79%) of women who inherit a harmful BRCA1 mutation and about 69% (95% CI: 61%-77%) of women who inherit a harmful BRCA2 mutation will develop breast cancer by the age of 80 (Kuchenbaecker *et al.*, 2017). In 2013, Angelina Jolie announced that she had undergone a preventative double mastectomy, because she had a family history of breast cancer and was carrying a harmful BRCA1 mutation. Thereby, DNA data can help identify individuals who are at high-risk for some diseases in order to target preventive actions.

In this introduction, we first introduce the context of our research and the type of data we work with. Then, we present the statistical methods that are widely used in our field, and how the field has moved from association testing to prediction. Finally, we present the main statistical and computational challenges that have driven our research. During this thesis, two peer-reviewed papers have been published and a third paper is currently available as a preprint.

1.1 Context

Today, clinical risk prediction for common adult-onset diseases often relies on demographic characteristics, such as age, gender and ethnicity; health parameters and lifestyle factors, such as body mass index, smoking status, alcohol consumption and physical activity; measurement of clinical risk factors linked to disease onset, such as blood pressure levels, blood chemistries or biomarkers indicative of ongoing disease processes; ascertainment of environmental exposures, such as air pollution, heavy metals and other environmental toxins; and family history (Torkamani *et al.*, 2018). Routine genetic profiling is absent from this list, often relegated to use only when testing clarifies individual-level risks in the context of a known family history for some common adult-onset diseases (Torkamani *et al.*, 2018).

1.1.1 Different types of diseases and mutations

How mutations affect diseases depends on the effect sizes of causal variants and on the allele frequencies of these variants (Figure 1.1). For example, harmful BRCA mutations are highly penetrant mutations, i.e. that most women carrying these mutations will develop breast cancer. Many mutations with large effect sizes have been identified and are referenced in an online database called OMIM (Hamosh *et al.*, 2005). Those mutations are often very rare; either they are associated with some very rare disease or they explain only a small proportion of common diseases incidence (Anglian Breast Cancer Study Group *et al.*, 2000). In this work, we focus on common diseases (e.g. breast cancer) and try to predict individuals' disease susceptibility based on common variants; the common disease–common variant hypothesis (Pritchard and Cox, 2002). This hypothesis further suggests that such diseases are likely caused by a large number of common variants, each contributing only a small risk and thereby evading negative evolutionary selection (Salari *et al.*, 2012). Indeed, selection might be responsible for keeping genetic effects low, since variants of large effect may be selected against and eventually disappear (Pritchard and Cox, 2002). One common form of variation across human genomes is called a single nucleotide polymorphism (SNP). SNPs are single base changes in the DNA. Genotyping technologies now exists to genotype hundreds of thousands of SNPs at once for around $50 only. Starting with the Wellcome Trust Case Control Consor-

tium (2007), these genotyping technologies have led to many genome-wide association studies.

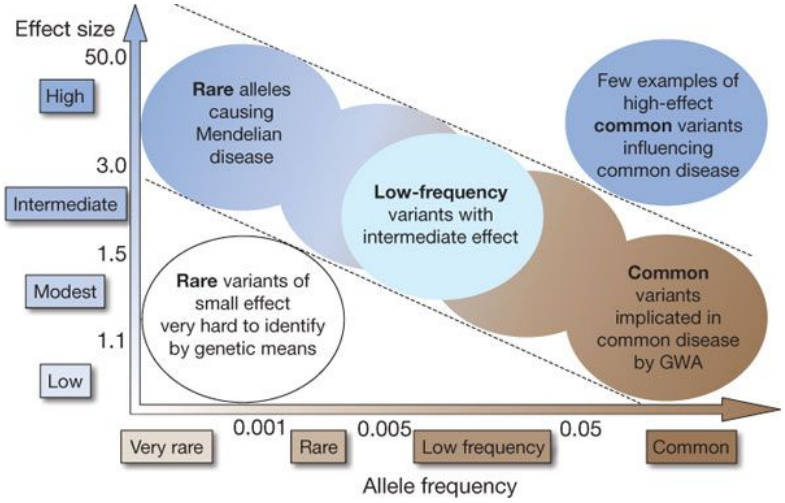


Figure 1.1: Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Source: Manolio *et al.* (2009).

1.1.2 Genome-Wide Association Studies (GWAS)

Visscher *et al.* (2017) provide a thorough review of the aims and outcomes of GWAS and Tam *et al.* (2019) talk extensively about the benefits and limitations of GWAS. The method behind GWAS is simple: test each variant one by one for association with a phenotype of interest. For a continuous phenotype (e.g. height), linear regression is used and, for each SNP j, a t-test is performed to look for an association between this SNP and the phenotype of interest ($\beta_j = 0$ vs $\beta_j \neq 0$), where

$$y = \alpha_j + \beta_j G_j + \gamma_j^{(1)} COV^{(1)} + \cdots + \gamma_j^{(K)} COV^{(K)} + \epsilon \,, \tag{1.1}$$

y is the continuous phenotype, α_j is the intercept, G_j is SNP j with effect β_j, $COV^{(1)}$, ..., $COV^{(K)}$ are K covariates with effects $\gamma_j^{(1)}$, ..., $\gamma_j^{(K)}$, including principal components and other covariates such as age and gender. Similarly, for a binary phenotype (e.g. disease status), logistic regression is used and a Z-test is performed on β_j for each SNP

j where

$$\log\left(\frac{p}{1-p}\right) = \alpha_j + \beta_j G_j + \gamma_j^{(1)} COV^{(1)} + \cdots + \gamma_j^{(K)} COV^{(K)}\,, \tag{1.2}$$

$p = \mathbb{P}(Y = 1)$ and Y denotes the binary phenotype.

It is well established that principal components of genotype data should be included as covariates in GWAS to account for the confounding effect of population structure (Price *et al.*, 2006). Indeed, principal components of genotype data capture well population structure (as shown in figure 1.2). To illustrate the importance of accounting for population structure, consider a dataset where there are 900 Finnish people and 100 Italian people. Because Finnish people are on average taller than Italian people, any SNP with a large difference in allele frequency between these two populations would be flagged as being associated with height, leading to many false positive associations. Thus, adding principal components as covariates aims at preventing those SNPs from being false positive reports.

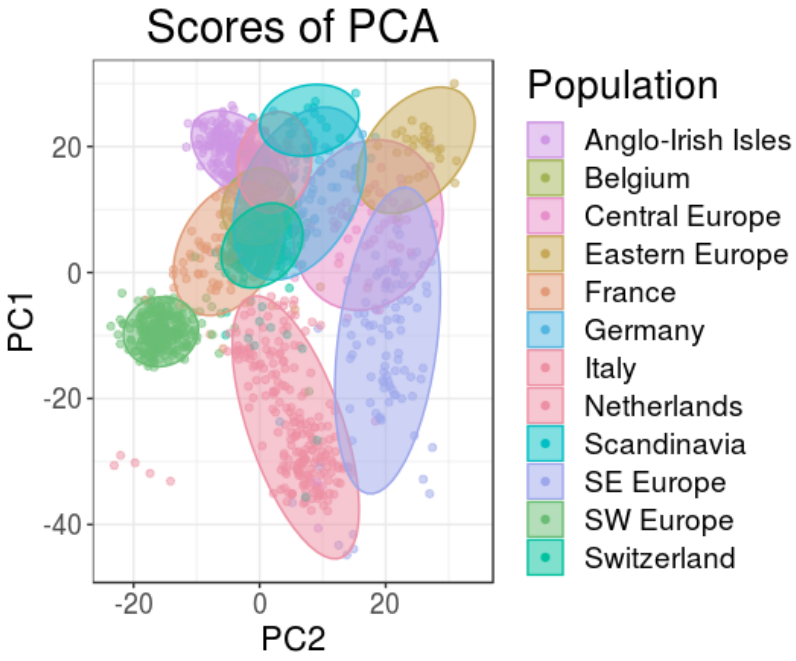


Figure 1.2: First two Principal Components of individuals from European populations using the POPRES dataset (Nelson *et al.*, 2008). PC1 correlates with latitude while PC2 correlates with longitude.

These simple tests can be used only if individuals are not related to one another. If they do, a common practice is to remove one individual from each pair of related individuals. Another strategy is to use Linear Mixed Models (LMM) to take into account both relatedness and population structure; these mixed models have also the potential to increase discovery power in association testing (Yang *et al.*, 2014).

In 2013, more than 10,000 strong associations had been reported between genetic variants and one or more complex traits (Welter *et al.*, 2013), where "strong" is defined as statistically significant at the genome-wide p-value threshold of 5×10^{-8}. This threshold corresponds to a type-I error of 5%, Bonferroni-corrected for one million independent tests (Pe'er *et al.*, 2008). Results of a GWAS are usually reported in a Manhattan plot (Figure 1.3). Manhattan plots show some association peaks (similar to skyscrapers in Manhattan) due to some local correlation between SNPs (Linkage Disequilibrium), with squared correlation roughly inversely proportional to genetic distance between SNPs (Hudson, 2001).

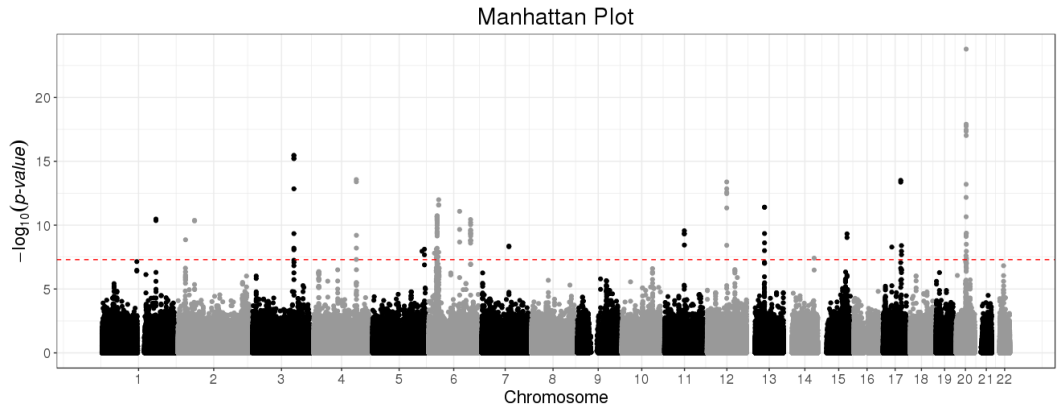


Figure 1.3: Manhattan plot from a GWAS of height based on 20,000 unrelated individuals from the UK Biobank dataset (Bycroft *et al.*, 2018).

1.1.3 GWAS data

There are mainly three types of individual-level data: genotyped SNPs from genotyping chips, imputed SNPs from reference panels, and Next Generation Sequencing (NGS) data. Genotyping chips enable a quick and cheap genotyping of 200K to 2M SNPs, mostly focusing on common variants (Minor Allele Frequency (MAF) larger than 1-

5%). Data resulting from genotyping can be coded as a matrix of 0s, 1s and 2s, counting the number of alternative alleles for each individual (row) and each genome position (column). There are usually few missing values (less than 5% in total) when using this technology.

Imputation has a different meaning in genetics than in other Data Science fields; it does not refer to filling those 5% missing values, but instead refers to adding completely new variants that were not genotyped with the chip used. This type of imputation is possible because genotypes of unobserved genetic variants can be predicted by haplotypes inferred from multiple observed SNPs (the ones that were genotyped) and haplotypes observed from a fully sequenced reference panel (Marchini and Howie, 2010; McCarthy *et al.*, 2016). Imputation now allows to have large GWAS datasets such as the UK Biobank: 90M imputed variants for each of 500K individuals who were initially genotyped at 800K SNPs only (Bycroft *et al.*, 2018).

Finally, NGS (also named Whole Genome Sequencing (WGS)) refers to fully sequenced data over more than 3M variants, including some rare variants. Yet, this technology is still very expensive, with a cost of around $1000 per genome but that could reduce to $100 in a few years[1]. GWAS to date have been based on SNP arrays designed to tag common variants in the genome. These arrays do not cover all genetic variants in the population, and it seems natural that future GWAS will be based on WGS. However, the price differential between SNP arrays and WGS is still substantial, and array technology remains more robust than sequencing (Visscher *et al.*, 2017). An in-between solution could be to use extremely low-coverage sequencing (Pasaniuc *et al.*, 2012).

Recently, some national biobank projects have emerged. For example, the UK Biobank has released to the international research community both genome-wide genotypes and rich phenotypic data on 500K individuals (Bycroft *et al.*, 2018). Yet, it is rare to have access to large individual-level genotype data. Usually, only summary statistics for a GWAS dataset are available, i.e. the estimated effect sizes and p-values for association of each variant of the dataset with a phenotype of interest (Table 1.1). Because of the availability of such data en masse, specific methods using those summary data have been developed for a wide range of applications such as imputation, polygenic prediction and heritability estimation (Pasaniuc *et al.*, 2014; Vilhjálmsson *et al.*, 2015;

[1] `https://www.bloomberg.com/news/articles/2019-02-27/a-100-genome-within-reach-illumina-ceo-asks-if-world-is-ready`

Bulik-Sullivan *et al.*, 2015; Pasaniuc and Price, 2017; Speed and Balding, 2018). The craze for such data can be explained by the fact that GWAS individual-level data cannot be easily shared publicly, as opposed to summary data (Lin and Zeng, 2010). In fact, modern large GWAS are meta-analyses of many smaller GWAS summary statistics. Moreover, methods using summary statistics data are usually fast and easy to use, making them even more appealing to researchers.

In this thesis, we have not used NGS data, but we have used genotyped SNPs, imputed SNPs and summary statistics to construct predictive models of disease risk for many common diseases.

Table 1.1: An example of summary statistics for type 2 diabetes (Scott *et al.*, 2017). Generally, effects and p-values are available for all SNPs in the GWAS, where there can be many millions of them (Editors of Nature Genetics, 2012).

Chr	Position	Allele1	Allele2	Effect	StdErr	P-value	TotalSampleSize
5	29439275	T	C	-0.000	0.015	0.990	111309
5	85928892	T	C	-0.008	0.031	0.790	111309
11	107819621	A	C	-0.110	0.200	0.590	87234
10	128341232	T	C	0.024	0.015	0.110	111309
8	66791719	A	G	0.069	0.120	0.560	99092
23	145616900	A	G	-0.011	0.060	0.860	19870
3	62707519	T	C	0.006	0.034	0.860	111308
2	80464120	T	G	0.110	0.057	0.062	108514
18	51112281	T	C	-0.011	0.016	0.490	111307
1	209652100	T	C	0.260	0.170	0.120	84836

1.2 From GWAS to Polygenic Risk Scores (PRS)

For thorough guides on how to perform PRS analyses, please refer to Wray *et al.* (2014); Chasioti *et al.* (2019); Choi *et al.* (2018).

1.2.1 The "Clumping + Thresholding" approach for computing PRS

The main method for computing Polygenic Risk Scores (PRS) is the widely used "Clumping + Thresholding" (C+T, also called "Pruning + Thresholding" in the literature) model based on univariate GWAS summary statistics as described in equations (1.1) and (1.2). Under the C+T model, a coefficient of regression is learned independently for each SNP along with a corresponding p-value (the GWAS part).

The SNPs are first clumped (C) so that there remains only SNPs that are weakly correlated with each other (S_{clumping}). Clumping looks at the most significant SNP first, computes correlation between this index SNP and nearby SNPs (within a genetic distance of e.g. 500kb) and remove all the nearby SNPs that are correlated with this index SNP beyond a particular threshold (e.g. $r^2 = 0.2$, Wray *et al.* (2014)). The clumping step aims at removing redundancy in included effects that is simply due to linkage disequilibrium (LD) between variants (see figures 1.4 and 1.5). Yet, this procedure may as well remove independently predictive variants in nearby regions.

Thresholding (T) consists in removing SNPs with a p-value larger than a p-value threshold p_T in order to reduce noise in the score. In figure 1.4, using no threshold corresponds to "C+T-all"; using the genome-wide threshold of 5×10^{-8} corresponds to "C+T-stringent". Generally, several p-value thresholds are tested to maximize prediction.

A polygenic risk score is finally defined as the sum of allele counts of the remaining SNPs (after clumping and thresholding) weighted by the corresponding GWAS effect sizes (Purcell *et al.*, 2009; Dudbridge, 2013; Wray *et al.*, 2014; Euesden *et al.*, 2015),

$$\mathrm{PRS_i} = \sum_{\substack{j \in S_{\text{clumping}} \\ p_j < p_T}} \hat{\beta}_\mathrm{j} \cdot \mathrm{G_{i,j}} \ ,$$

where $\hat{\beta}_j$ (p_j) are the effect sizes (p-values) estimated from the GWAS and $G_{i,j}$ is the allele count (genotype) for individual i and SNP j.

1.2.2 PRS for epidemiology

Polygenic Risk Scores (PRS) have been used for epidemiology before being used for prediction. The steps for a PRS analysis are illustrated in figure 1.6 and have two goals. First, PRS can be used when there is no SNP detected (at 5×10^{-8}) in a GWAS in order to show that there is still a significant polygenic contribution to the phenotype of interest. For example, in 2009, a GWAS for schizophrenia by Purcell *et al.* (2009) found only a single significantly associated SNP, although this disease is known to be highly heritable. Yet, by constructing a PRS using these GWAS results and testing this polygenic score for association with schizophrenia in another independent dataset, Purcell

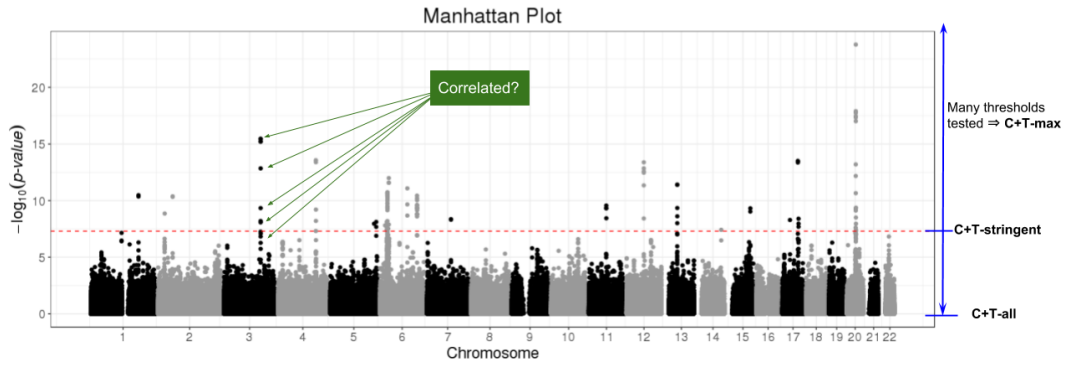


Figure 1.4: Illustration of C+T looking at a Manhattan plot from a GWAS of height based on 20,000 unrelated individuals from the UK Biobank dataset (Bycroft *et al.*, 2018). **Clumping** removes nearby SNPs that are too correlated with one another because indirect associations due to Linkage Disequilibrium provide only redundant information (see figure 1.5). **Thresholding** includes SNPs if they are significant enough ($p_j < p_T$) in order to reduce noise in the polygenic score.

et al. (2009) proved that there is a polygenic contribution to schizophrenia (Figure 1.7). Thus, polygenic analysis was central in demonstrating that the first phase of GWAS was underpowered, which justified the need for larger sample sizes that is now starting to pay off (Wray *et al.*, 2014).

Another use of PRS for epidemiology is to test the PRS for association with a phenotype that is different from the one used to compute the summary statistics. This technique enables researchers to prove that there is a common genetic contribution between two traits. For example, it was shown that there is a common genetic contribution

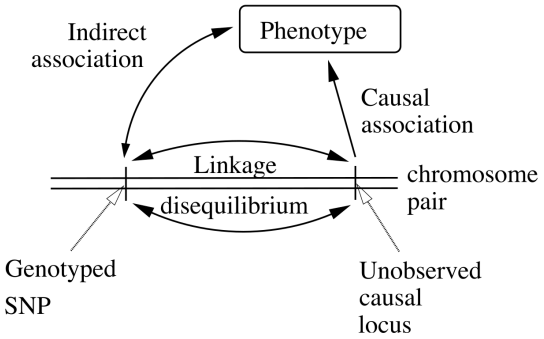


Figure 1.5: Illustration of an indirect association with a phenotype due to Linkage Disequilibrium between SNPs. Source: Astle *et al.* (2009).

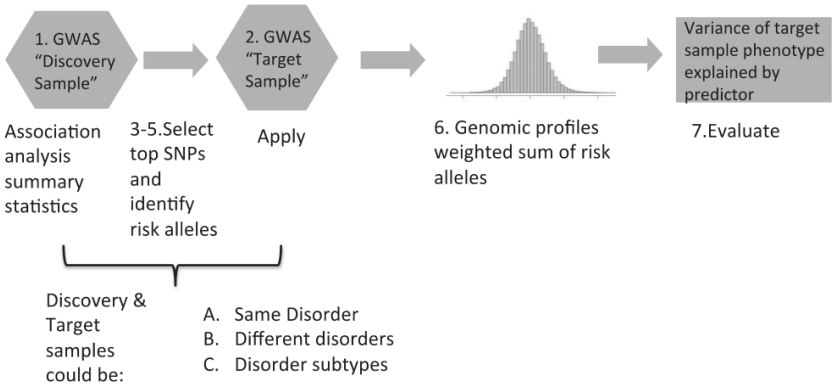


Figure 1.6: Illustration of the steps in genomic profile risk scoring. Source: Wray *et al.* (2014).

between schizophrenia and bipolar disorder (Figure 1.7).

1.2.3 The differing goals of association testing and risk prediction

Association testing (GWAS) and prediction have very different goals. First, GWAS aims at identifying highly replicable disease-associated variants by using a highly stringent p-value threshold to prevent false discoveries. However, using only hits from GWAS results in PRS of low predictive value (see section 1.3.1). A common mistake is to report highly significant findings with large odds ratios as useful predictors of disease. Thus, people have been reminded over the years that GWAS findings are often not predictive on their own even if they are highly associated with the disease of interest, and that we would need scores that combine many SNPs in order to have a decent predictor of disease, i.e. polygenic scores (Pepe *et al.*, 2004; Janssens *et al.*, 2006; Jakobsdottir *et al.*, 2009; Wald and Old, 2019).

Finally, it should be noted that population stratification, usually considered an unwelcome confounder in GWAS, may be useful in risk prediction and may be leveraged to produce better models (Golan and Rosset, 2014; Abraham and Inouye, 2015). Indeed, for predictive purposes, the objective is to provide the best possible prediction and confounding is not an issue.

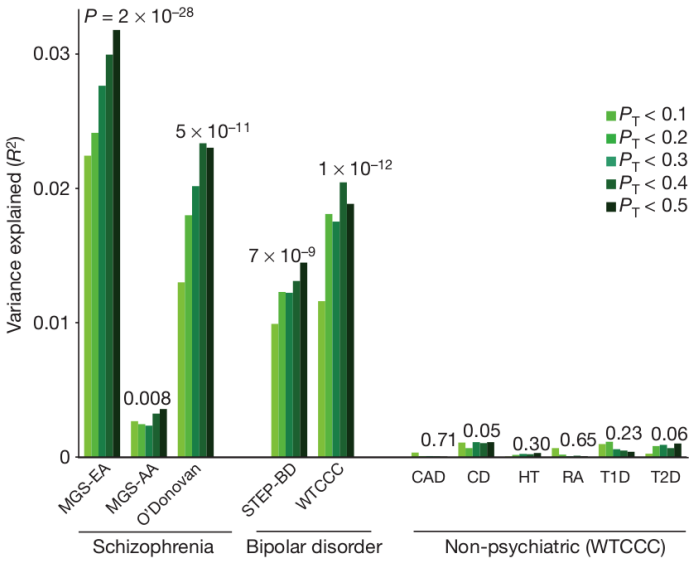


Figure 1.7: Replication of the polygenic component derived by the International Schizophrenia Consortium in independent schizophrenia and bipolar disorder samples. A PRS was computed using summary statistics from a GWAS of schizophrenia, and this polygenic score was tested for association with schizophrenia, bipolar disorder and other diseases in independent datasets. This proved that there was a polygenic contribution to schizophrenia, a common genetic contribution between schizophrenia and bipolar disorder, but no apparent common genetic contribution between schizophrenia and other diseases such as coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis and diabetes. Associations were maximized for $p_T = 0.5$, i.e. including more than half of all SNPs. Source: Purcell *et al.* (2009).

1.3 Polygenic prediction

1.3.1 Heritability and missing heritability

The basic components of disease risk are usually broken down into genetic susceptibility, environmental exposures and lifestyle factors. Thus, all disease incidence cannot be predicted by genetic factors only. For a quantitative phenotype, we call heritability (h^2) the proportion of phenotypic variation that is attributable to genetic factors among a population (Visscher *et al.*, 2008). Methods now enable the estimation of chip-heritability (also called SNP-heritability: h^2_{SNP}) using linear mixed models (LMM) and residual maximum likelihood (ReML). For example, for a chip of 300K SNPs, it was shown

that those SNPs could account for 45% of the variance of height (Yang *et al.*, 2010). Note that the heritability of height is estimated to be around 80% (Silventoinen, 2003; Visscher *et al.*, 2006); the difference between these two values can be explained by the fact that 300K SNPs cannot capture the same variation in height as the 3 billion base pairs of DNA. This difference can also reflect an overestimation of heritability (Visscher *et al.*, 2008). Authors of a recent preprint claim that they can recover the full heritability for height and BMI using both rare and common variants from WGS data (Wainschtein *et al.*, 2019).

Heritability is the upper bound in terms of prediction power (when measured with R^2) that we can get using a model from genetic variants only. The difference between R^2 and h^2 has been termed "missing heritability" (Manolio *et al.*, 2009). So, the main goal of my thesis is to get best possible predictions based on genetic data in order to reduce this missing heritability.

The gap between predictions and heritability estimates was very large in the first years of GWAS. For example, first GWAS found only 12 associated SNPs for type 2 diabetes and only 2 for prostate cancer, explaining a very small proportion of heritability for these diseases (Jakobsdottir *et al.*, 2009). Likewise, in 2008, only 40 genome-wide-significant SNPs had been identified for height, and together they explained about 5% of the heritability of height (Manolio *et al.*, 2009). In 2014, the number of associated SNPs had increased to around 700 for height, explaining 20% of its heritability (Wood *et al.*, 2014). Since many of the identified associated SNPs have an effect size close to the limit dictated by the power of the studies, a likely explanation, at least in part, is that there are many common polymorphisms with effects that are too small to be identified at the stringent significance threshold of current GWAS (Wray *et al.*, 2008). Therefore, as results from multiple GWAS are combined to increase sample size, a larger fraction of the genetic variance is likely to be explained and accurate prediction of genetic risk to disease will become possible even though the risks conveyed by individual variants are small (Wray *et al.*, 2008, 2018). These findings have also led people to use not only genome-wide significant SNPs, but many other SNPs, sometimes not even marginally significant (i.e. with a p-value > 5%) in order to maximize predictive power (Purcell *et al.*, 2009; Dudbridge, 2013; Wray *et al.*, 2014).

1.3.2 Methods for polygenic prediction

Several methods have been developed to predict disease status based on genetic data. We can divide these methods in two categories: the ones that use summary statistics and the ones that use individual-level data only.

When summary statistics are available, the most widely used method is called "Clumping + Thresholding" (C+T), which has been described in section 1.2.1. More recently, researchers have focused their efforts on implementing more elegant and potentially more optimal ways to account for LD, as a replacement of clumping that simply discards SNPs (Vilhjálmsson *et al.*, 2015; Mak *et al.*, 2017; Chun *et al.*, 2019; Ge *et al.*, 2019). Take the solution of a linear regression $y = X\beta + \epsilon$, $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$. This vector of effect sizes $\hat{\beta}$, estimated from all variables at once, can be decomposed in two parts: $X^T y$ that represents the marginal effects, i.e. the effects of each variable when learned independently (up to some scaling); and $\left(X^T X\right)^{-1}$, some rotation of the effects that account for the correlation between variables. Then, the first element can be replaced by summary statistics and the second element can be replaced by an estimation of LD obtained e.g. from a reference panel.

Moreover, these methods handle weights differently than C+T that directly uses GWAS effect sizes as weights in the PRS, or weights of 0 for SNPs not passing the clumping and thresholding steps. Instead, these methods usually shrink effects towards 0. Apart from "lassosum" of Mak *et al.* (2017), the other methods do not perform variable selection at all. This means that if you use GWAS summary statistics for 10M variants as input, you would get a predictive model composed of 10M variables (Janssens and Joyner, 2019).

When using individual-level data only, the problem boils down to a standard classification problem. Thus, some statistical learning methods have been used to derive PRS for complex human diseases by jointly estimating SNP effects. Such methods include joint logistic regression, Support Vector Machine (SVM) and random forests (Wei *et al.*, 2009; Abraham *et al.*, 2012, 2014; Botta *et al.*, 2014; Okser *et al.*, 2014). 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 (Yang *et al.*, 2010). However, these methods and their derivatives are often computationally very demanding, both in terms of memory and time required (Zhou *et al.*,

2013; Golan and Rosset, 2014; Speed and Balding, 2014; Maier *et al.*, 2015). Recently, two methods named BOLT-LMM and SAIGE have been developed to handle very large datasets (Loh *et al.*, 2018; Zhou *et al.*, 2018). BOLT-LMM and SAIGE were primarily designed for association testing but can also be used for prediction purposes based on individual-level data.

1.3.3 Objective and main difficulties of the thesis

We want to use genetic data to help distinguish between cases and controls for a given disease, or at least to stratify people in the population in order to improve early detection of diseases and prevention for high-risk individuals. Genomic data are usually very large and highly dimensional with hundreds of thousands of variables to many millions, for thousands or hundreds of thousands individuals. Thanks to large sample sizes of recent GWAS studies, many robust associations between DNA variants and many diseases have been identified. Yet, individually, these variants generally have a small effect on disease susceptibility, explaining a small fraction of the total heritability of the diseases studied. In order to have predictive models useful in clinical settings, we need to combine the information from a multitude of DNA variants (polygenic models), coming from multiple studies and in diverse formats (e.g. individual-level data and summary statistics).

To improve current disease predictions from Polygenic Risk Scores (PRS), we have focused on using methods from the statistical learning community, which have received only moderate attention in the "predictive human genetics" field. The main difficulty in using these methods is that they do not necessarily scale well with the large-scale data we now have in this field. For example, the UK Biobank is composed of 500K individuals from which 90M variants are available (Bycroft *et al.*, 2018). When analyzing these large-scale datasets, only a few methods can be used. Most of them are being developed in a separate piece of software that does a specific analysis. Yet, if you want to do some exploratory analyses and test new ideas, it becomes increasingly difficult to do so.

Thus, the first part of our work has been dedicated to developing two R packages that could handle very large datasets, while being simple and flexible to use for both standard and exploratory analyses. Our second paper has been dedicated to implementing penalized regressions as a replacement to more simple, less optimal methods, and

that could be used for very large individual-level datasets. Finally, because lots of summary statistics data are available while individual-level data are still scarce, we worked on making the most of the Clumping and Thresholding (C+T) method since it proved to be a simple and effective method for constructing PRS based on large GWAS summary statistics and smaller individual-level datasets.

Chapter 2

Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr

2.1 Summary of the article

2.1.1 Introduction

Sample size of GWAS data has rapidly grown due to the reduction in genotyping costs over the years. Moreover, thanks to the imputation of many non-genotyped SNPs, the number of available SNPs for a given dataset has grown to millions. In 2007, there were datasets with 2000 cases and 3000 controls, genotyped over 300K SNPs (Wellcome Trust Case Control Consortium, 2007). Now, there are datasets of 500K individuals, genotyped over 800K SNPs, and imputed over 90M SNPs (Bycroft *et al.*, 2018). Genotype data are the first data of the omics family to have grown to such large scale. To analyze these datasets, software have been consistently produced or updated over the years to keep up with growing sizes. I think this is one of a few fields where producing software is really recognized as an important part of research to help advance the field. An obvious example in genetics is PLINK, a command line piece of software whose first version has been cited more than 17K times since 2007 and whose second version has already been cited more than 1500 times since 2015 (Purcell *et al.*, 2007; Chang

et al., 2015). This software is useful for file conversions as well as many types of SNP data analyses and is used in plant, animal and human genetics alike.

I wanted to use R to analyze data from this field as it provides excellent tools for exploratory analyses. R is a programming language that makes it easy to tie together existing or new functions to be used as part of large, interactive and reproducible analyses (R Core Team, 2018). Yet, most of the R packages that have been developed in human genetics are now obsolete because they cannot scale to the size of the data we currently have in the field. The first problem there is to solve is to actually store the data. For example, a standard R matrix of size 500K x 800K would require 3TB of RAM just to access it in memory. The second problem concerns computation time; if all functions provided by a package take two weeks to run, it is not really useful.

2.1.2 Methods

We developed two R packages called bigstatsr and bigsnpr. To solve the memory issue, we use a data format stored as a binary file on disk but that can be accessed almost as if it were a standard R matrix in memory. To provide functions with a reasonable computation time, I spent thousands of hours on the performance of code. Moreover, most of the functions provided in these packages are parallelized, which is facilitated by the fact that the data is stored on disk, therefore accessible by each process without the need of any copying. The R packages makes extensive use of some C++ code in order to fully optimize key parts of the available functions.

Specifically, package bigstatsr provides the on-disk data format and some standard statistical algorithms such as Principal Component Analysis (PCA), multiple association testing (GWAS, EWAS, TWAS, etc.), matrix products, etc. for this data format. This package is not specific to genetic data and can be used by other fields. Package bigsnpr builds on top of package bigstatsr and provides algorithms specific to GWAS data. It also provides wrappers to widely used software such as PLINK in order to perform all analyses within R, making it both simple and reproducible[1]. To save some disk space and make accesses faster, we store genotype matrices using one byte per element only, instead of eight bytes per element for a standard R matrix. With a special format, we

[1] `https://hackseq.github.io/2017_project_5/all-in-R.html` (Grande *et al.*, 2018)

are able to store both hard calls (0s, 1s, 2s and NAs) and dosages (expected values from imputation probabilities, $d = 0 \times \mathbb{P}(0) + 1 \times \mathbb{P}(1) + 2 \times \mathbb{P}(2)$).

We also developed two new algorithms by building on existing R packages. One algorithm is used for the imputation of missing values inside a genotype matrix. Generally, there are less than 1% of missing data in a genotype matrix, and current algorithms for filling these blanks relies on complex inference algorithms. Notably, these algorithms rely on a first step of phasing, which consists in inferring haplotypes from genotypes. Phasing is very computationally demanding, so that we propose an algorithm based on XGBoost (Chen and Guestrin, 2016), an efficient algorithm for building decision trees using extreme gradient boosting, which allows for reconstructing data for one SNP based on non-linear combinations of nearby SNPs. The other algorithm we developed infer consecutive loadings that capture the structure of long-range LD regions instead of capturing population structure when performing PCA on SNP data. This new algorithm relies on pcadapt, an algorithm that find outlier loadings in PCA (Luu *et al.*, 2017).

2.1.3 Results

We show that our two R packages are very efficient and can perform standard analyses as fast as dedicated command line software such as PLINK, and much faster than previously developed R packages. We also show that commonly used software for computing principal components of genomic data are not accurate enough in some cases. Finally, we show that, thanks to our two newly developed algorithms, we are able to quickly impute the few missing values in a genotype matrix while being almost as accurate as more complex and computationally demanding software. We also show that our PCA algorithm is able to detect and remove long-range LD regions, which makes it possible to automatically retrieve population structure without capturing any LD structure in PCA of SNP data.

2.1.4 Discussion

We developed two very fast R packages for analyzing large genomic data. One of them, bigstatsr, is not specific to SNP data so that it could be used by other fields that need to

analyze large matrices. Moreover, we think these packages are simple to use, very well tested and easily maintainable because of relatively simple code. The two R packages use a matrix-like format, which makes it easy to develop new functions in order to experiment and develop new ideas. Integration in R makes it possible to take advantage of the vast and diverse R packages.

2.2 Article 1 and supplementary materials

The following article is published in *Bioinformatics* [2].

[2] https://doi.org/10.1093/bioinformatics/bty185

Bioinformatics, 34(16), 2018, 2781–2787
doi: 10.1093/bioinformatics/bty185
Advance Access Publication Date: 30 March 2018
Original Paper

Genetics and population analysis

Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr

Florian Privé[1,*], Hugues Aschard[2,3], Andrey Ziyatdinov[3] and Michael G.B. Blum[1,*]

[1]Laboratoire TIMC-IMAG, UMR 5525, CNRS, Université Grenoble Alpes, 38058 Grenoble, France, [2]Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, 75724 Paris, France and [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on October 6, 2017; revised on February 2, 2018; editorial decision on March 22, 2018; accepted on March 29, 2018

Abstract

Motivation: Genome-wide datasets produced for association studies have dramatically increased in size over the past few years, with modern datasets commonly including millions of variants measured in dozens of thousands of individuals. This increase in data size is a major challenge severely slowing down genomic analyses, leading to some software becoming obsolete and researchers having limited access to diverse analysis tools.

Results: Here we present two R packages, bigstatsr and bigsnpr, allowing for the analysis of large scale genomic data to be performed within R. To address large data size, the packages use memory-mapping for accessing data matrices stored on disk instead of in RAM. To perform data pre-processing and data analysis, the packages integrate most of the tools that are commonly used, either through transparent system calls to existing software, or through updated or improved implementation of existing methods. In particular, the packages implement fast and accurate computations of principal component analysis and association studies, functions to remove single nucleotide polymorphisms in linkage disequilibrium and algorithms to learn polygenic risk scores on millions of single nucleotide polymorphisms. We illustrate applications of the two R packages by analyzing a case–control genomic dataset for celiac disease, performing an association study and computing polygenic risk scores. Finally, we demonstrate the scalability of the R packages by analyzing a simulated genome-wide dataset including 500 000 individuals and 1 million markers on a single desktop computer.

Availability and implementation: https://privefl.github.io/bigstatsr/ and https://privefl.github.io/bigsnpr/.

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

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide datasets produced for association studies have dramatically increased in size over the past few years, with modern datasets commonly including millions of variants measured in dozens of thousands of individuals. As a consequence, most existing software and algorithms have to be continuously optimized in order to avoid obsolescence. For computing principal component analysis (PCA), commonly performed to account for population stratification in association, a fast mode named FastPCA has been added to the software EIGENSOFT, and FlashPCA has been replaced by FlashPCA2 (Abraham and Inouye, 2014; Abraham *et al.*, 2016; Galinsky *et al.*, 2016; Price *et al.*, 2006). PLINK 1.07, which has

been a central tool in the analysis of genotype data, has been replaced by PLINK 1.9 to speed-up computations, and there is also an alpha version of PLINK 2.0 that will handle more data types (Chang *et al.*, 2015; Purcell *et al.*, 2007).

Increasing size of genetic datasets is a source of major computational challenges and many analytical tools would be restricted by the amount of memory (RAM) available on computers. This is particularly a burden for commonly used analysis languages such as R. For analyzing genotype datasets in R, a range of software are available, including for example the popular R packages GenABEL, SNPRelate and GWASTools (Aulchenko *et al.*, 2007; Gogarten *et al.*, 2012; Zheng *et al.*, 2012b). Solving memory issues for languages such as R would give access to a broad range of already implemented tools for data analysis. Fortunately, strategies have been developed to avoid loading large datasets in RAM. For storing and accessing matrices, memory-mapping is very attractive because it is seamless and usually much faster to use than direct read or write operations. Storing large matrices on disk and accessing them via memory-mapping has been available for several years in R through 'big.matrix' objects implemented in the R package bigmemory (Kane *et al.*, 2013).

2 Approach

In order to perform analyses of large-scale genomic data in R, we developed two R packages, bigstatsr and bigsnpr, that provide a wide-range of building blocks which are parts of standard analyses. R is a programming language that makes it easy to tie together existing or new functions to be used as part of large, interactive and reproducible analyses (R Core Team, 2017). We provide a similar format as file-backed 'big.matrix' objects that we called 'Filebacked Big Matrices (FBMs)'. Thanks to this matrix-like format, algorithms in R/C++ can be developed or adapted for large genotype data. This data format is a particularly good trade-off between easiness of use and computation efficiency, making our code both simple and fast. Package bigstatsr implements many statistical tools for several types of FBMs (unsigned char, unsigned short, integer and double). This includes implementation of multivariate sparse linear models, PCA, association tests, matrix operations and numerical summaries. The statistical tools developed in bigstatsr can be used for other types of data as long as they can be represented as matrices. Package bigsnpr depends on bigstatsr, using a special type of filebacked big matrix (FBM) object to store the genotypes, called 'FBM.code256'. Package bigsnpr implements algorithms which are specific to the analysis of single nucleotide polymorphism (SNP) arrays, such as calls to external software for processing steps, Input/Output (I/O) operations from binary PLINK files and data analysis operations on SNP data (thinning, testing, predicting and plotting). We use both a real case–control genomic dataset for celiac disease and large-scale simulated data to illustrate application of the two R packages, including two association studies and the computation of polygenic risk scores (PRS). We compare results from bigstatsr and bigsnpr with those obtained by using command-line software PLINK, EIGENSOFT and PRSice, and R packages SNPRelate and GWASTools. We report execution times along with the code to perform major computational tasks. For a comprehensive comparison between R packages bigstatsr and bigmemory, see Supplementary notebook 'bigstatsr-and-bigmemory'.

3 Materials and methods

3.1 Memory-mapped files

The two R packages do not use standard read operations on a file nor load the genotype matrix entirely in memory. They use a hybrid solution: memory-mapping. Memory-mapping is used to access data, possibly stored on disk, as if it were in memory. This solution is made available within R through the BH package, providing access to Boost C++ Header Files (http://www.boost.org/).

We are aware of the software library SNPFile that uses memory-mapped files to store and efficiently access genotype data, coded in C++ (Nielsen and Mailund, 2008) and of the R package BEDMatrix (https://github.com/QuantGen/BEDMatrix) which provides memory-mapping directly for binary PLINK files. With the two packages we developed, we made this solution available in R and in C++ via package Rcpp (Eddelbuettel and François, 2011). The major advantage of manipulating genotype data within R, almost as if it were a standard matrix in memory, is the possibility of using most of the other tools that have been developed in R (R Core Team, 2017). For example, we provide sparse multivariate linear models and an efficient algorithm for PCA based on adaptations from R packages biglasso and RSpectra (Qiu and Mei, 2016; Zeng and Breheny, 2017).

Memory-mapping provides transparent and faster access than standard read/write operations. When an element is needed, a small chunk of the genotype matrix, containing this element, is accessed in memory. When the system needs more memory, some chunks of the matrix are freed from the memory in order to make space for others. All this is managed by the operating system so that it is seamless and efficient. It means that if the same chunks of data are used repeatedly, it will be very fast the second time they are accessed, the third time and so on. Of course, if the memory size of the computer is larger than the size of the dataset, the file could fit entirely in memory and every second access would be fast.

3.2 Data management, pre-processing and imputation

We developed a special FBM object, called 'FBM.code256', that can be used to seamlessly store up to 256 arbitrary different values, while having a relatively efficient storage. Indeed, each element is stored in one byte which requires eight times less disk storage than double-precision numbers but four times more space than the binary PLINK format '.bed' which can store only genotype calls. With these 256 values, the matrix can store genotype calls and missing values (four values), best guess genotypes (three values) and genotype dosages (likelihoods) rounded to two decimal places (201 values). So, we use a single data format that can store both genotype calls and dosages.

For pre-processing steps, PLINK is a widely-used software. For the sake of reproducibility, one could use PLINK directly from R via systems calls. We therefore provide wrappers as R functions that use system calls to PLINK for conversion and quality control and a variety of formats can be used as input (e.g. vcf, bed/bim/fam, ped/map) and bed/bim/fam files as output (Supplementary Fig. S1). Package bigsnpr provides fast conversions between bed/bim/fam PLINK files and the 'bigSNP' object, which contains the genotype FBM (FBM.code256), a data frame with information on samples and another data frame with information on SNPs. We also provide another function which could be used to read from tabular-like text files in order to create a genotype in the format 'FBM'. Finally, we provide two methods for converting dosage data to the format 'bigSNP' (Supplementary notebook 'dosage').

Most modern SNP chips provide genotype data with large call-rates. For example, the celiac data we use in this paper presents only 0.04% of missing values after quality control. Yet, most of the functions in bigstatsr and bigsnpr do not handle missing values. So, we provide two functions for imputing missing values of genotyped

SNPs. Note that we do not impute completely missing SNPs which would require the use of reference panels and could be performed via e.g. imputation servers for human data (McCarthy *et al.*, 2016). The first function is a wrapper to PLINK and Beagle (Browning and Browning, 2007) which takes bed files as input and return bed files without missing values, and should therefore be used before reading the data in R (Supplementary Fig. S2). The second function is a new algorithm we developed in order to have a fast imputation method without losing much of imputation accuracy. This function also provides an estimator of the imputation error rate by SNP for post-quality control. This algorithm is based on machine learning approaches for genetic imputation (Wang *et al.*, 2012) and does not use phasing, thus allowing for a dramatic decrease in computation time. It only relies on some local XGBoost models (Chen and Guestrin, 2016). XGBoost, which is available in R, builds decision trees that can detect non-linear interactions, partially reconstructing phase, making it well suited for imputing genotype matrices. Our algorithm is the following: for each SNP, we divide the individuals in the ones which have a missing genotype (test set) and the ones which have a non-missing genotype for this particular SNP. Those latter individuals are further separated in a training set and a validation set (e.g. 80% training and 20% validation). The training set is used to build the XGBoost model for predicting missing data. The prediction model is then evaluated on the validation set for which we know the true genotype values, providing an estimator of the number of genotypes that have been wrongly imputed for that particular SNP. The prediction model is also projected on the test set (missing values) in order to impute them.

3.3 Population structure and SNP thinning based on linkage disequilibrium

For computing principal components (PCs) of a large-scale genotype matrix, we provide several functions related to SNP thinning and two functions for computing a partial singular value decomposition (SVD), one based on eigenvalue decomposition and the other one based on randomized projections, respectively named big_SVD and big_randomSVD (Fig. 1). While the function based on eigenvalue decomposition is at least quadratic in the smallest dimension, the function based on randomized projections runs in linear time in all dimensions (Lehoucq and Sorensen, 1996). Package bigstatsr uses the same PCA algorithm as FlashPCA2 called implicitly restarted Arnoldi method (IRAM), which is implemented in R package RSpectra. The main difference between the two implementations is that FlashPCA2 computes vector-matrix multiplications with the genotype matrix based on the binary PLINK file whereas bigstatsr computes these multiplications based on the FBM format, which enables parallel computations and easier subsetting.

SNP thinning improves ascertainment of population structure with PCA (Abdellaoui *et al.*, 2013). There are at least three different approaches to thin SNPs based on linkage disequilibrium. Two of them, named pruning and clumping, address SNPs in LD close to each other's because of recombination events, while the third one address long-range regions with a complex LD pattern due to other biological events such as inversions (Price *et al.*, 2008). First, pruning is an algorithm that sequentially scan the genome for nearby SNPs in LD, performing pairwise thinning based on a given threshold of correlation. Clumping is useful if a statistic is available for sorting the SNPs by importance. Clumping is usually used to post-process results of genome-wide association studies (GWAS) in order to keep only the most significant SNP per region of the genome. For PCA, the thinning procedure should remain unsupervised (no

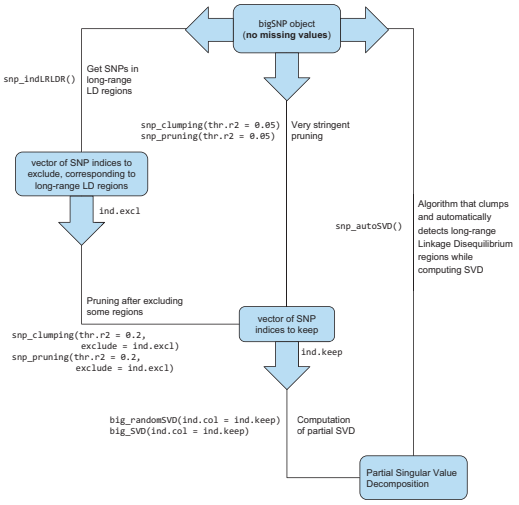


Fig. 1. Functions available in packages bigstatsr and bigsnpr for the computation of a partial singular value decomposition of a genotype array, with three different methods for thinning SNPs

phenotype must be used) and we therefore propose to use the minor allele frequency (MAF) as the statistic of importance. This choice is consistent with the pruning algorithm of PLINK; when two nearby SNPs are correlated, PLINK keeps only the one with the highest MAF. Yet, in some worst-case scenario, the pruning algorithm can leave regions of the genome without any representative SNP at all (Supplementary notebook 'pruning-vs-clumping'). So, we suggest to use clumping instead of pruning, using the MAF as the statistic of importance, which is the default in function snp_clumping of package bigsnpr. In practice, for the three datasets we considered, the clumping algorithm with the MAF provides similar sets of SNPs as when using the pruning algorithm (results not shown).

The third approach, which is generally combined with pruning, consists of removing SNPs in long-range LD regions (Price *et al.*, 2008). Long-range LD regions for the human genome are available as an online table (https://goo.gl/8TngVE) that package bigsnpr can use to discard SNPs in these regions before computing PCs. However, the pattern of LD might be population specific, so we developed an iterative algorithm that automatically detects these long-range LD regions and removes them. This algorithm consists in the following steps: first, PCA is performed using a subset of SNP remaining after clumping (with MAFs), then outliers SNPs are detected using the robust Mahalanobis distance as implemented in method pcadapt (Luu *et al.*, 2017). Finally, the algorithm considers that consecutive outlier SNPs are in long-range LD regions. Indeed, a long-range LD region would cause SNPs in this region to have strong consecutive weights (loadings) in the PCA. This algorithm is implemented in function snp_autoSVD of package bigsnpr and will be referred by this name in the rest of the paper.

3.4 Association tests and polygenic risk scores

Any test statistic that is based on counts could be easily implemented because we provide fast counting summaries. Among these tests, the Armitage trend test and the MAX3 test statistic are already provided for binary outcomes in bigsnpr (Zheng *et al.*, 2012a). Package bigstatsr implements statistical tests based on linear and logistic

regressions. For linear regression, a t-test is performed for each SNP j on $\beta^{(j)}$ where

$$\widehat{y} = \alpha^{(j)} + \beta^{(j)} SNP^{(j)} + \gamma_1^{(j)} PC_1 + \cdots + \gamma_K^{(j)} PC_K + \delta_1^{(j)} COV_1 + \cdots + \delta_L^{(j)} COV_L, \quad (1)$$

and K is the number of PCs and L is the number of other covariates (such as age and gender). Similarly, for logistic regression, a Z-test is performed for each SNP j on $\beta^{(j)}$ where

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = \alpha^{(j)} + \beta^{(j)} SNP^{(j)} + \gamma_1^{(j)} PC_1 + \cdots + \gamma_K^{(j)} PC_K + \delta_1^{(j)} COV_1 + \cdots + \delta_L^{(j)} COV_L, \quad (2)$$

and $\widehat{p} = \mathbb{P}(Y = 1)$ and Y denotes the binary phenotype. These tests can be used to perform GWAS and are very fast due to the use of optimized implementations, partly based on previous work by Sikorska *et al.* (2013).

The R packages also implement functions for computing PRS using two methods. The first method is the widely-used 'Clumping + Thresholding' (C + T, also called 'Pruning + Thresholding' in the literature) model based on univariate GWAS summary statistics as described in previous equations. Under the C + T model, a coefficient of regression is learned independently for each SNP along with a corresponding P-value (the GWAS part). The SNPs are first clumped (C) so that there remains only SNPs that are weakly correlated with each other. Thresholding (T) consists in removing SNPs that are under a certain level of significance (P-value threshold to be determined). A PRS is defined as the sum of allele counts of the remaining SNPs weighted by the corresponding regression coefficients (Chatterjee *et al.*, 2013; Dudbridge, 2013; Euesden *et al.*, 2015). On the contrary, the second approach does not use univariate summary statistics but instead train a multivariate model on all the SNPs and covariables *at once*, optimally accounting for correlation between predictors (Abraham *et al.*, 2012). The currently available models are very fast sparse linear and logistic regressions. These models include lasso and elastic-net regularizations, which reduce the number of predictors (SNPs) included in the predictive models (Friedman *et al.*, 2010; Tibshirani, 1996; Zou and Hastie, 2005). Package bigstatsr provides a fast implementation of these models by using efficient rules to discard most of the predictors (Tibshirani *et al.*, 2012). The implementation of these algorithms is based on modified versions of functions available in the R package biglasso (Zeng and Breheny, 2017). These modifications allow to include covariates in the models, to use these algorithms on the special type of FBM called 'FBM.code256' used in bigsnpr and to remove the need of choosing the regularization parameter.

3.5 Data analyzed

In this paper, two datasets are analyzed: the celiac disease cohort and POPRES (Dubois *et al.*, 2010; Nelson *et al.*, 2008). The celiac dataset is composed of 15 283 individuals of European ancestry genotyped on 295 453 SNPs. The POPRES dataset is composed of 1385 individuals of European ancestry genotyped on 447 245 SNPs. For computation time comparisons, we replicated individuals in the celiac dataset 5 and 10 times in order to increase sample size while keeping the same eigen decomposition (up to a constant) and pairwise SNP correlations as the original dataset. To assess scalability of the packages for a biobank-scale genotype dataset, we formed another dataset of 500 000 individuals and 1 million SNPs, also through replication of the celiac dataset.

3.6 Reproducibility

All the code used in this paper along with results, such as execution times and figures, are available as HTML R notebooks in the Supplementary materials. In Supplementary notebook 'public-data', we provide some open-access data of domestic dogs so that users can test our code and functions on a moderate size dataset with 4342 samples and 145 596 SNPs (Hayward *et al.*, 2016).

4 Results

4.1 Overview

We present the results of four different analyses. First, we illustrate the application of R packages bigstatsr and bigsnpr. Second, by performing two GWAS, we compare the performance of bigstatsr and bigsnpr to the performance obtained with FastPCA (EIGENSOFT 6.1.4) and PLINK 1.9, and also two R packages SNPRelate and GWASTools (Chang *et al.*, 2015; Galinsky *et al.*, 2016; Gogarten *et al.*, 2012; Zheng *et al.*, 2012b). PCA is a computationally intensive step of the GWAS, so that we further compare PCA methods on larger datasets. Third, by performing a PRS analysis with summary statistics, we compare the performance of bigstatsr and bigsnpr to the performance obtained with PRSice-2 (Euesden *et al.*, 2015). Finally, we present results of the two new methods implemented in bigsnpr, one method for the automatic detection and removal of long-range LD regions in PCA and another for the in-sample imputation of missing genotypes (i.e. for genotyped SNPs only). We compare performance on two computers, a desktop computer with 64 GB of RAM and 12 cores (six physical cores), and a laptop with only 8 GB of RAM and 4 cores (two physical cores). For the functions that enable parallelism, we use half of the cores available on the corresponding computer. We present a table summarizing the features of different software in Supplementary Table S5.

4.2 Application

The data were pre-processed following steps from Supplementary Figure S1, removing individuals and SNPs with more than 5% of missing values, non-autosomal SNPs, SNPs with a MAF lower than 0.05 or a P-value for the Hardy–Weinberg exact test lower than 10^{-10}, and finally, removing the first individual in each pair of individuals with a proportion of alleles shared IBD >0.08 (Purcell *et al.*, 2007). For the POPRES dataset, this resulted in 1382 individuals and 344 614 SNPs with no missing value. For the celiac dataset, this resulted in 15 155 individuals and 281 122 SNPs with an overall genotyping rate of 99.96%. The 0.04% missing genotype values were imputed with the XGBoost method. If we would have used a standard R matrix to store the genotypes, this data would have required 32 GB of memory. On the disk, the '.bed' file requires 1 GB and the '.bk' file (storing the FBM) requires 4 GB.

We used bigstatsr and bigsnpr R functions to compute the first PCs of the celiac genotype matrix and to visualize them (Fig. 2). We then performed a GWAS investigating how SNPs are associated with celiac disease, while adjusting for PCs, and plotted the results as a Manhattan plot (Fig. 3). As illustrated in the Supplementary data, the whole pipeline is user-friendly, requires only 20 lines of R code and there is no need to write temporary files or objects because functions of packages bigstatsr and bigsnpr have parameters which enable subsetting of the genotype matrix without having to copy it.

To illustrate the scalability of the two R packages, we performed a GWAS analysis on 500 K individuals and 1 M SNPs. The GWAS analysis completed in ∼11 h using the aforementioned desktop computer. The GWAS analysis was composed of four main steps.

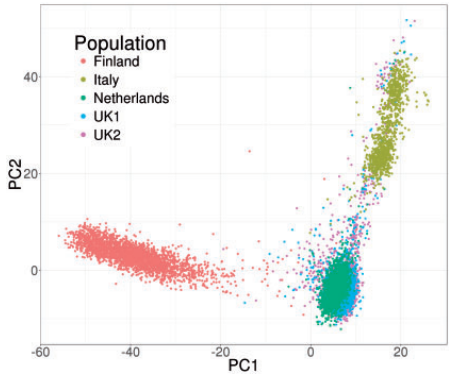


Fig. 2. Principal components of the celiac cohort genotype matrix produced by package bigstatsr

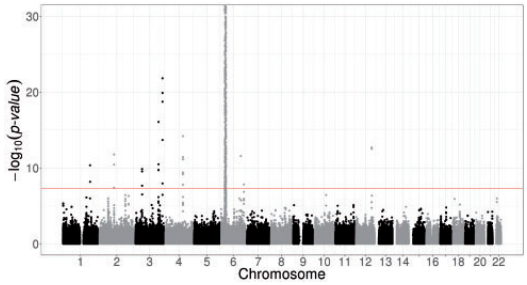


Fig. 3. Manhattan plot of the celiac disease cohort produced by package bigsnpr. Some SNPs in chromosome 6 have *P*-values smaller than the 10^{-30} threshold used for visualization purposes

First we converted binary PLINK files in the format 'bigSNP' in 1 h. Then, we removed SNPs in long-range LD regions and used SNP clumping, leaving 93 083 SNPs in 5.4 h. Then, the 10 first PCs were computed on the 500 K individuals and these remaining SNPs in 1.8 h. Finally, we performed a linear association test on the complete 500 K dataset for each of the 1 M SNPs, using the 10 first PCs as covariables in 2.9 h.

4.3 Performance and precision comparisons

First, we compared the GWAS computations obtained with bigstatsr and bigsnpr to the ones obtained with PLINK 1.9 and EIGENSOFT 6.1.4, and also two R packages SNPRelate and GWASTools. For most functions, multithreading is not available yet in PLINK, nevertheless, PLINK-specific algorithms that use bitwise parallelism (e.g. pruning) are still faster than the parallel algorithms reimplemented in package bigsnpr (Table 1). Overall, performing a GWAS on a binary outcome with bigstatsr and bigsnpr is as fast as when using EIGENSOFT and PLINK, and 19–45 times faster than when using R packages SNPRelate and GWASTools. For performing an association study on a continuous outcome, we report a dramatic increase in performance by using bigstatsr and bigsnpr, making it possible to perform such analysis in <2 min for a relatively large dataset such as the celiac dataset. This analysis was 7–19 times faster as compared to PLINK 1.9 and 28–74 times faster as compared to SNPRelate and GWASTools (Table 1). Note that the PC scores obtained are more accurate as compared to PLINK (see the last paragraph of this subsection), which is also the case for the *P*-values computed for the two GWAS (see Supplementary notebook 'GWAS-comparison').

Table 1. Execution times with bigstatsr and bigsnpr compared to PLINK 1.9 and FastPCA (EIGENSOFT) and also to R packages SNPRelate and GWASTools for making a GWAS for the celiac dataset (15 155 individuals and 281 122 SNPs). The first execution time is with a desktop computer (6 cores used and 64 GB of RAM) and the second one is with a laptop (2 cores used and 8 GB of RAM)

Operation\software	Execution times (in seconds)		
	FastPCA PLINK 1.9	bigstatsr bigsnpr	SNPRelate GWASTools
Converting PLINK files	n/a	6/20	13/33
Pruning	4/4	14/52	33/32
Computing 10 PCs	305/314	58/183	323/535
GWAS (binary phenotype)	337/284	291/682	16 220/17 425
GWAS (continuous phenotype)	1348/1633	10/23	6115/7101
Total (binary)	646/602	369/937	16 589/18 025
Total (continuous)	1657/1951	88/278	6484/7701

Table 2. Execution times with bigstatsr and bigsnpr compared to PRSice for making a PRS on the celiac dataset based on summary statistics for height. The first execution time is with a desktop computer (6 cores used and 64 GB of RAM) and the second one is with a laptop (2 cores used and 8 GB of RAM)

Operation\software	Execution times (in seconds)	
	PRSice	bigstatsr and bigsnpr
Converting PLINK files	—	6/20
Reading summary stats	—	4/6
Clumping	—	9/31
PRS	—	2/33
Compute *P*-values	—	1/1
Total	22/29	22/91

Second, we compared the PRS analysis performed with the R packages to the one using PRSice-2. There are five main steps in such an analysis (Table 2), including four steps handled with functions of packages bigstatsr and bigsnpr. The remaining step is the reading of summary statistics which can be performed with the widely used function fread of R package data.table. Using bigstatsr and bigsnpr results in an analysis as fast as with PRSice-2 when using our desktop computer, and three times slower when using our laptop (Table 2).

Finally, on our desktop computer, we compared the computation times of FastPCA (fast mode of EIGENSOFT), FlashPCA2 and PLINK 2.0 (approx mode) to the similar function big_randomSVD implemented in bigstatsr. For each comparison, we used the 93 083 SNPs which were remaining after pruning and we computed 10 PCs. We used the datasets of growing size simulated from the celiac dataset (from 15 155 to 151 550 individuals). Overall, function big_randomSVD is almost twice as fast as FastPCA and FlashPCA2 and eight times as fast as when using parallelism with six cores, an option not currently available in either FastPCA or FlashPCA2 (Fig. 4). PLINK 2.0 is faster than bigstatsr with a decrease in time of 20–40%. We also compared results in terms of precision by comparing squared correlation between approximated PCs and 'true' PCs provided by an exact eigen decomposition obtained with PLINK 2.0 (exact mode). Package bigstatsr and FlashPCA2 (that use the same algorithm) infer all PCs with a squared correlation of more than

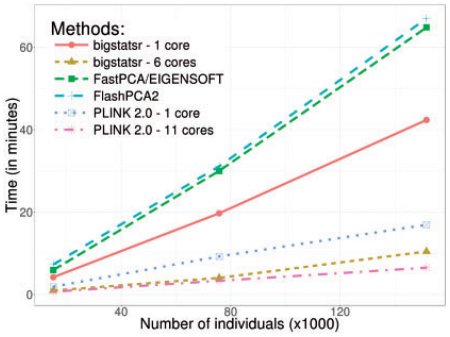


Fig. 4. Benchmark comparisons between randomized partial singular value decomposition available in FlashPCA2, FastPCA (fast mode of SmartPCA/EIGENSOFT), PLINK 2.0 (approx mode) and package bigstatsr. It shows the computation time in minutes as a function of the number of samples. The first 10 principal components have been computed based on the 93 083 SNPs which remained after thinning

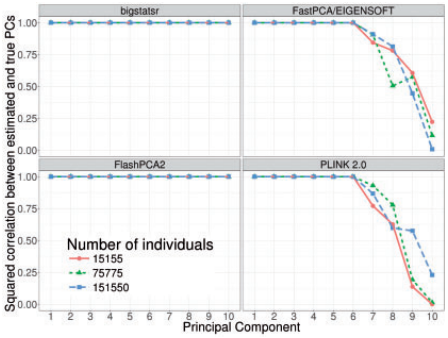


Fig. 5. Precision comparisons between randomized partial singular value decomposition available in FlashPCA2, FastPCA (fast mode of SmartPCA/EIGENSOFT), PLINK 2.0 (approx mode) and package bigstatsr. It shows the squared correlation between approximated PCs and 'true' PCs (produced by the exact mode of PLINK 2.0) of the celiac dataset (whose individuals have been repeated 1, 5 and 10 times)

0.999 between true PCs and approximated ones (Fig. 5). Yet, FastPCA (fast mode of EIGENSOFT) and PLINK 2.0 (that use the same algorithm) infer the true first six PCs but the squared correlation between true PCs and approximated ones decreases for further PCs (Fig. 5).

4.4 Automatic detection of long-range LD regions

For detecting long-range LD regions during the computation of PCA, we tested the function snp_autoSVD on both the celiac and POPRES datasets. For the POPRES dataset, the algorithm converged in two iterations. The first iterations found three long-range LD regions in chromosomes 2, 6 and 8 (Supplementary Table S1). We compared the PCs of genotypes obtained after applying snp_autoSVD with the PCs obtained after removing pre-determined long-range LD regions (https://goo.gl/8TngVE) and found a mean correlation of 89.6% between PCs, mainly due to a rotation of PC7 and PC8 (Supplementary Table S2). For the celiac dataset, we found five long-range LD regions (Supplementary Table S3) and a mean correlation of 98.6% between PCs obtained with snp_autoSVD and the ones obtained by clumping and removing pre-determined long-range LD regions (Supplementary Table S4).

For the celiac dataset, we further compared results of PCA obtained when using snp_autoSVD and when computing PCA without removing any long range LD region (only clumping at $R^2 > 0.2$). When not removing any long range LD region, we show that PC4 and PC5 do not capture population structure and correspond to a long-range LD region in chromosome 8 (Supplementary Figs S3 and S4). When automatically removing some long-range LD regions with snp_autoSVD, we show that PC4 and PC5 reflect population structure (Supplementary Fig. S3). Moreover, loadings are more equally distributed among SNPs after removal of long-range LD regions (Supplementary Fig. S4). This is confirmed by Gini coefficients (measure of dispersion) of each squared loadings that are significantly smaller when computing PCA with snp_autoSVD than when no long-range LD region is removed (Supplementary Fig. S5).

4.5 Imputation of missing values for genotyped SNPs

For the imputation method based on XGBoost, we compared the imputation accuracy and computation times with Beagle on the POPRES dataset (with no missing value). The histogram of the MAFs of this dataset is provided in Supplementary Figure S6. We used a beta-binomial distribution to simulate the number of missing values by SNP and then randomly introduced missing values according to these numbers, resulting in ∼3% of missing values overall (Supplementary Fig. S7). Imputation was compared between function snp_fastImpute of package bigsnpr and Beagle 4.1 (version of January 21, 2017) by counting the percentage of imputation errors (when the imputed genotype is different from the true genotype). Overall, in three runs, snp_fastImpute made only 4.7% of imputation errors and Beagle made only 3.1% of errors. Yet, it took Beagle 14.6 h to complete while snp_fastImpute only took 42 min (20 times less). We also note that snp_fastImpute made less 0/2 switching errors, i.e. imputing with a homozygous referent where the true genotype is a homozygous variant, or the contrary (Supplementary notebook 'imputation'). We also show that the estimation of the number of imputation errors provided by function snp_fastImpute is accurate (Supplementary Fig. S8), which can be useful for post-processing the imputation by removing SNPs with too many errors (Supplementary Fig. S9). For the celiac dataset in which there were already missing values, in order to further compare computation times, we report that snp_fastImpute took <10 h to complete for the whole genome whereas Beagle did not finish imputing chromosome 1 in 48 h.

5 Discussion

We have developed two R packages, bigstatsr and bigsnpr, which enable multiple analyses of large-scale genotype datasets in R thanks to memory-mapping. Linkage disequilibrium pruning, PCA, association tests and computation of PRS are made available in these software. Implemented algorithms are both fast and memory-efficient, allowing the use of laptops or desktop computers to make genome-wide analyses. Technically, bigstatsr and bigsnpr could handle any size of datasets. However, if the OS has to often swap between the file and the memory for accessing the data, this would slow down data analysis. For example, the PCA algorithm in bigstatsr is iterative so that the matrix has to be sequentially accessed over a hundred times. If the number of samples times the number of SNPs remaining after pruning is larger than the available memory, this slowdown would happen. For instance, a 32 GB computer would be

slow when computing PCs on more than 100K samples and 300K SNPs remaining after LD thinning.

The two R packages use a matrix-like format, which makes it easy to develop new functions in order to experiment and develop new ideas. Integration in R makes it possible to take advantage of the vast and diverse R libraries. For example, we developed a fast and accurate imputation algorithm for genotyped SNPs using the widely-used machine learning algorithm XGBoost available in the R package xgboost. Other functions, not presented here, are also available and all the functions available within the package bigstatsr are not specific to SNP arrays, so that they could be used for other omic data or in other fields of research.

We think that the two R packages and the corresponding data format could help researchers to develop new ideas and algorithms to analyze genome-wide data. For example, we wish to use these packages to train much more accurate predictive models than the standard C + T model currently in use for computing PRS. As a second example, multiple imputation has been shown to be a very promising method for increasing statistical power of a GWAS (Palmer and Pe'er, 2016), and it could be implemented with the data format 'FBM.code256' without having to write multiple files.

Funding

Authors acknowledge LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01). Authors also acknowledge the Grenoble Alpes Data Institute that is supported by the French National Research Agency under the 'Investissements d'avenir' program (ANR-15-IDEX-02).

Acknowledgements

Authors would like to thank the reviewers of this paper because their comments and suggestions have led to a significant improvement of this paper.

Conflict of Interest: none declared.

References

Abdellaoui,A. *et al.* (2013) Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.*, **21**, 1277–1285.

Abraham,G. and Inouye,M. (2014) Fast principal component analysis of large-scale genome-wide data. *PLoS ONE*, **9**, e93766.

Abraham,G. *et al.* (2012) SparSNP: fast and memory-efficient analysis of all SNPs for phenotype prediction. *BMC Bioinformatics*, **13**, 88.

Abraham,G. *et al.* (2016) FlashPCA2: principal component analysis of biobank-scale genotype datasets. *bioRxiv*, **12**, 2014–2017.

Aulchenko,Y.S. *et al.* (2007) Genabel: an r library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.

Chang,C.C. *et al.* (2015) Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.

Chatterjee,N. *et al.* (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genet.*, **45**, 400–405. 405e1–3.

Chen,T. and Guestrin,C. (2016) XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 785–794.

Dubois,P.C.A. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nature Genet.*, **42**, 295–302.

Dudbridge,F. (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, **9**, e1003348.

Eddelbuettel,D. and François,R. (2011) Rcpp: seamless R and C ++ integration. *J. Stat. Softw.*, **40**, 1–18.

Euesden,J. *et al.* (2015) PRSice: Polygenic Risk Score software. *Bioinformatics*, **31**, 1466–1468.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Galinsky,K.J. *et al.* (2016) Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.*, **98**, 456–472.

Gogarten,S.M. *et al.* (2012) GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, **28**, 3329–3331.

Hayward,J.J. *et al.* (2016) Complex disease and phenotype mapping in the domestic dog. *Nat. Commun.*, **7**, 10460.

Kane,M.J. *et al.* (2013) Scalable strategies for computing with massive data. *J. Stat. Softw.*, **55**, 1–19.

Lehoucq,R.B. and Sorensen,D.C. (1996) Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM J. Matrix Anal. Appl.*, **17**, 789–821.

Luu,K. *et al.* (2017) pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.*, **17**, 67–77.

McCarthy,S. *et al.* (2016) A reference panel of 64, 976 haplotypes for genotype imputation. *Nature Genet.*, **48**, 1279.

Nelson,M.R. *et al.* (2008) The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.*, **83**, 347–358.

Nielsen,J. and Mailund,T. (2008) SNPFile–a software library and file format for large scale association mapping and population genetics studies. *BMC Bioinformatics*, **9**, 526.

Palmer,C. and Pe'er,I. (2016) Bias characterization in probabilistic genotype data and improved signal detection with multiple imputation. *PLoS Genet.*, **12**, e1006091.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.*, **38**, 904–909.

Price,A.L. *et al.* (2008) Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.*, **83**, 132–135.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Qiu,Y. and Mei,J. (2016) *RSpectra: Solvers for Large Scale Eigenvalue and SVD Problems*. R package version 0.12-0.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sikorska,K. *et al.* (2013) Gwas on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics*, **14**, 166.

Tibshirani,R. (1996) Regression selection and shrinkage via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Tibshirani,R. *et al.* (2012) Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Series B Stat. Methodol.*, **74**, 245–266.

Wang,Y. *et al.* (2012) Fast accurate missing SNP genotype local imputation. *BMC Res. Notes*, **5**, 404.

Zeng,Y. and Breheny,P. (2017) The biglasso package: a memory-and computation-efficient solver for lasso model fitting with big data in R. *arXiv preprint arXiv: 1701.05936*.

Zheng,G. *et al.* (2012a) *Analysis of Genetic Association Studies. Statistics for Biology and Health*. Springer, Boston, MA, USA.

Zheng,X. *et al.* (2012b) A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, **28**, 3326–3328.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, **67**, 301–320.

1 Supplementary Data

1.1 Pre-processing

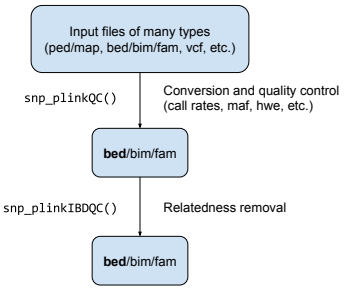


Figure S1: Conversion and Quality Control preprocessing functions available in package bigsnpr via system calls to PLINK.

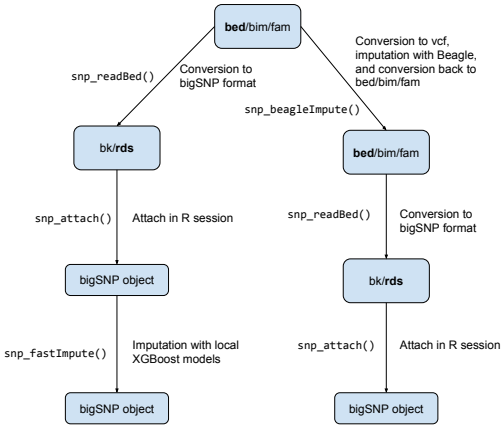


Figure S2: Imputation and reading functions available in package bigsnpr.

1.2 Long-range LD regions

	Chromosome	Start (Mb)	Stop (Mb)
1	2	134.7 (134.5)	137.3 (138)
2	6	27.5 (25.5)	33.1 (33.5)
3	8	6.6 (8)	13.2 (12)

Table S1: Regions found by snp_autoSVD for the POPRES dataset. Numbers in parentheses correspond to regions referenced in [Price *et al.*(2008)Price, Weale, Patterson, Myers, Need, Shianna, Ge, Rotter, Torres, Taylor, Goldst

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PC1	100.0	-0.1	-0.0	0.1	-0.1	0.0	0.0	0.0	-0.0	-0.0
PC2	0.1	100.0	-0.0	0.1	-0.0	-0.0	-0.0	0.2	-0.1	-0.0
PC3	0.0	-0.0	99.9	0.9	0.1	-0.1	-0.3	0.2	0.4	0.1
PC4	-0.1	-0.1	-0.9	99.7	-1.0	0.7	0.6	0.2	0.3	0.9
PC5	0.1	0.0	-0.1	1.1	99.3	1.3	-0.8	1.3	-4.2	-2.4
PC6	-0.0	0.0	0.1	-0.7	-1.0	97.7	-3.5	6.1	7.9	-6.2
PC7	-0.0	-0.1	0.2	-0.3	-1.7	0.3	58.3	73.2	-25.9	9.1
PC8	0.1	-0.1	-0.3	0.4	-0.5	-5.3	-73.5	59.5	15.8	13.2
PC9	0.0	0.1	-0.4	-0.8	5.0	-7.6	27.8	11.0	91.9	9.0
PC10	0.1	0.0	0.0	-0.9	1.6	10.2	3.9	-19.6	-6.3	89.2

Table S2: Correlation between scores of PCA for the POPRES dataset when automatically removing long-range LD regions and when removing them based on a predefined table.

	Chromosome	Start (Mb)	Stop (Mb)
1	2	134.4 (134.5)	138.1 (138)
2	6	23.8 (25.5)	35.8 (33.5)
3	8	6.3 (8)	13.5 (12)
4	3	163.1 (n/a)	164.9 (n/a)
5	14	46.6 (n/a)	47.5 (n/a)

Table S3: Regions found by snp_autoSVD for the celiac dataset. Numbers in parentheses correspond to regions referenced in [Price *et al.*(2008)Price, Weale, Patterson, Myers, Need, Shianna, Ge, Rotter, Torres, Taylor, Goldst

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PC1	100.0	-0.1	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PC2	0.1	100.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0
PC3	0.1	-0.0	99.9	0.2	-0.0	0.1	0.1	0.1	0.0	-0.1
PC4	-0.0	-0.0	-0.3	99.9	-0.1	0.1	-0.1	0.0	0.1	0.1
PC5	0.0	0.0	0.0	0.1	99.7	0.9	-0.3	0.1	-0.8	-0.6
PC6	-0.0	0.0	-0.1	-0.2	-0.8	99.6	0.5	-0.5	-0.2	-0.4
PC7	-0.0	0.0	-0.1	0.0	0.5	-0.4	98.9	3.1	0.7	1.6
PC8	0.0	0.0	-0.2	-0.0	-0.2	0.5	-3.2	98.4	-4.5	-1.5
PC9	-0.0	-0.0	-0.0	0.0	0.6	0.1	-0.7	4.6	96.9	-10.7
PC10	-0.0	-0.0	0.1	-0.1	0.3	0.1	-1.2	1.5	8.6	92.7

Table S4: Correlation between scores of PCA for the Celiac dataset when automatically removing long-range LD regions and when removing them based on a predefined table.

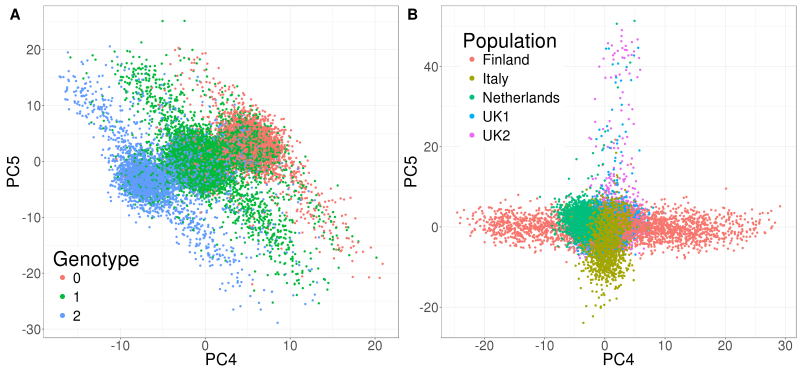


Figure S3: PC4 and PC5 of the celiac disease dataset. Left panel, PC scores obtained without removing any long range LD region (only clumping at $R^2 > 0.2$). Individuals are coloured according to their genotype at the SNP that has the highest loading for PC4. Right panel, PC scores obtained with the automatic detection and removal of long-range LD regions. Individuals are coloured according to their population of origin.

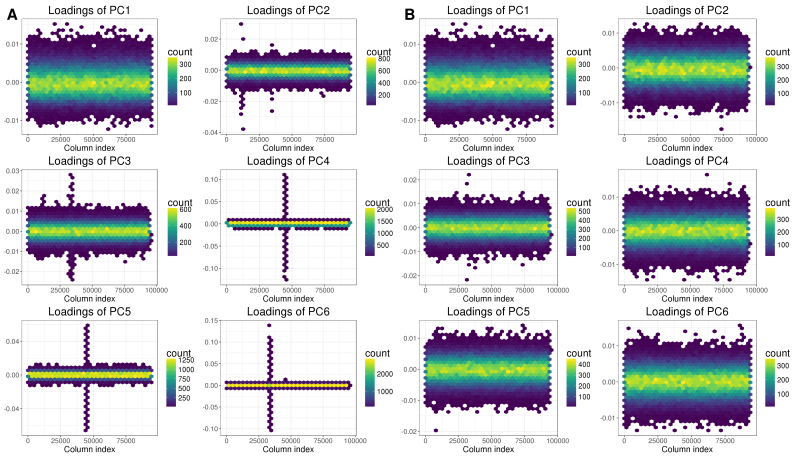


Figure S4: Loadings of first 6 PCs of the celiac disease dataset plotted as hexbins (2-D histogram with hexagonal cells). On the left, without removing any long range LD region (only clumping at $R^2 > 0.2$). On the right, with the automatic detection and removal of long-range LD regions.

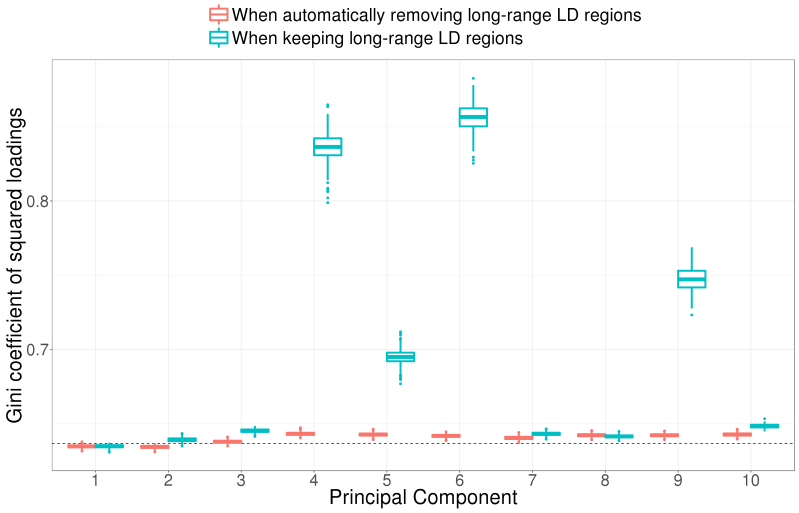


Figure S5: Boxplots of 1000 bootstrapped Gini coefficients (measure of statistical dispersion) of squared loadings without removing any long range LD region (only clumping at $R^2 > 0.2$) and with the automatic detection and removal of long-range LD regions. The dashed line corresponds to the theoretical value for gaussian loadings.

1.3 Imputation

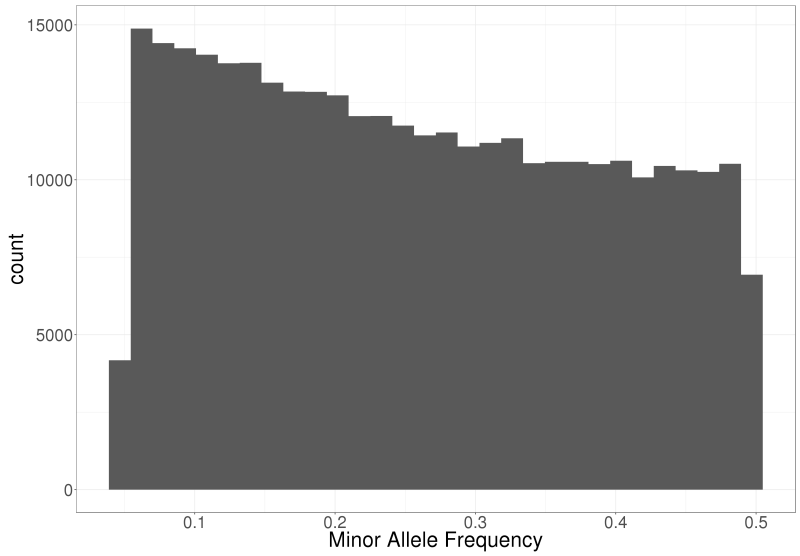


Figure S6: Histogram of the minor allele frequencies of the POPRES dataset used for comparing imputation methods.

References

[Chen and Guestrin(2016)Chen and Guestrin] Chen, T. and Guestrin, C. (2016). XGBoost : Reliable Large-scale Tree Boosting System. *arXiv*, pages 1–6.

[Price *et al.*(2008)Price, Weale, Patterson, Myers, Need, Shianna, Ge, Rotter, Torres, Taylor, Goldst Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D. D., Goldstein, D. B., and Reich, D. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations.

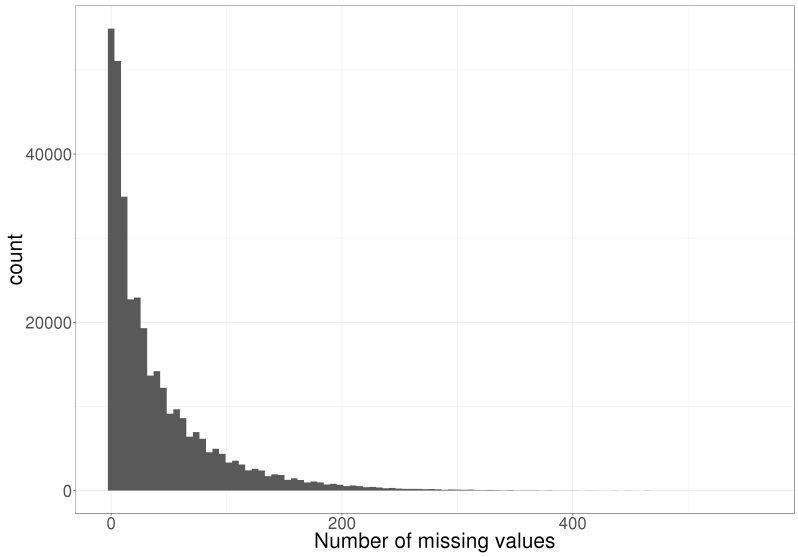


Figure S7: Histogram of the number of missing values by SNP. These numbers were generated using a Beta-binomial distribution.

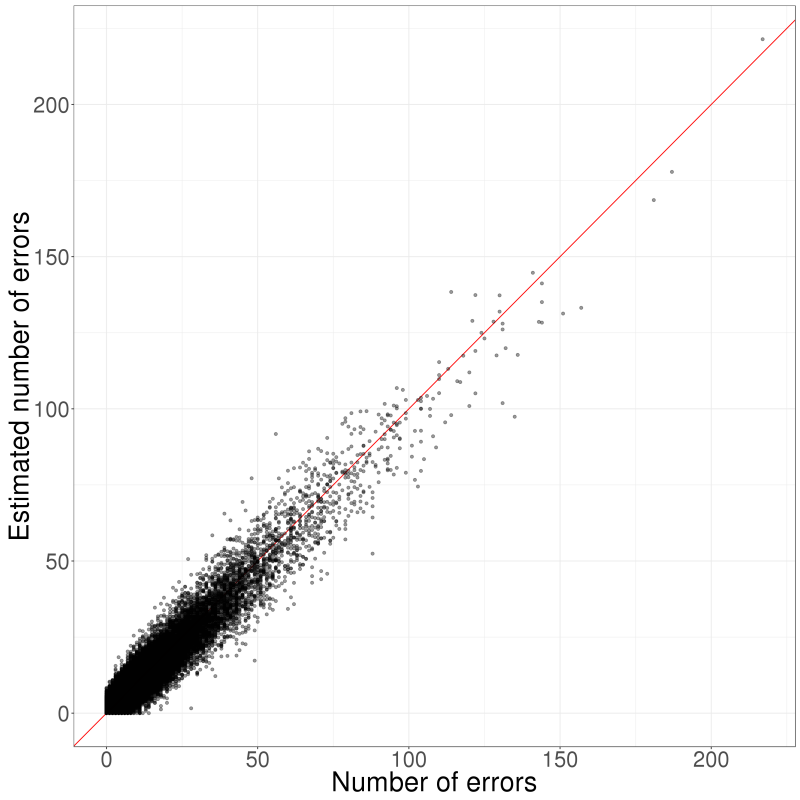


Figure S8: Number of imputation errors vs the estimated number of imputation errors by SNP. For each SNP with missing data, the number of imputation errors corresponds to the number of individuals for which imputation is incorrect. The estimated number of errors is a quantity that is returned when imputing with snp_fastimpute, which is based on XGBoost [Chen and Guestrin(2016)Chen and Guestrin,].

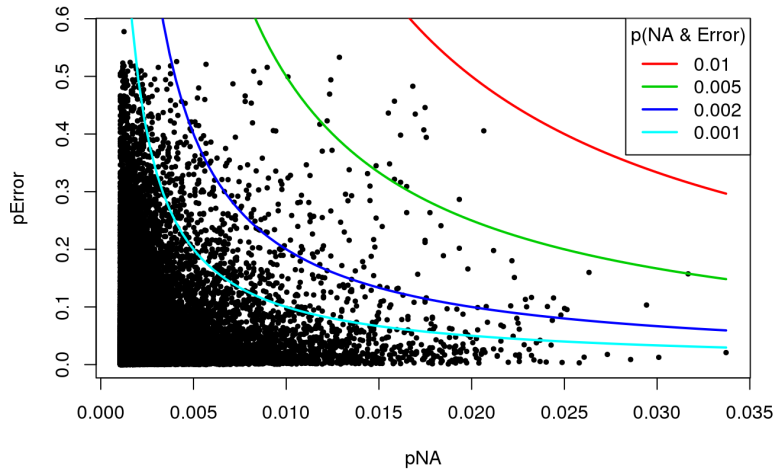


Figure S9: For each SNP (point), the estimated proportion of imputation errors ($\hat{p}(error \mid NA)$) vs the proportion of missing values ($p(NA) > 0.001$). These results come from the imputation of the Celiac dataset with function snp_fastImpute (Supplementary notebook "preprocessing"). Colored curves are representing the estimated proportion of wrong genotypes per SNP ($\hat{p}(error \ \& \ NA) = \hat{p}(error \mid NA) \cdot p(NA)$). This particularly shows that no SNP has more than 1% of wrong genotypes, allowing for no post-processing.

Chapter 3

Efficient Implementation of Penalized Regression for Genetic Risk Prediction

3.1 Summary of the article

3.1.1 Introduction

"Clumping+Thresholding" (C+T) is the most common method to derive Polygenic Risk Scores (PRS). C+T uses only GWAS summary statistics with a (small) individual-level data reference panel to account for linkage disequilibrium (LD). However, previous work showed that jointly estimating SNP effects has the potential to substantially improve predictive performance of PRS as compared to C+T (Abraham *et al.*, 2013). Moreover, now that large individual-level datasets such as the UK Biobank are available, it would be a waste of information to not use them to their full potential (Bycroft *et al.*, 2018). Indeed, in order for PRS to be useful in clinical settings, it should be as predictive as possible.

3.1.2 Methods

We included some efficient implementation of penalized (linear and logistic) regressions in R package bigstatsr. This implementation is not specific to genotype data at all, but this paper focuses on its application to predicting disease status based on large genotype data. We recall that bigstatsr uses some matrix format stored on disk instead

of memory, so that functions of this package can be very memory efficient. To make the algorithm very efficient, we based our implementation on existing implementations that use mathematical rules to quickly discard many variables as they will not enter the final model (Tibshirani *et al.*, 2012). These rules can be used when fitting penalized regression with either lasso or elastic net regularizations. To facilitate the choice of the two hyper-parameters of the elastic net regularization, we develop a procedure that we call Cross-Model Selection and Averaging (CMSA). CMSA is somehow similar to cross-validation but allows to derive an early stopping criterion that further increases the efficiency of the implementation.

We compare the penalized regressions with C+T and another method based on decision trees. We use extensive simulations to compare methods for different disease architectures, sample sizes and number of variables. We also compare methods in models with non-additive effects and show how to extend penalized regression to account for recessive and dominant effects on top of additive effects. Finally, we compare performance of methods using the UK Biobank, training models on 350K individuals and using 656K genotyped SNPs.

3.1.3 Results

We show that penalized regressions can provide large improvements in predictive performance as compared to C+T. When SNP effect sizes are small and sample size is small compared to the number of SNPs, PLR performs worse than C+T, but all methods provide poor predictive performance (AUC lower than 0.6) in this context. In contrast, when sample size is large enough, when there are some moderately large effects, or when there are some correlation between causal variants, using PLR substantially improves predictive performance as compared to C+T. By using some feature engineering technique, we are able to capture not only additive effects, but also recessive and dominant effects in penalized regressions. Finally, we show that our implementation of penalized regressions is scalable to datasets such as the UK Biobank, including hundreds of thousands of both observations and variables.

3.1.4 Discussion

In this paper, we demonstrate the feasibility and relevance of using penalized regressions for PRS computation when large individual-level datasets are available. Indeed, first, we show that the larger is the data, the larger is the gain in predictive performance of PLR over C+T. Second, we show that our implementation of PLR is scalable to very large datasets such as the UK Biobank. We discuss what makes our implementation scalable to very large datasets by explaining the algorithm and its memory requirements. Computation time is a function of the sample size and the number of variables with a predictive effect.

3.2 Article 2 and supplementary materials

The following article is published in *Genetics* [1].

[1] https://doi.org/10.1534/genetics.119.302019

HIGHLIGHTED ARTICLE
GENETICS | GENOMIC PREDICTION

Efficient Implementation of Penalized Regression for Genetic Risk Prediction

Florian Privé,*,[1] Hugues Aschard,[†] and Michael G. B. Blum*,[1]
*Laboratoire TIMC-IMAG, UMR 5525, University of Grenoble Alpes, CNRS, 38700 La Tronche, France and [†]Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, 75015 Paris, France

ABSTRACT Polygenic Risk Scores (PRS) combine genotype information across many single-nucleotide polymorphisms (SNPs) to give a score reflecting the genetic risk of developing a disease. PRS might have a major impact on public health, possibly allowing for screening campaigns to identify high-genetic risk individuals for a given disease. The "Clumping+Thresholding" (C+T) approach is the most common method to derive PRS. C+T uses only univariate genome-wide association studies (GWAS) summary statistics, which makes it fast and easy to use. However, previous work showed that jointly estimating SNP effects for computing PRS has the potential to significantly improve the predictive performance of PRS as compared to C+T. In this paper, we present an efficient method for the joint estimation of SNP effects using individual-level data, allowing for practical application of penalized logistic regression (PLR) on modern datasets including hundreds of thousands of individuals. Moreover, our implementation of PLR directly includes automatic choices for hyper-parameters. We also provide an implementation of penalized linear regression for quantitative traits. We compare the performance of PLR, C+T and a derivation of random forests using both real and simulated data. Overall, we find that PLR achieves equal or higher predictive performance than C+T in most scenarios considered, while being scalable to biobank data. In particular, we find that improvement in predictive performance is more pronounced when there are few effects located in nearby genomic regions with correlated SNPs; for instance, in simulations, AUC values increase from 83% with the best prediction of C+T to 92.5% with PLR. We confirm these results in a data analysis of a case-control study for celiac disease where PLR and the standard C+T method achieve AUC values of 89% and of 82.5%. Applying penalized linear regression to 350,000 individuals of the UK Biobank, we predict height with a larger correlation than with the best prediction of C+T (∼65% instead of ∼55%), further demonstrating its scalability and strong predictive power, even for highly polygenic traits. Moreover, using 150,000 individuals of the UK Biobank, we are able to predict breast cancer better than C+T, fitting PLR in a few minutes only. In conclusion, this paper demonstrates the feasibility and relevance of using penalized regression for PRS computation when large individual-level datasets are available, thanks to the efficient implementation available in our R package bigstatsr.

KEYWORDS polygenic risk scores; SNP; LASSO; genomic prediction; GenPred; shared data resources

POLYGENIC risk scores (PRS) combine genotype information across many single-nucleotide polymorphisms (SNPs) to give a score reflecting the genetic risk of developing a disease. PRS are useful for genetic epidemiology when testing polygenicity of diseases and finding a common genetic contribution between two diseases (Purcell *et al.* 2009). Personalized medicine is another major application of PRS. Personalized medicine envisions to use PRS in screening campaigns in order to identify high-risk individuals for a given disease (Chatterjee *et al.* 2016). As an example of practical application, targeting screening of 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). However, in order to be used in clinical settings, PRS should discriminate well enough between cases and controls. For screening high-risk individuals and for presymptomatic diagnosis of the general population, it is suggested that, for a

doi: https://doi.org/10.1534/genetics.119.302019
Manuscript received October 11, 2018; accepted for publication February 22, 2019; published Early Online February 26, 2019.
Available freely online through the author-supported open access option.

Supplemental material available at https://figshare.com/articles/code/7178750.
[1]Corresponding authors: Laboratoire TIMC-IMAG, UMR 5525, Université Grenoble Alpes, CNRS, 5 Ave. du Grand Sablon, 38700 La Tronche, France. E-mail: florian.prive@univ-grenoble-alpes.fr; and michael.blum@univ-grenoble-alpes.fr

10% disease prevalence, the AUC must be >75% and 99%, respectively (Janssens *et al.* 2007).

Several methods have been developed to predict disease status, or any phenotype, based on SNP information. A commonly used method often called "P+T" or "C+T" (which stands for "Clumping and Thresholding") is used to derive PRS from results of Genome-Wide Association Studies (GWAS) (Wray *et al.* 2007; Evans *et al.* 2009; Purcell *et al.* 2009; Chatterjee *et al.* 2013; Dudbridge 2013). This technique uses GWAS summary statistics, allowing for a fast implementation of C+T. However, C+T also has several limitations; for instance, previous studies have shown that predictive performance of C+T is very sensitive to the threshold of inclusion of SNPs, depending on the disease architecture (Ware *et al.* 2017). In parallel, statistical learning methods have also been used to derive PRS for complex human diseases by jointly estimating SNP effects. Such methods include joint logistic regression, Support Vector Machine (SVM) and random forests (Wei *et al.* 2009; Abraham *et al.* 2012, 2014; Botta *et al.* 2014; Okser *et al.* 2014; Lello *et al.* 2018; Mavaddat *et al.* 2019). Finally, Linear Mixed-Models (LMMs) are another widely used method in fields such as plant and animal breeding, or for predicting highly polygenic quantitative human phenotypes such as height (Yang *et al.* 2010). Yet, predictions resulting from LMM, known *e.g.*, as "gBLUP," have not proven as efficient as other methods for predicting several complex diseases based on genotypes [see table 2 of Abraham *et al.* (2013)].

We recently developed two R packages, bigstatsr and bigsnpr, for efficiently analyzing large-scale genome-wide data (Privé *et al.* 2018). Package bigstatsr now includes an efficient algorithm with a new implementation for computing sparse linear and logistic regressions on huge datasets as large as the UK Biobank (Bycroft *et al.* 2018). In this paper, we present a comprehensive comparative study of our implementation of penalized logistic regression (PLR), which we compare to the C+T method and the T-Trees algorithm, a derivation of random forests that has shown high predictive performance (Botta *et al.* 2014). In this comparison, we do not include any LMM method, yet, L2-PLR should be very similar to LMM methods. Moreover, we do not include any SVM method because it is expected to give similar results to logistic regression (Abraham *et al.* 2012). For C+T, we report results for a large grid of hyper-parameters. For PLR, the choice of hyper-parameters is included in the algorithm so that we report only one model for each simulation. We also use a modified version of PLR in order to capture not only linear effects, but also recessive and dominant effects.

To perform simulations, we use real genotype data and simulate new phenotypes. In order to make our comparison as comprehensive as possible, we compare different disease architectures by varying the number, size and location of causal effects as well as disease heritability. We also compare two different models for simulating phenotypes, one with additive effects only, and one that combines additive, dominant and interaction-type effects. Overall, we find that PLR achieves higher predictive performance than C+T except in highly underpowered cases (AUC values lower than 0.6), while being scalable to biobank data.

Materials and Methods

Genotype data

We use real genotypes of European individuals from a case-control study for celiac disease (Dubois *et al.* 2010). This dataset is presented in Supplemental Material, Table S1. Details of quality control and imputation for this dataset are available in Privé *et al.* (2018). For simulations presented later, we first restrict this dataset to controls from UK in order to remove the genetic structure induced by the celiac disease status and population structure. This filtering process results in a sample of 7100 individuals (see supplemental notebook "preprocessing"). We also use this dataset for real data application, in this case keeping all 15,155 individuals (4496 cases and 10,659 controls). Both datasets contain 281,122 SNPs.

Simulations of phenotypes

We simulate binary phenotypes using a Liability Threshold Model (LTM) with a prevalence of 30% (Falconer 1965). We vary simulation parameters in order to match a range of genetic architectures from low to high polygenicity. This is achieved by varying the number of causal variants and their location (30, 300, or 3000 anywhere in all 22 autosomal chromosomes or 30 in the HLA region of chromosome 6), and the disease heritability h^2 (50 or 80%). Liability scores are computed either from a model with additive effects only ("ADD") or a more complex model that combines additive, dominant and interaction-type effects ("COMP"). For model "ADD," we compute the liability score of the i-th individual as

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

where S_{causal} is the set of causal SNPs, w_j are weights generated from a Gaussian distribution $N(0, h^2/|S_{\text{causal}}|)$ or a Laplace distribution $Laplace(0, \sqrt{h^2/(2|S_{\text{causal}}|)})$, $G_{i,j}$ is the allele count of individual i for SNP j, $\widetilde{G_{i,j}}$ corresponds to its standardized version (zero mean and unit variance for all SNPs), and ϵ follows a Gaussian distribution $N(0, 1 - h^2)$. For model "COMP," we simulate liability scores using additive, dominant and interaction-type effects (see Supplemental Materials).

We implement three different simulation scenarios, summarized in Table 1. Scenario N°1 uses the whole dataset (all 22 autosomal chromosomes – 281,122 SNPs) and a training set of size 6000. For each combination of the remaining parameters, results are based on 100 simulations except when comparing PLR with T-Trees, which relies on five simulations only because of a much higher computational burden of T-Trees as compared to other methods. Scenario N°2 consists of 100 simulations per combination of parameters on a dataset composed of chromosome six only (18,941 SNPs).

Reducing the number of SNPs increases the polygenicity (the proportion of causal SNPs) of the simulated models. Reducing the number of SNPs (p) is also equivalent to increasing the sample size (n) as predictive power increases as a function of n/p (Dudbridge 2013; Vilhjálmsson *et al.* 2015). For this scenario, we use the additive model only, but continue to vary all other simulation parameters. Finally, scenario N°3 uses the whole dataset as in scenario N°1 while varying the size of the training set in order to assess how the sample size affects predictive performance of methods. A total of 100 simulations per combination of parameters are run using 300 causal SNPs randomly chosen on the genome.

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) (Lusted 1971; Fawcett 2006). In the case of our study, the AUC is the probability that the PRS of a case is greater than the PRS of a control. This measure indicates the extent to which we can distinguish between cases and controls using PRS. As a second measure, we also report the partial AUC for specificities between 90 and 100% (McClish 1989; Dodd and Pepe 2003). This measure is similar to the AUC, but focuses on high specificities, which is the most useful part of the ROC curve in clinical settings. When reporting AUC results of simulations, we also report maximum achievable AUC values of 84% and 94% for heritabilities of 50% and 80%, respectively. These estimates are based on three different yet consistent estimations (see Supplemental Materials).

Methods compared

In this paper, we compare three different types of methods: the C+T method, T-Trees and PLR.

The C+T method directly derives PRS from the results of Genome-Wide Associations Studies (GWAS). In GWAS, a coefficient of regression (*i.e.*, the estimated effect size $\hat{\beta}_j$) is learned independently for each SNP j along with a corresponding P-value p_j. The SNPs are first clumped (C) so that there remain only loci that are weakly correlated with one another (this set of SNPs is denoted S_{clumping}). Then, thresholding (T) consists in removing SNPs with P-values larger than a user-defined threshold p_T. Finally, the PRS for individual i is defined as the sum of allele counts of the remaining SNPs weighted by the corresponding effect coefficients

$$\text{PRS}_\text{i} = \sum_{\substack{j \in S_{\text{clumping}} \\ p_j < p_T}} \hat{\beta}_\text{j} \cdot \text{G}_\text{i,j},$$

where $\hat{\beta}_j$ (p_j) are the effect sizes (P-values) learned from the GWAS. In this study, we mostly report scores for a clumping threshold at $r^2 > 0.2$ within regions of 500 kb, but we also investigate thresholds of 0.05 and 0.8. We report three different scores of prediction: one including all the SNPs remaining after clumping (denoted "C+T-all"), one including only the SNPs remaining after clumping and that have a P-value under the GWAS threshold of significance ($P < 5 \cdot 10^{-8}$, "C+T-stringent"), and one that maximizes the AUC ("C+T-max") for 102 P-value thresholds between 1 and 10^{-100} (Table S2). As we report the optimal threshold based on the test set, the AUC for "C+T-max" is an upper bound of the AUC for the C+T method. Here, the GWAS part uses the training set while clumping uses the test set (all individuals not included in the training set).

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 (Botta *et al.* 2014). We use the same parameters as reported in table 4 of Botta *et al.* (2014), except that we use 100 trees instead of 1000. Using 1000 trees provides a minimal increase of AUC while requiring a disproportionately long processing time (*e.g.*, AUC of 81.5% instead of 81%, data not shown).

Finally, for PLR, we find regression coefficients β_0 and β that minimize the following regularized loss function

$$\begin{aligned} L(\lambda, \alpha) = & \underbrace{-\sum_{i=1}^{n} (y_i \log(z_i) + (1-y_i)\log(1-z_i))}_{\text{Loss function}} \\ & + \underbrace{\lambda \left((1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1 \right)}_{\text{Penalization}}, \end{aligned} \tag{1}$$

where $z_i = 1/(1+\exp(-(\beta_0 + x_i^T\beta)))$, x denotes the genotypes and covariables (*e.g.*, principal components), y is the disease status to predict, λ and α are two regularization hyper-parameters that need to be chosen. 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 equal to zero and can be used as a means of variable selection, leading to sparse 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 $p \gg n$ situation (p is the number of SNPs), or any situation involving many correlated predictors (corresponds to $0 < \alpha < 1$) (Friedman *et al.* 2010). In this study, we use a grid search over $\alpha \in \{1, 0.5, 0.05, 0.001\}$. This grid-search is directly embedded in our PLR implementation for simplicity. Using $\alpha = 0.001$ should result in a model very similar to gBLUP.

To fit PLR, we use an efficient algorithm (Friedman *et al.* 2010; Tibshirani *et al.* 2012; Zeng and Breheny 2017) from which we derived our own implementation in R package bigstatsr. This algorithm builds predictions for many values of λ, which is called a "regularization path." To obtain an algorithm that does not require to choose this hyper-parameter λ, we developed a procedure that we call Cross-Model

Table 1 Summary of all simulations

Number of scenario	Dataset (number of SNPs)	Sample size of training set	Causal SNPs (number and location)	Distribution of effects	Heritability	Simulation model	Methods
1	All 22 chromosomes (281,122 SNPs)	6000	30 in HLA 30 in all 300 in all 3000 in all	Gaussian Laplace	0.5 0.8	ADD COMP	C+T PLR PLR3 (T-Trees)
2	Chromosome 6 only (18,941 SNPs)	—[a]	—[a]	—[a]	—[a]	ADD	C+T PLR
3	All 22 chromosomes (281,122 SNPs)	1000 2000 3000 4000 5000	300 in all	—[a]	—[a]	—[a]	—[a]

[a] Parameters are the same as the ones in the upper box.

Selection and Averaging (CMSA, Figure S1). Because of L1-regularization, the resulting vector of estimated effect sizes is sparse. We refer to this method as "PLR" in the results section.

To capture recessive and dominant effects on top of additive effects in PLR, we use simple feature engineering: we construct a separate dataset with three times as many variables as the initial one. For each SNP variable, we add two more variables coding for recessive and dominant effects: one variable is coded 1 if homozygous variant and 0 otherwise, and the other is coded 0 for homozygous referent and 1 otherwise. We then apply our PLR implementation to this dataset with three times as many variables as the initial one; we refer to this method as "PLR3" in the rest of the paper.

Evaluating predictive performance for celiac data

We use Monte Carlo cross-validation to compute AUC, partial AUC, the number of predictors, and execution time for the original Celiac dataset with the observed case-control status: we randomly split 100 times the dataset in a training set of 12,000 individuals and a test set composed of the remaining 3155 individuals.

Data availability

Supplemental Data include a PDF with two sections of methods, two tables and eight figures. Supplemental data also include six HTML R notebooks including all code and results used in this paper, for reproducibility purposes, and available at https://figshare.com/articles/code/7178750. Additional analyses of the UK Biobank are available as three R scripts at https://figshare.com/articles/code_UKB/7531559. Results of simulations are available at https://figshare.com/articles/results_zip/7126964. A tutorial on how to start with R packages bigstatsr and bigsnpr is available at https://privefl.github.io/bigsnpr/articles/demo.html. The two R packages are available on GitHub.

Results

Joint estimation improves predictive performance

We compared PLR with the C+T method using simulations of scenario N°1 (Table 1). When simulating a model with 30 causal SNPs and a heritability of 80%, PLR provides AUC of 93%, nearly reaching the maximum achievable AUC of 94% for this setting (Figure 1). Moreover, PLR consistently provides higher predictive performance than C+T across all scenarios considered, except in some cases of high polygenicity and small sample size, where all methods perform poorly (AUC values below 60% – Figure 1 and Figure 3). PLR provides particularly higher predictive performance than C+T when there are correlations between predictors, *i.e.*, when we choose causal SNPs to be in the HLA region. In this situation, the mean AUC reaches 92.5% for PLR and 84% for "C+T-max" (Figure 1). For the simulations, we do not report results in terms of partial AUC because partial AUC values have a Spearman correlation of 98% with the AUC results for all methods (Figure S3).

Importance of hyper-parameters

In practice, a particular value of the threshold of inclusion of SNPs should be chosen for the C+T method, and this choice can dramatically impact the predictive performance of C+T. For example, in a model with 30 causal SNPs, AUC ranges from $<60\%$ when using all SNPs passing clumping to 90% *if* choosing the optimal *P*-value threshold (Figure S4).

Concerning the r^2 threshold of the clumping step in C+T, we mostly used the common value of 0.2. Yet, using a more stringent value of 0.05 provides equal or higher predictive performance than using 0.2 in most of the cases we considered (Figure 2 and Figure 3).

Our implementation of PLR that automatically chooses hyper-parameter λ provides similar predictive performance than the best predictive performance of 100 models corresponding to different values of λ (Figure S8).

Nonlinear effects

We tested the T-Trees method in scenario N°1. As compared to PLR, T-Trees perform worse in terms of predictive ability, while taking much longer to run (Figure S5). Even for simulations with model "COMP" in which there are dominant and interaction-type effects that T-Trees should be able to handle,

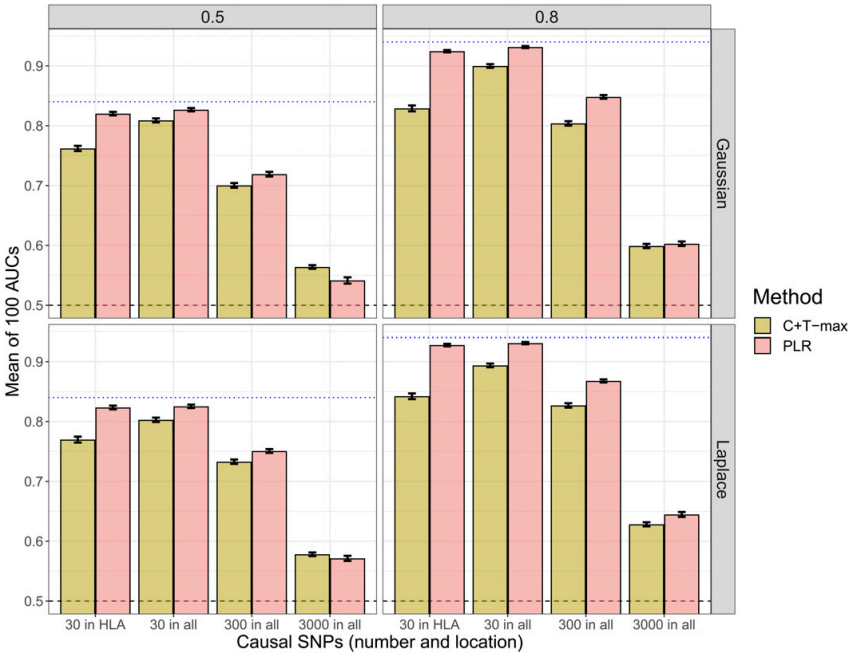


Figure 1 Main comparison of C+T and PLR when simulating phenotypes with additive effects (scenario N°1, model "ADD"). Mean AUC over 100 simulations for PLR and the maximum AUC reported with "C+T-max" (clumping threshold at $r^2 > 0.2$). Upper (lower) panels present results for effects following a Gaussian (Laplace) distribution, and left (right) panels present results for a heritability of 0.5 (0.8). Error bars are representing ± 2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC.

AUC is still lower when using T-Trees than when using PLR (Figure S5).

We also compared the two PLRs in scenario N°1: PLR *vs.* PLR3 that uses additional features (variables) coding for recessive and dominant effects. Predictive performance of PLR3 are nearly as good as PLR when there are additive effects only (differences of AUC are always <2%) and can lead to significantly greater results when there are also dominant and interactions effects (Figures S6 and S7). For model "COMP," PLR3 provides AUC values at least 3.5% higher than PLR, except when there are 3000 causal SNPs. Yet, PLR3 takes two to three times as much time to run and requires three times as much disk storage as PLR.

Simulations varying number of SNPs and sample size

First, when reproducing simulations of scenario N°1 using chromosome six only (scenario N°2), the predictive performance of PLR always increase (Figure 2). There is a particularly large increase when simulating 3000 causal SNPs: AUC from PLR increases from 60% to nearly 80% for Gaussian effects and a disease heritability of 80%. On the contrary, when simulating only 30 or 300 causal SNPs with the corresponding dataset, AUC of "C+T-max" does not increase, and even decreases for a heritability of 80% (Figure 2). Second, when varying the training size (scenario N°3), we report an increase of AUC with a larger training size, with a faster increase of AUC for PLR as compared to "C+T-max" (Figure 3).

Polygenic scores for celiac disease

Joint PLRs also provide higher AUC values for the Celiac data: 88.7% with PLR and 89.1% with PLR3 as compared to 82.5% with "C+T-max" (Figure S2 and Table 2). The relative increase in partial AUC, for specificities larger than 90%, is even larger (42 and 47%) with partial AUC values of 0.0411, 0.0426, and 0.0289 obtained with PLR, PLR3, and "C+T-max," respectively. Moreover, logistic regressions use less predictors, respectively, at 1570, 2260, and 8360. In terms of computation time, we show that PLR, while learning jointly on all SNPs at once and testing four different values for hyper-parameter α, is almost as fast as the C+T method (190 *vs.* 130 sec), and PLR3 takes less than twice as long as PLR (296 *vs.* 190 sec).

Polygenic scores for the UK Biobank

We tested our implementation on 656K genotyped SNPs of the UK Biobank, keeping only Caucasian individuals and removing related individuals (excluding the second individual in each pair with a kinship coefficient >0.08). Results are presented in Table 3.

Our implementation of L1-penalized linear regression runs in <1 day for 350K individuals (training set), achieving a correlation of >65.5% with true height for each sex in the remaining 24K individuals (test set). By comparison, the best C+T model achieves a correlation of 55% for women and 56% for men (in the test set), and the GWAS part takes 1 hr (for the training set). If using only the top 100,000 SNPs from a GWAS on the training set to fit our L1-PLR,

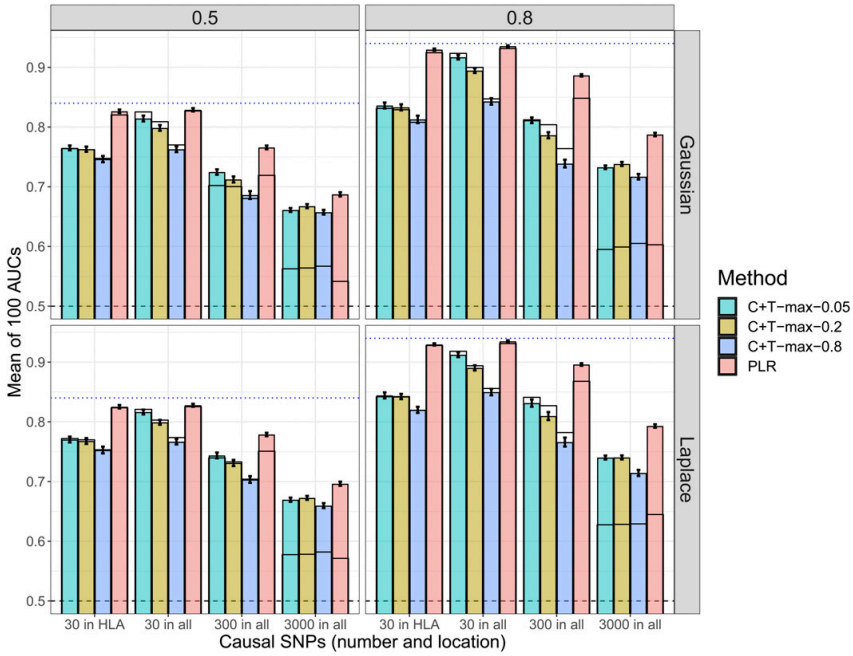


Figure 2 Comparison of methods when simulating phenotypes with additive effects and using chromosome six only (scenario N°2). Thinner lines represent results in scenario N°1. Mean AUC over 100 simulations for PLR and the maximum values of C+T for three different r^2 thresholds (0.05, 0.2, and 0.8) as a function of the number and location of causal SNPs. Upper (lower) panels present results for effects following a Gaussian (Laplace) distribution and left (right) panels present results for a heritability of 0.5 (0.8). Error bars representing ±2SD of 10^5 nonparametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC.

correlation between predicted and true heights drops at 63.4% for women and 64.3% for men. Our L1-PLR on breast cancer runs in 13 min for 150K women, achieving an AUC of 0.598 in the remaining 39K women, while the best C+T model achieves an AUC of 0.589, and the GWAS part takes 15 hr.

Discussion

Joint estimation improves predictive performance

In this comparative study, we present a computationally efficient implementation of PLR. This model can be used to build PRS based on very large individual-level SNP datasets such as the UK biobank (Bycroft *et al.* 2018). In agreement with previous work (Abraham *et al.* 2013), we show that jointly estimating SNP effects has the potential to substantially improve predictive performance as compared to the standard C+T approach in which SNP effects are learned independently. PLR always outperforms the C+T method, except in some highly underpowered cases (AUC values always <0.6), and the benefits of using PLR are more pronounced with an increasing sample size or when causal SNPs are correlated with one another.

When there are many small effects and a small sample size, PLR performs worse than (the best result for) C+T. For example, this situation occurs when there are many causal variants (3K) to distinguish among many typed variants (280K) while using a small sample size (6K). In such underpowered scenarios, it is difficult to detect true causal variants, which makes PLR too conservative, whereas the best strategy is to include nearly all SNPs (Purcell *et al.* 2009).

When increasing sample size (scenario N°3), PLR achieves higher predictive performance than C+T and the benefits of using PLR over C+T increase with an increasing sample size (Figure 3). Moreover, when decreasing the search space (total number of candidate SNPs) in scenario N°2, we increase the proportion of causal variants and we virtually increase the sample size (Dudbridge 2013). In this scenario N°2, even when there are small effects and a high polygenicity (3000 causal variants out of 18,941), PLR gets a large increase in predictive performance, now consistently higher than C+T (Figure 2).

Importance of hyper-parameters

The choice of hyper-parameter values is very important since it can greatly impact the performance of methods. In the C+T method, there are two main hyper-parameters: the r^2 and the p_T thresholds that control how stringent are the C+T steps. For the clumping step, appropriately choosing the r^2 threshold is important. Indeed, on the one hand, choosing a low value for this threshold may discard informative SNPs that are correlated. On the other hand, when choosing a high value for this threshold, too much redundant information is included in the model, which adds noise to the PRS. Based on the simulations, we find that using a stringent threshold ($r^2 = 0.05$) leads to higher predictive performance, even when causal SNPs are correlated. It means that, in most cases tested in this paper, avoiding redundant information in C+T is more important than including all causal SNPs. The choice

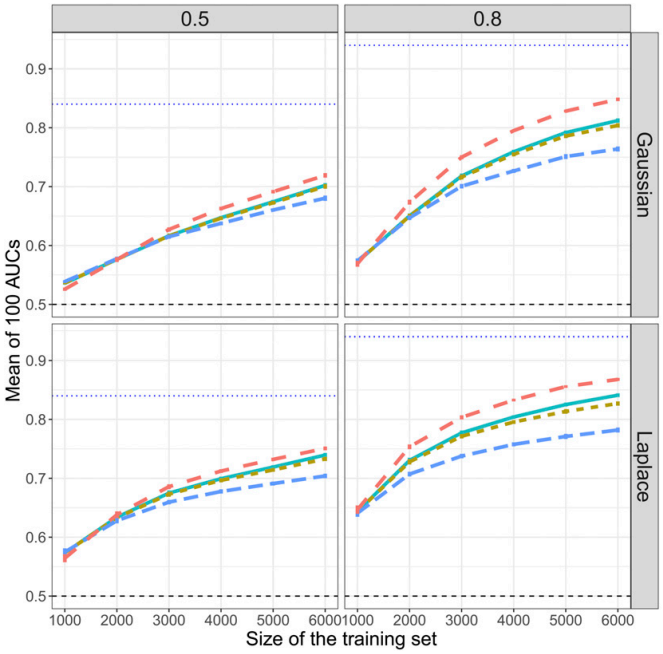


Figure 3 Comparison of methods when simulating 300 causal SNPs with additive effects and when varying sample size (scenario N°3). Mean AUC over 100 simulations for the maximum values of C+T for three different r^2 thresholds (0.05, 0.2, and 0.8) and PLR as a function of the training size. Upper (lower) panels are presenting results for effects following a Gaussian (Laplace) distribution and left (right) panels are presenting results for a heritability of 0.5 (0.8). Error bars represent $\pm$2SD of 10^5 nonparametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC.

of the p_T threshold is also very important as it can greatly impact the predictive performance of the C+T method, which we confirm in this study (Ware *et al.* 2017). In this paper, we reported the maximum AUC of 102 different *P*-value thresholds, a threshold that should normally be learned on the training set only. To our knowledge, there is no clear standard on how to choose these two critical hyper-parameters for C+T. So, for C+T, we report the best AUC value on the test set, even if it leads to overoptimistic results for C+T as compared to PLR.

In contrast, for PLR, we developed an automatic procedure called CMSA that releases investigators from the burden of choosing hyper-parameter λ. Not only this procedure provides near-optimal results, but it also accelerates the model training thanks to the development of an early stopping criterion. Usually, cross-validation is used to choose hyper-parameter values and then the model is trained again with these particular hyper-parameter values (Hastie *et al.* 2008; Wei *et al.* 2013). Yet, performing cross-validation and retraining the model is computationally demanding; CMSA offers a less burdensome alternative. Concerning hyper-parameter α that accounts for the relative importance of the L1 and L2 regularizations, we use a grid search directly embedded in the CMSA procedure.

Nonlinear effects

We also explored how to capture nonlinear effects. For this, we introduced a simple feature engineering technique that enables PLR to detect and learn not only additive effects, but also dominant and recessive effects. This technique improves the predictive performance of PLR when there are nonlinear effects in the simulations, while providing nearly the same predictive performance when there are additive effects only. Moreover, it also improves predictive performance for the celiac disease.

Yet, this approach is not able to detect interaction-type effects. In order to capture interaction-type effects, we tested T-Trees, a method that is able to exploit SNP correlations and interactions thanks to special decision trees (Botta *et al.* 2014). However, predictive performance of T-Trees are consistently lower than with PLR, even when simulating a model with dominant and interaction-type effects that T-Trees should be able to handle.

Time and memory requirements

The computation time of our PLR implementation mainly depends on the sample size and the number of candidate variables (variables that are included in the gradient descent). Indeed, the algorithm is composed of two steps: first, for each variable, the algorithm computes an univariate statistic that is used to decide if the variable is included in the model (for each value of λ). This first step is very fast. Then, the algorithm iterates over a regularization path of decreasing values of λ, which progressively enables variables to enter the model (Figure S1). In the second step, the number of variables increases and computations stop when an early stopping criterion is reached (when prediction is getting worse on the corresponding validation set, see Figure S1).

Table 2 Results for the real celiac dataset

Method	AUC	pAUC	# predictors	Execution time (s)
C+T-max	0.825 (0.000664)	0.0289 (0.000187)	8360 (744)	130 (0.143)
PLR	0.887 (0.00061)	0.0411 (0.000224)	1570 (46.4)	190 (1.21)
PLR3	0.891 (0.000628)	0.0426 (0.000219)	2260 (56.1)	296 (2.03)

The results are averaged over 100 runs where the training step is randomly composed of 12,000 individuals. In the parentheses is reported the SD of 10^5 bootstrap samples of the mean of the corresponding variable. Results are reported with three significant digits.

Table 3 Results for the UK Biobank

Trait	Method	*r* (women/men)	# Predictors	Execution time
Height	PLR	0.656/0.657	115,997	21 hr
Height	C+T-max	0.549/0.561	45,570	69 min
Trait	**Method**	**AUC**	**# Predictors**	**Execution time**
Breast cancer	PLR	0.598	2653	13 min
Breast cancer	C+T-max	0.589	21	15 hr

The sizes of training/test sets for height (resp. breast cancer) are 350,000/24,131 (resp. 150,000/38,628). For height, *r* (correlation between predicted and true heights) is reported within women/men separately; for breast cancer, AUC is reported.

For highly polygenic traits such as height and when using huge datasets such as the UK Biobank, the algorithm might iterate over >100,000 variables, which is computationally demanding. On the contrary, for traits like celiac disease or breast cancer that are less polygenic, the number of variables included in the model is much smaller so that fitting is very fast (only 13 min for 150K women of the UK Biobank for breast cancer).

Memory requirements are tightly linked to computation time. Indeed, variables are accessed in memory thanks to memory-mapping when they are used (Privé *et al.* 2018). When there is not enough memory left, the operating system (OS) frees some memory for new incoming variables. Yet, if too many variables are used in the gradient descent, the OS would regularly swap memory between disk and RAM, severely slowing down computations. A possible approach to reduce computational burden is to apply penalized regression on a subset of SNPs by prioritizing SNPs using univariate tests (GWAS computed from the same dataset). Yet, this strategy was shown to reduce predictive power (Abraham *et al.* 2013; Lello *et al.* 2018), which we also confirm in this paper. Indeed, when using only the 100K most significantly associated SNPs, correlation between predicted and true heights is reduced from 0.656/0.657 to 0.634/0.643 within women/men. A key advantage of our implementation of PLR is that prior filtering of variables is no more required for computational feasibility, thanks to the use of sequential strong rules and early stopping criteria.

Limitations

Our approach has one major limitation: the main advantage of the C+T method is its direct applicability to summary statistics, allowing to leverage the largest GWAS results to date, even when individual cohort data cannot be merged because of practical or legal reasons. Our implementation of PLR does not allow yet for the analysis of summary data, but this represents an important future direction. The current version is of particular interest for the analysis of modern individual-level datasets including hundreds of thousands of individuals.

Finally, in this comparative study, we did not consider the problem of population structure (Vilhjálmsson *et al.* 2015; Márquez-Luna *et al.* 2017; Martin *et al.* 2017), and also did not consider nongenetic data such as environmental and clinical data (Van Vliet *et al.* 2012; Dey *et al.* 2013).

Conclusions

In this comparative study, we have presented a computationally efficient implementation of PLR that can be used to predict disease status based on genotypes. A similar penalized linear regression for quantitative traits is also available in R package bigstatsr. Our approach solves the dramatic memory and computational burdens faced by standard implementations, thus allowing for the analysis of large-scale datasets such as the UK biobank (Bycroft *et al.* 2018).

We also demonstrated in simulations and real datasets that our implementation of penalized regressions is highly effective over a broad range of disease architectures. It can be appropriate for predicting autoimmune diseases with a few strong effects (*e.g.*, celiac disease), as well as highly polygenic traits (*e.g.*, standing height) provided that sample size is not too small. Finally, PLR as implemented in bigstatsr can also be used to predict phenotypes based on other omics data, since our implementation is not specific to genotype data.

Acknowledgments

We are grateful to Félix Balazard for useful discussions about T-Trees, and to Yaohui Zeng for useful discussions about R package biglasso. We are grateful to the two anonymous reviewers who contributed to improving this paper. The authors acknowledge LabEx Pervasive Systems and Algorithms (PERSYVAL)-Lab [Agence Nationale de Recherche (ANR)-11-LABX-0025-01] and ANR project French Regional Origins in Genetics for Health (FROGH) (ANR-16-CE12-0033). The authors also acknowledge the Grenoble Alpes Data Institute, which is supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02). This research was conducted using the UK Biobank Resource under Application Number 25589.

Literature Cited

Abraham, G., A. Kowalczyk, J. Zobel, and M. Inouye, 2012 Sparsnp: fast and memory-efficient analysis of all snps for phenotype prediction. BMC Bioinformatics 13: 88. https://doi.org/10.1186/1471-2105-13-88

Abraham, G., A. Kowalczyk, J. Zobel, and M. Inouye, 2013 Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genet. Epidemiol. 37: 184–195. https://doi.org/10.1002/gepi.21698

Abraham, G., J. A. Tye-Din, O. G. Bhalala, A. Kowalczyk, J. Zobel *et al.*, 2014 Accurate and robust genomic prediction of celiac disease using statistical learning. PLoS Genet. 10: e1004137 (erratum: PLoS Genet. 10: e1004374). https://doi.org/10.1371/journal.pgen.1004137

Botta, V., G. Louppe, P. Geurts, and L. Wehenkel, 2014 Exploiting SNP correlations within random forest for genome-wide association studies. PLoS One 9: e93379. https://doi.org/10.1371/journal.pone.0093379

Breiman, L., 2001 Random forests. Mach. Learn. 45: 5–32. https://doi.org/10.1023/A:1010933404324

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott *et al.*, 2018 The UK biobank resource with deep phenotyping and genomic data. Nature 562: 203–209. https://doi.org/10.1038/s41586-018-0579-z

Chatterjee, N., B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock *et al.*, 2013 Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. 45: 400–405. https://doi.org/10.1038/ng.2579

Chatterjee, N., J. Shi, and M. García-Closas, 2016 Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet. 17: 392–406. https://doi.org/10.1038/nrg.2016.27

Dey, S., R. Gupta, M. Steinbach, and V. Kumar, 2013 Integration of clinical and genomic data: a methodological survey. Technical Report TR13005. Department of Computer Science and Engineering, University of Minnesota.

Dodd, L. E., and M. S. Pepe, 2003 Partial AUC estimation and regression. Biometrics 59: 614–623. https://doi.org/10.1111/1541-0420.00071

Dubois, P. C., G. Trynka, L. Franke, K. A. Hunt, J. Romanos *et al.*, 2010 Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 42: 295–302 (erratum: Nat. Genet. 42: 465). https://doi.org/10.1038/ng.543

Dudbridge, F., 2013 Power and predictive accuracy of polygenic risk scores. PLoS Genet. 9: e1003348 (erratum: PLoS Genet. 9). https://doi.org/10.1371/journal.pgen.1003348

Evans, D. M., P. M. Visscher, and N. R. Wray, 2009 Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum. Mol. Genet. 18: 3525–3531. https://doi.org/10.1093/hmg/ddp295

Falconer, D. S., 1965 The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet. 29: 51–76. https://doi.org/10.1111/j.1469-1809.1965.tb00500.x

Fawcett, T., 2006 An introduction to roc analysis. Pattern Recognit. Lett. 27: 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33: 1. https://doi.org/10.18637/jss.v033.i01

Hastie, T., R. Tibshirani, and J. Friedman, 2008 Model assessment and selection, pp. 219–259 in *The Elements of Statistical Learning*. Springer, New York.

Hoerl, A. E., and R. W. Kennard, 1970 Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12: 55–67. https://doi.org/10.1080/00401706.1970.10488634

Janssens, A. C. J., R. Moonesinghe, Q. Yang, E. W. Steyerberg, C. M. van Duijn *et al.*, 2007 The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. Genet. Med. 9: 528–535. https://doi.org/10.1097/GIM.0b013e31812eece0

Lello, L., S. G. Avery, L. Tellier, A. I. Vazquez, G. de los Campos *et al.*, 2018 Accurate genomic prediction of human height. Genetics 210: 477–497. https://doi.org/10.1534/genetics.118.301267

Lusted, L. B., 1971 Signal detectability and medical decision-making. Science 171: 1217–1219. https://doi.org/10.1126/science.171.3977.1217

Márquez-Luna, C., P.-R. Loh, and A. L. Price, 2017 Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. 41: 811–823. https://doi.org/10.1002/gepi.22083

Martin, A. R., C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale *et al.*, 2017 Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. 100: 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

Mavaddat, N., K. Michailidou, J. Dennis, M. Lush, L. Fachal *et al.*, 2019 Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am. J. Hum. Genet. 104: 21–34. https://doi.org/10.1016/j.ajhg.2018.11.002

McClish, D. K., 1989 Analyzing a portion of the roc curve. Med. Decis. Making 9: 190–195. https://doi.org/10.1177/0272989X8900900307

Okser, S., T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti *et al.*, 2014 Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. 10: e1004754. https://doi.org/10.1371/journal.pgen.1004754

Pashayan, N., S. W. Duffy, D. E. Neal, F. C. Hamdy, J. L. Donovan *et al.*, 2015 Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. Genet. Med. 17: 789–795. https://doi.org/10.1038/gim.2014.192

Privé, F., H. Aschard, A. Ziyatdinov, and M. G. B. Blum, 2018 Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics 34: 2781–2787. https://doi.org/10.1093/bioinformatics/bty185

Purcell, S. M., N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan *et al.*, 2009 Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752. https://doi.org/10.1038/nature08185

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 58: 267–288.

Tibshirani, R., J. Bien, J. Friedman, T. Hastie, N. Simon *et al.*, 2012 Strong rules for discarding predictors in lasso-type problems. J. R. Stat. Soc. Series B Stat. Methodol. 74: 245–266. https://doi.org/10.1111/j.1467-9868.2011.01004.x

Van Vliet, M. H., H. M. Horlings, M. J. Van De Vijver, M. J. Reinders, and L. F. Wessels, 2012 Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. PLoS One 7: e40358. https://doi.org/10.1371/journal.pone.0040358

Vilhjálmsson, B. J., J. Yang, H. K. Finucane, A. Gusev, S. Lindström *et al.*, 2015 Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. 97: 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Ware, E. B., L. L. Schmitz, J. D. Faul, A. Gard, C. Mitchell *et al.*, 2017 Heterogeneity in polygenic scores for common human traits. bioRxiv 106062.

Wei, Z., K. Wang, H.-Q. Qu, H. Zhang, J. Bradfield *et al.*, 2009 From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 5: e1000678. https://doi.org/10.1371/journal.pgen.1000678

Wei, Z., W. Wang, J. Bradfield, J. Li, C. Cardinale *et al.*, 2013 Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am. J. Hum. Genet. 92: 1008–1012. https://doi.org/10.1016/j.ajhg.2013.05.002

Wray, N. R., M. E. Goddard, and P. M. Visscher, 2007 Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 17: 1520–1528. https://doi.org/10.1101/gr.6665407

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common snps explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569. https://doi.org/10.1038/ng.608

Zeng, Y., and P. Breheny, 2017 The biglasso package: a memory- and computation-efficient solver for lasso model fitting with big data in R. arXiv:1701.05936.

Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Communicating editor: N. Wray

Supplementary Material and Methods

Model "COMP"

For model "COMP", we separate the causal SNPs in three equal sets $S^{(1)}_{\text{causal}}$, $S^{(2)}_{\text{causal}}$ and $S^{(3)}_{\text{causal}}$; $S^{(3)}_{\text{causal}}$ is further separated in two equal sets, $S^{(3.1)}_{\text{causal}}$ and $S^{(3.2)}_{\text{causal}}$. We then compute

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

where w_j are weights generated from a Gaussian or a Laplace distribution, $G_{i,j}$ is the allele count of individual i for SNP j, $\widetilde{G_{i,j}}$ corresponds to its standardized version (zero mean and unit variance for all SNPs), $D_{i,j} = \mathbb{1}\{G_{i,j} \neq 0\}$, ϵ follows a Gaussian distribution $N(0, 1-h^2)$ and $S^{(q)}_{\text{causal}} = \left\{e^{(q)}_k,\ k \in \left\{1, \ldots, \left|S^{(q)}_{\text{causal}}\right|\right\}\right\}$.

Maximum AUCs

We use three different ways to estimate the maximum achievable AUC for simulations (see supplementary notebook "oracle").

First, we use the estimation from equation (3) of Wray *et al.* (2010). For a prevalence fixed at 30% and an heritability of 50% (respectively 80%), the approximated theoretical values of AUC are 84.1% (respectively 93.0%). Note that this approximation is reported to be less accurate for high heritabilities.

Secondly, if we assume that the genetic part of the liabilities follows a Gaussian distribution $N(0, h^2)$ and that the environmental part follows a Gaussian distribution $N(0, 1-h^2)$, we can estimate the theoretical value of the AUC that can be achieved given the disease heritability h^2 (and prevalence K) through Monte Carlo simulations. We report AUCs of 84.1% and 94.1% for heritabilities of 50% and 80%, respectively.

Thirdly, we reproduce the exact same procedure of simulations and, for each combination of parameters (Table 2), we estimate the AUC of the "oracle", i.e. the true simulated genetic part of the liabilities, through 100 replicates. For every combination of parameters, AUC values of oracles vary between 83.2% and 84.2% for an heritability of 50% and between 93.2% and 94.1% for an heritability of 80%.

Given all these estimates of maximal achievable AUC and for the sake of simplicity, we report maximum AUCs of 84% (94%) for heritabilities of 50% (80%) whatever are the simulation parameters.

References

Sachs, M. C. *et al.* (2017). plotroc: A tool for plotting roc curves. *Journal of Statistical Software*, **79**(c02).

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

Zeng, Y. and Breheny, P. (2017). The biglasso package: A memory-and computation-efficient solver for lasso model fitting with big data in R. *arXiv preprint arXiv:1701.05936*.

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 in the celiac disease case-control study (after quality control, genotyped on 281,122 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 for the C+T method for this study.

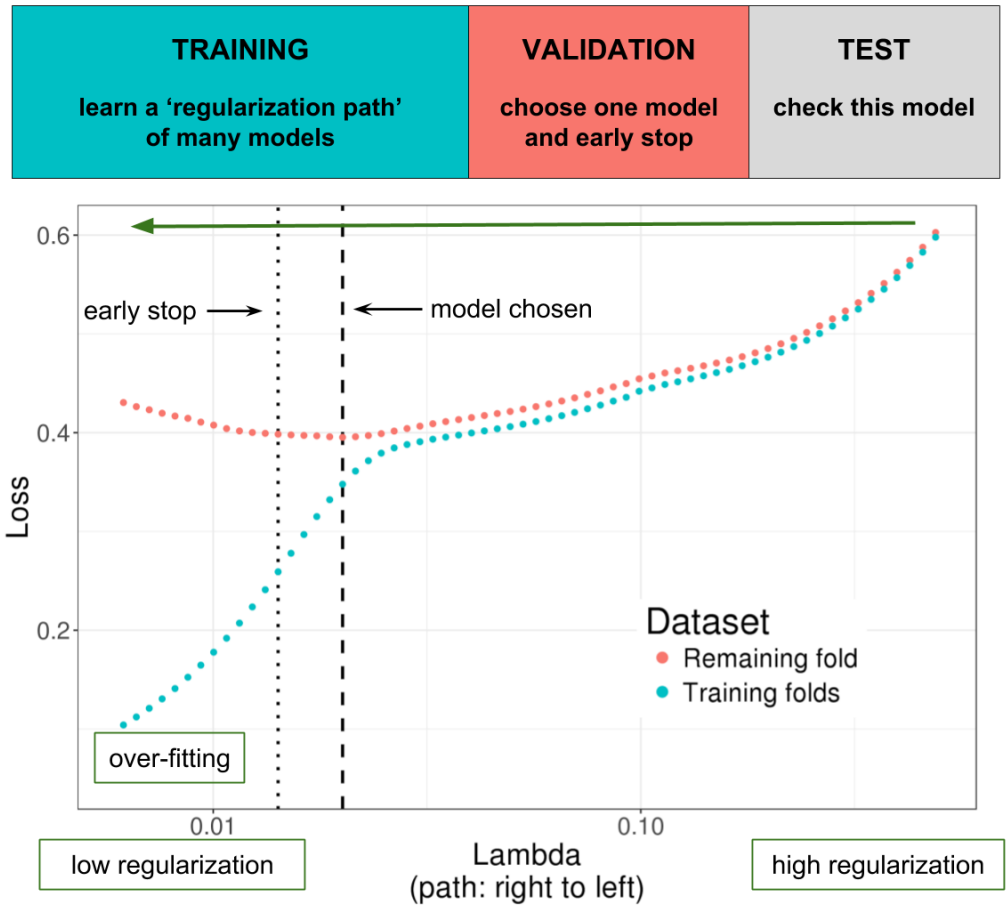


Figure S1: Illustration of one turn of the Cross-Model Selection and Averaging (CMSA) procedure. First, this procedure separates the training set in K folds (e.g. 10 folds). Secondly, in turn, each fold is considered as an inner validation set (red) and the other $(K-1)$ folds form an inner training set (blue). A "regularization path" of models is trained on the inner training set and the corresponding predictions (scores) for the inner validation set are computed. The model that minimizes the loss on the inner validation set is selected. Finally, the K resulting models are averaged. We also use this procedure to derive an early stopping criterion so that the algorithm does not need to evaluate the whole regularization paths, making this procedure much faster.

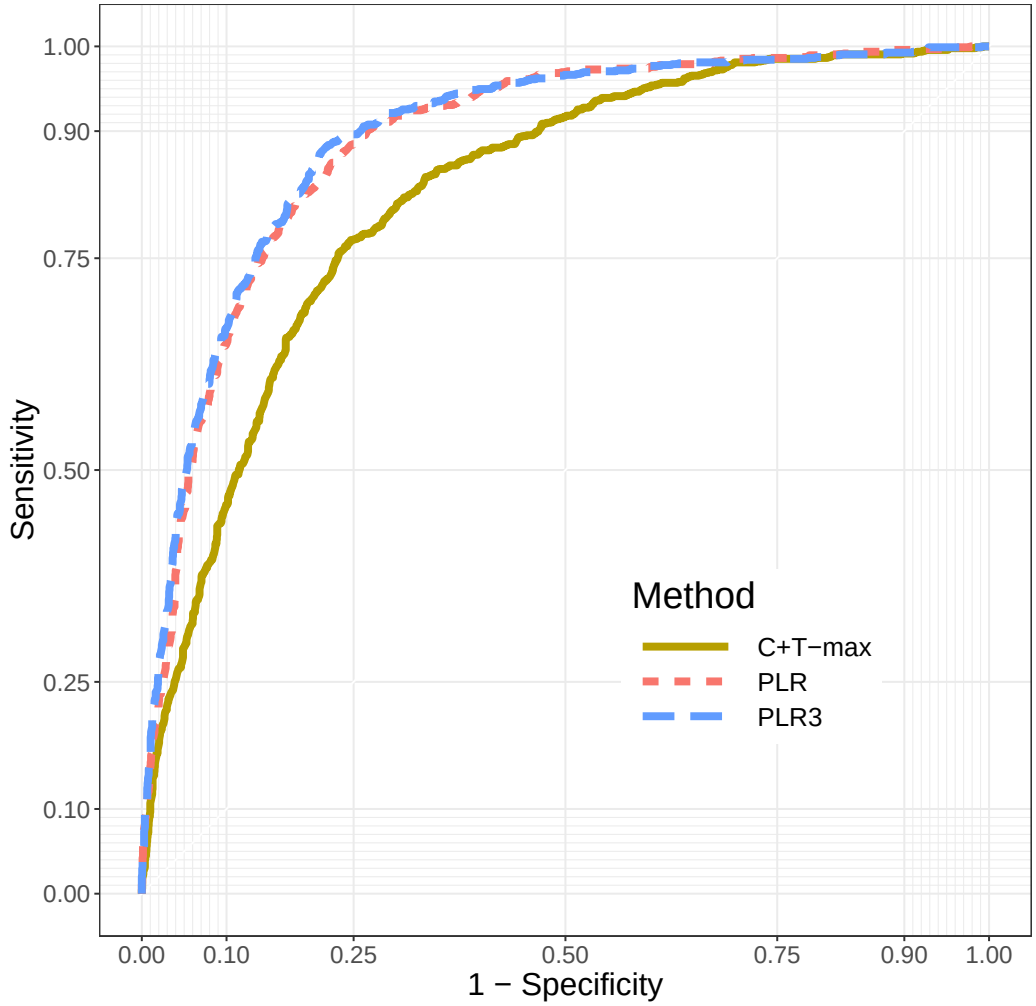


Figure S2: ROC Curves for C+T-max, PLR and PLR3 for the celiac disease dataset. Models were trained using 12,000 individuals. These are results projecting these models on the remaining 3155 individuals. The figure is plotted using R package plotROC (Sachs *et al.* 2017).

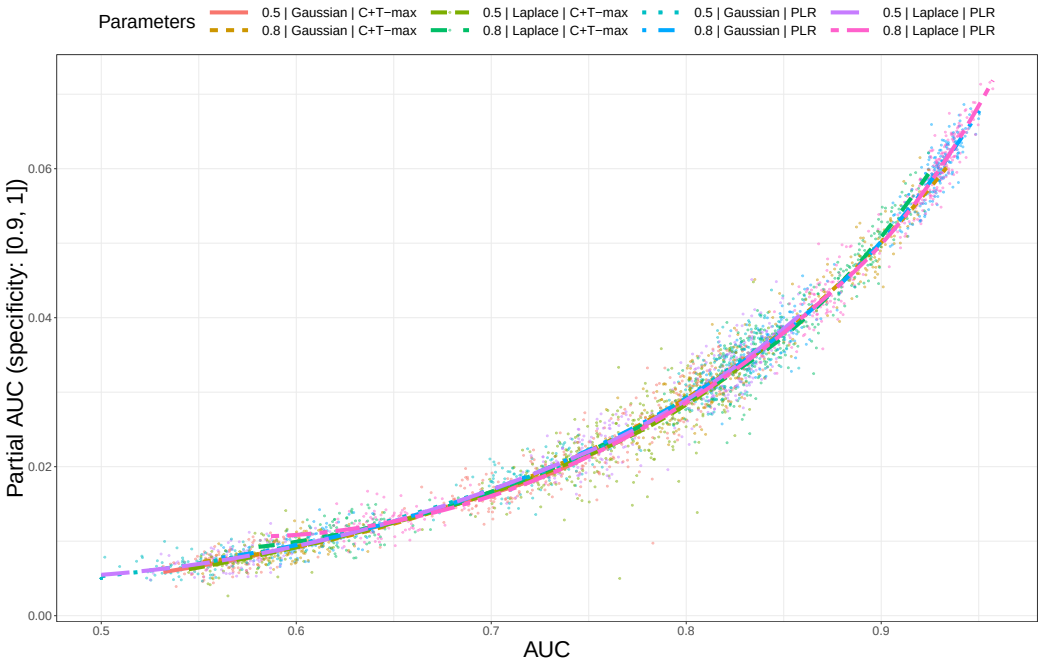


Figure S3: Correlation between AUC and partial AUC values in scenario №1. There is a Spearman correlation of 98% between values of AUC and partial AUC. The relation between the two values are the same whatever are the disease heritability, distribution of effects and method used.

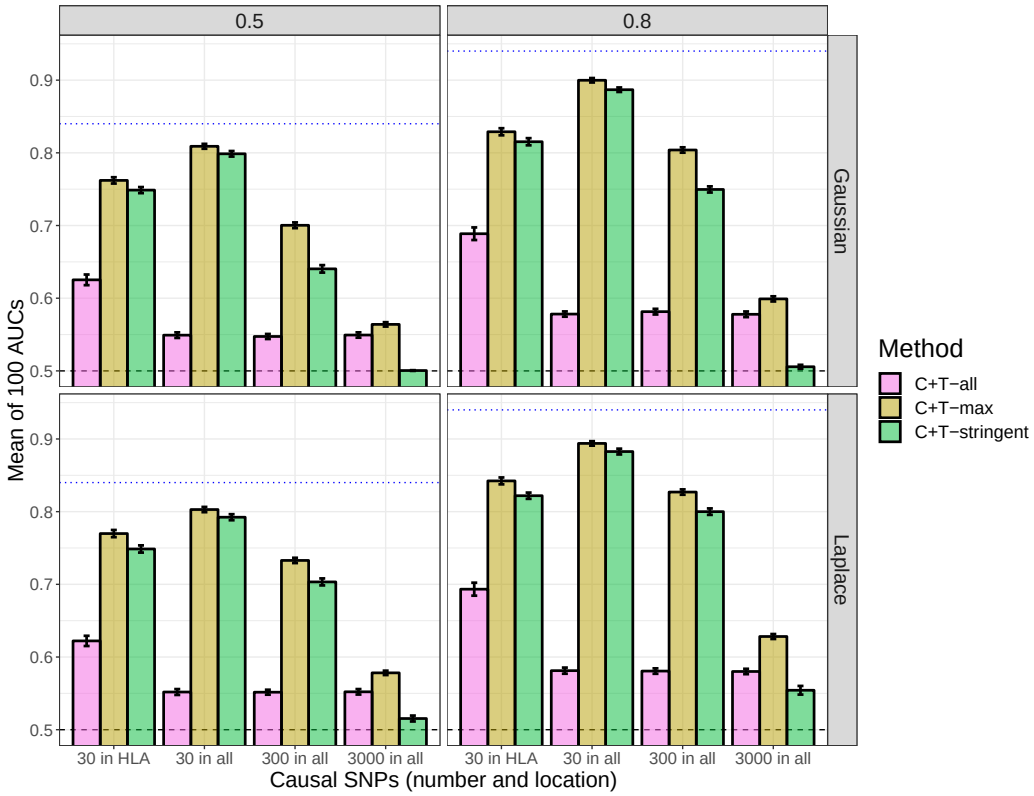


Figure S4: Comparison of three different p-value thresholds used in the C+T method in scenario №1 for model "ADD". Mean AUC over 100 simulations. Upper (lower) panels are presenting results for effets following a Gaussian (Laplace) distribution and left (right) panels are presenting results for an heritability of 0.5 (0.8). Error bars are representing $\pm$2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC.

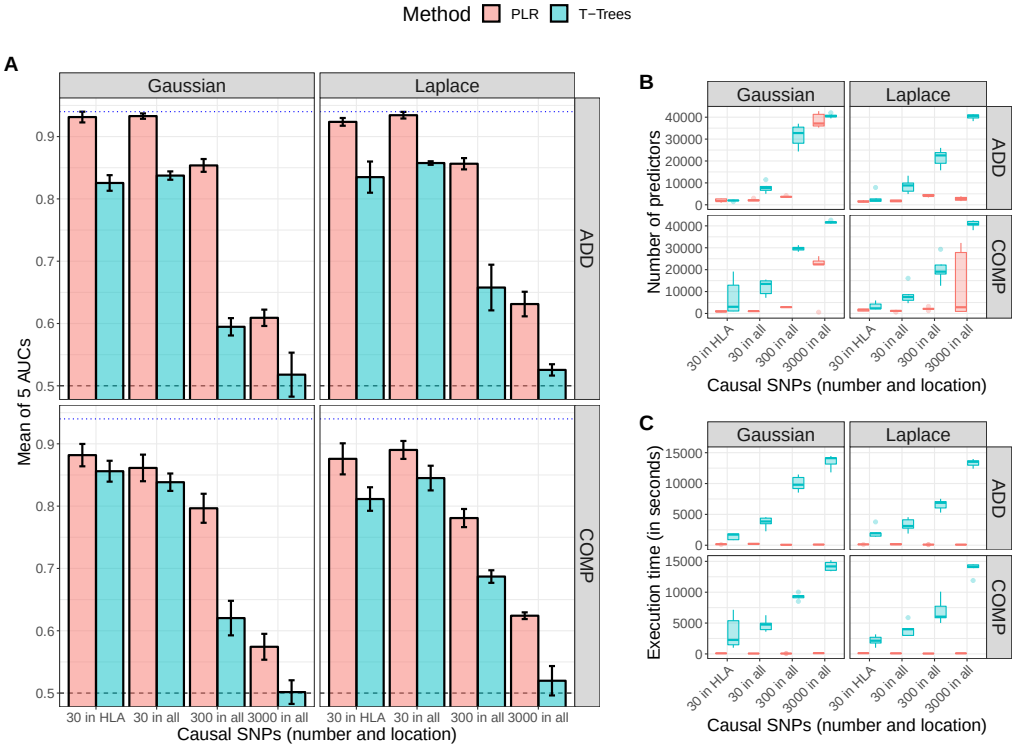


Figure S5: Comparison of T-Trees and PLR in scenario №1 for an heritability of 80%. Vertical panels are presenting results for effects following a Gaussian or Laplace distribution. Horizontal panels are presenting results for models "ADD" and "COMP" for simulating phenotypes. **A:** Mean AUC over 5 simulations. Error bars are representing ± 2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC. **B:** Boxplots of numbers of predictors used by the methods for 5 simulations. **C:** Boxplots of execution times for 5 simulations.

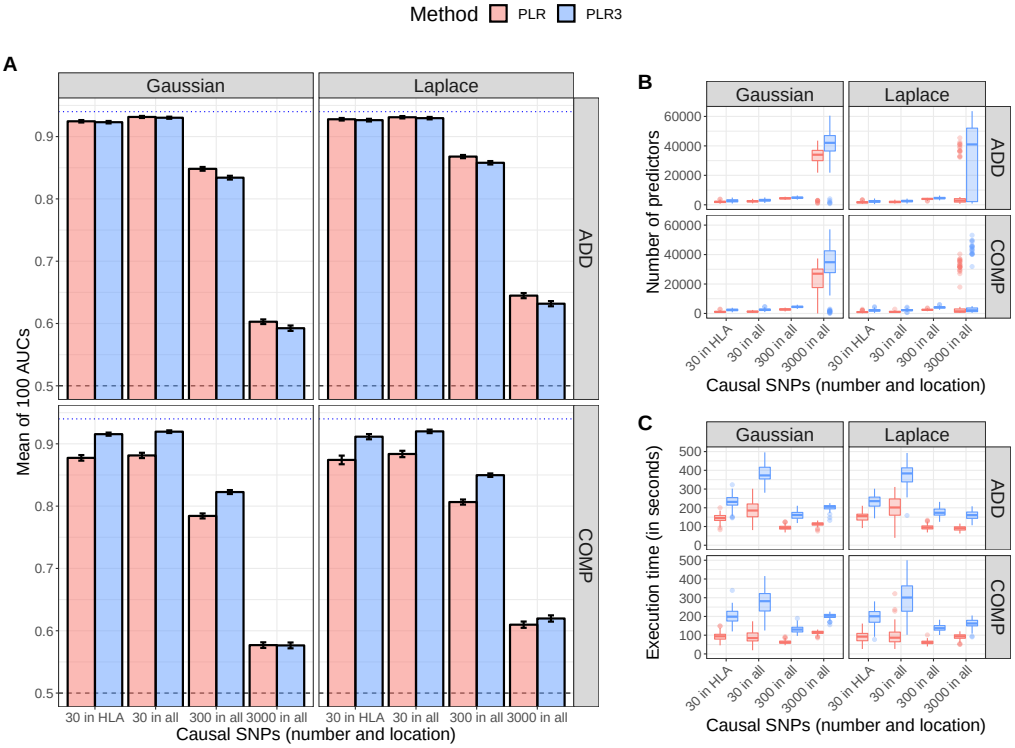


Figure S6: Comparison of PLR3 and PLR in scenario №1 for an heritability of 80%. Vertical panels are presenting results for effects following a Gaussian or Laplace distribution. Horizontal panels are presenting results for models "ADD" and "COMP" for simulating phenotypes. **A:** Mean AUC over 100 simulations. Error bars are representing ± 2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC. **B:** Boxplots of numbers of predictors used by the methods for 100 simulations. **C:** Boxplots of execution times for 100 simulations.

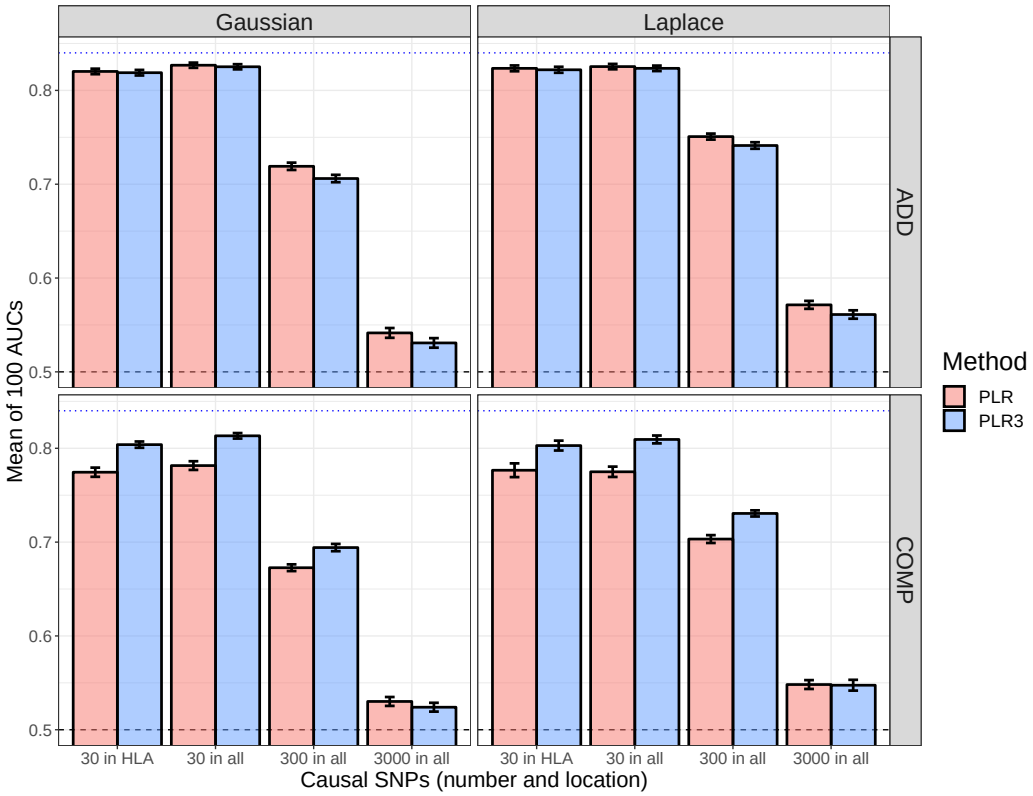


Figure S7: Comparison of PLR3 and PLR in scenario №1 for an heritability of 50%. Vertical panels are presenting results for effects following a Gaussian or Laplace distribution. Horizontal panels are presenting results for models "ADD" and "COMP" for simulating phenotypes. **A:** Mean AUC over 100 simulations. Error bars are representing ± 2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC.

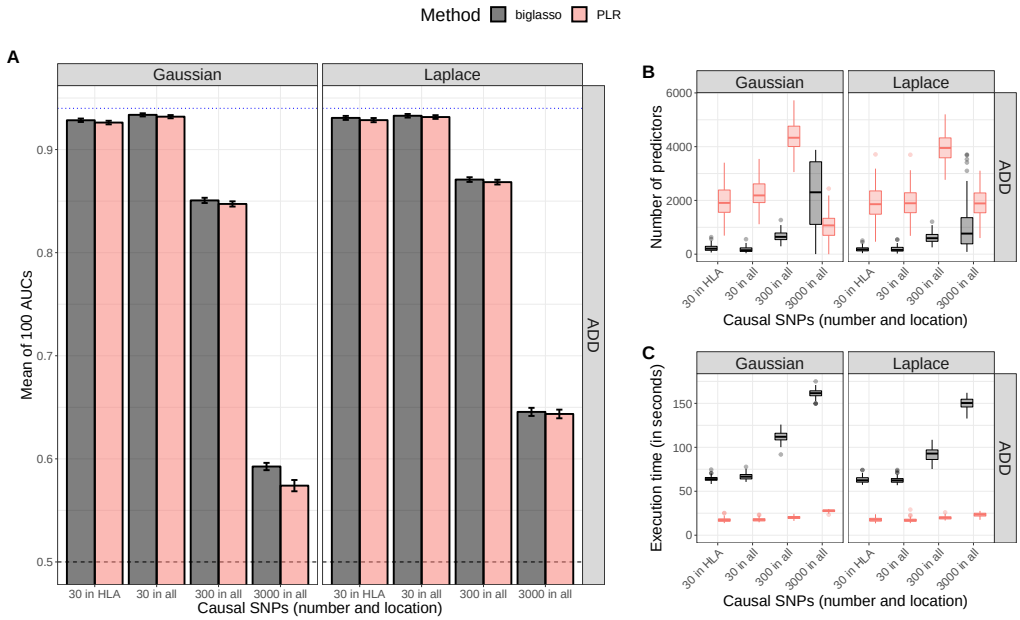


Figure S8: Comparison of PLR and the best prediction (among 100 tested λ values) for "biglasso" (another implementation of penalized logistic regression – Zeng and Breheny (2017)) in scenario №1. Simulations use model "ADD", an heritability of 80% and $\alpha = 1$. Vertical panels are presenting results for effects following a Gaussian or Laplace distribution. **A:** Mean AUC over 100 simulations. Error bars are representing $\pm$2SD of 10^5 non-parametric bootstrap of the mean AUC. The blue dotted line represents the maximum achievable AUC. **B:** Boxplots of numbers of predictors used by the methods for 100 simulations. **C:** Boxplots of execution times for 100 simulations.

Chapter 4

Making the most of Clumping and Thresholding for polygenic scores

4.1 Summary of the article

4.1.1 Introduction

Most of the time, only summary statistics for a GWAS dataset are available, i.e. the estimated effect sizes and p-values for each variant of the dataset. Because of the availability of such data en masse, specific methods using those summary data have been developed for a wide range of applications (Pasaniuc *et al.*, 2014; Vilhjálmsson *et al.*, 2015; Bulik-Sullivan *et al.*, 2015; Pasaniuc and Price, 2017; Speed and Balding, 2018). Moreover, methods using summary statistics data are usually fast and easy to use, making them even more appealing to researchers. One of these summary statistics based methods applicable for polygenic prediction is Clumping and Thresholding (C+T). When only limited sample size of individual-level data are available (as opposed to summary statistics), C+T provides a competitive method for deriving predictive polygenic risk scores (Privé *et al.*, 2019).

C+T is the simplest and most widely-used method for constructing PRS based on summary statistics and has been used for many years now. The idea behind C+T is simple because it directly uses weights learned from GWAS; it further removes SNPs as one does when reporting hits from GWAS, i.e. only SNPs that pass the genome-wide

threshold (p-value thresholding) and that are independent association findings (clumping) are reported. In GWAS, it is commonly accepted to use a p-value threshold of 5×10^{-8} when reporting significant findings, yet for prediction purposes, including less significant SNPs can substantially improve predictive performance (Purcell *et al.*, 2009).

Therefore, when using C+T, one has to choose a p-value threshold that balances between removing informative variants when using a stringent p-value threshold and adding too much noise in the score by including too many variants with no effect. The clumping step aims at removing redundancy in included effects that is simply due to linkage disequilibrium (LD) between variants. Yet, clumping may as well remove independently predictive variants in nearby regions; to balance this, C+T uses as hyper-parameter a threshold on correlation between variants included. Thus, C+T users must choose hyper-parameters of C+T well if they want to maximize predictive performance of the polygenic score derived. Most of the time, people use default values for these parameters, expect for the p-value threshold, for which they look at different values and choose the one maximizing predictive ability in a training set.

4.1.2 Methods

We implement an efficient way to compute many C+T scores corresponding to many different sets of hyper-parameters for C+T. This is now part of R package bigsnpr (Privé *et al.*, 2018). The 4 parameters we vary are the correlation threshold of clumping, the window size for looking at correlation, the p-value threshold and the imputation accuracy threshold when using imputed variants. In total, we investigate 5600 different sets of hyper-parameters for C+T.

We also derive C+T scores for each chromosome separately, resulting in 123,200 different scores. We propose to use stacking, i.e. we fit a penalized regression of these scores and learn an optimal linear combination of those scores instead of only choosing the best one (Breiman, 1996). We hypothesize that Stacked Clumping and Thresholding (SCT) has the potential to make C+T more flexible and to increase its predictive performance. Moreover, SCT results in a linear model from which we can derive an unique vector of coefficients to be used for testing in unseen individuals.

4.1.3 Results

We test 6 different simulation scenarios using the UK Biobank dataset. We also derive PRS for 8 common diseases using external summary statistics from published GWAS and dividing the UK Biobank data into training and test sets. Investigating more hyper-parameters for C+T (we call this maxCT) instead of using default values for these hyper-parameters (we call this stdCT) consistently improves predictive performance in simulations and real data applications. This makes C+T competitive to state-of-the-art methods like lassosum (Mak *et al.*, 2017). Moreover, SCT often provides substantial predictive performance improvement over maxCT by using different weights from those reported from the GWAS.

4.1.4 Discussion

We provide an efficient way to compute C+T scores for many different hyper-parameters values in R package bigsnpr. Investigating 8 different diseases, we show that the optimal C+T hyper-parameters for those traits are very different, probably because these diseases have different architectures. Therefore, fine-tuning hyper-parameters of C+T improves its predictive performance as compared to using default values for clumping.

Instead of choosing one set of hyper-parameters that maximizes predictive performance in a training set, we propose instead to learn a combination of many C+T scores, corresponding to different sets of hyper-parameters. This extension of C+T that we call SCT (Stacked C+T) makes C+T more flexible. Moreover, we implement the possibility for an user of SCT to define their own groups of variants. This opens many possibilities for SCT. For example, we could derive and stack C+T scores for two related but different GWAS summary statistics, we could use external information such as functional annotations, or we could learn to differentiate between two genetically different phenotypes with similar symptoms such as type 1 and type 2 diabetes.

4.2 Article 3 and supplementary materials

The following article is available as a preprint in *bioRxiv* [1].

[1] https://doi.org/10.1101/653204

Making the most of Clumping and Thresholding for polygenic scores

Florian Privé,[1,*] Bjarni J. Vilhjálmsson,[2] Hugues Aschard[3] and Michael G.B. Blum[1,*]

[1]Laboratoire TIMC-IMAG, UMR 5525, Univ. Grenoble Alpes, CNRS, La Tronche, France,
[2]National Center for Register-based Research (NCRR), Aarhus University, Denmark.
[3]Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France,
[*]To whom correspondence should be addressed.

Contacts:

- florian.prive@univ-grenoble-alpes.fr
- bjv@econ.au.dk
- hugues.aschard@pasteur.fr
- michael.blum@univ-grenoble-alpes.fr

Abstract

Polygenic prediction has the potential to contribute to precision medicine. Clumping and Thresholding (C+T) is a widely used method to derive polygenic scores. When using C+T, people usually test several p-value thresholds to maximize predictive ability of derived polygenic scores. Along with this p-value threshold, we propose to tune 3 other hyper-parameters for C+T. We implement an efficient way to derive C+T scores corresponding to many different sets of hyper-parameters. For example, you can now derive thousands of different C+T scores for 300K individuals and 1M variants in less than one day. We show that tuning 4 hyper-parameters of C+T consistently improves its predictive performance in both simulations and real data applications as compared to tuning only the p-value threshold.

Using this grid of computed C+T scores, we further extend C+T with stacking. More precisely, instead of choosing one set of hyper-parameters that maximizes prediction in some training set, we propose to learn an optimal linear combination of all these C+T scores using an efficient penalized regression. We call this method Stacked Clumping and Thresholding (SCT) and show that this makes C+T more flexible. When the training set is large enough, SCT can provide much larger predictive performance as compared to any of the C+T scores individually.

1 Introduction

The ability to predict disease risk accurately is a principal aim of modern precision medicine. As more population-scale genetic datasets become available, polygenic risk scores (PRS) are expected to become more accurate and clinically relevant. The most commonly used method for computing polygenic scores is Clumping and Thresholding (C+T), also known as pruning and thresholding (P+T). The C+T polygenic score is defined as the sum of allele counts (genotypes), weighted by estimated effect sizes obtained from genome-wide association studies, where two filtering steps have been applied (Wray *et al.* 2007; Purcell *et al.* 2009; Dudbridge 2013; Wray *et al.* 2014; Euesden *et al.* 2014; Chatterjee *et al.* 2016). More precisely, the variants are first clumped (C) so that only variants that are weakly correlated with one another are retained. Clumping looks at the most significant variant iteratively, computes correlation between this index variant and nearby variants within some genetic distance w_c, and removes all the nearby variants that are correlated with this index variant beyond a particular value r_c^2. Thresholding (T) consists in removing variants with a p-value larger than a chosen level of significance ($p > p_T$). Both steps, clumping and thresholding, represent a statistical compromise between signal and noise. The clumping step prunes redundant correlated effects caused by linkage disequilibrium (LD) between variants. However, this procedure may also remove independently predictive variants in nearby LD regions. Similarly, thresholding must balance between including predictive variants and reducing noise in the score by excluding null effects. This is why hyper-parameters of clumping and thresholding must be chosen with care when using C+T in order to maximize its predictive ability.

When applying C+T, one has 3 hyper-parameters to select, namely the squared correlation threshold r_c^2 and the window size w_c of clumping, along with the p-value threshold p_T. Usually, C+T users assign default values for clumping, such as r_c^2 of 0.1 (default of PRSice), 0.2 or 0.5 (default of PLINK), and w_c of 250kb (default of PRSice and PLINK) or 500kb, and test several values for p_T ranging from 1 to 10^{-8} (Purcell *et al.* 2009; Wray *et al.* 2014; Euesden *et al.* 2014; Chang *et al.* 2015). Moreover, to match the variants of summary statistics and to compute the PRS, the target sample genotypes are usually imputed to some degree of precision. Liberal inclusion of imputed variants is common, assuming that using more variants in the model yields better prediction, whatever the imputation accuracy of these variants. Here, we explore the validity of this approach and suggest an additional INFO_T threshold on the quality of imputation (often called the INFO score) as a fourth parameter of the C+T method.

We implement an efficient way to compute C+T scores for many different parameters (LD, window size, p-value and INFO score) in R package bigsnpr (Privé *et al.* 2018). Using a training set, one could therefore choose the best predictive C+T model among a large set of C+T models with many different parameters, and then evaluate this model in a test set. Moreover, instead of choosing one set of parameters that corresponds to the best prediction, we propose to use stacking, i.e. we learn an optimal linear combination of all computed C+T scores using an efficient penalized regression to improve prediction beyond the best prediction provided by any of these scores (Breiman 1996). We call this method SCT, which stands for Stacked Clumping and Thresholding. Using the UK Biobank

data (Bycroft *et al.* 2018) and external summary statistics for simulated and real data analyses, we show that testing a larger grid of parameters consistently improves predictions as compared to using some default parameters for C+T. We also show that SCT consistently improves prediction compared to any single C+T model when sample size of the training set is large enough.

2 Material and Methods

2.1 Clumping and Thresholding (C+T) and Stacked C+T (SCT)

We compute C+T scores *for each chromosome separately* and for several parameters:

- Threshold on imputation INFO score INFO_T within $\{0.3, 0.6, 0.9, 0.95\}$.
- Squared correlation threshold of clumping r_c^2 within $\{0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 0.95\}$.
- Base size of clumping window within $\{50, 100, 200, 500\}$. The window size w_c is then computed as the base size divided by r_c^2. For example, for $r_c^2 = 0.2$, we test values of w_c within $\{250, 500, 1000, 2500\}$ (in kb). This is motivated by the fact that linkage disequilibrium is inversely proportional to genetic distance between variants (Pritchard and Przeworski 2001).
- A sequence of 50 thresholds on p-values between the least and the most significant p-values, equally spaced on a log-log scale.

Thus, for individual i, chromosome k and the four hyper-parameters INFO_T, r_c^2, w_c and p_T, we compute C+T predictions

$$V_i^{(k)}\left(\text{INFO}_T,\, r_c^2,\, w_c,\, p_T\right) = \sum_{j \in S_{\text{clumping}}(k,\, \text{INFO}_T,\, r_c^2,\, w_c)} \hat{\beta}_j \cdot G_{i,j} \cdot \mathbb{1}\{p_j < p_T\}\,,$$

where $\hat{\beta}_j$ (p_j) are the effect sizes (p-values) estimated from the GWAS, $G_{i,j}$ is the dosage for individual i and variant j, and the set $S_{\text{clumping}}(k,\, \text{INFO}_T,\, r_c^2,\, w_c)$ corresponds to first restricting to variants of chromosome k with an INFO score $\geq \text{INFO}_T$ and that further result from clumping with parameters r_c^2 and w_c.

Overall, we compute $22 \times 4 \times 7 \times 4 \times 50 = 123200$ vectors of polygenic scores. Then, we stack all these polygenic scores (for individuals in the training set) by using these scores as explanatory variables and the phenotype as the outcome in a regression setting (Breiman 1996). In other words, we fit weights for each C+T scores using an efficient penalized logistic regression available in R package bigstatsr (Privé *et al.* 2019). This results in a linear combination of C+T scores, where C+T scores are merely linear combinations of variants, so that we can derive a single vector of effect sizes corresponding to each variant. The single vector of new variant effects resulting from stacking C+T scores is used for evaluation in the test set. We refer to this method as "SCT" in the rest of the paper.

From this grid of 123,200 vectors of polygenic scores, we also derive two C+T scores for comparison. First, "stdCT" is the standard C+T score using some default parameters, i.e. with r_c^2 = 0.2, w_c = 500, a liberal threshold of 0.3 on imputation INFO score, and choosing the p-value threshold ($\geq 10^{-8}$)

maximizing the AUC on the training set (Wray *et al.* 2014). Second, "maxCT" is the C+T score maximizing the AUC on the training set among the 5600 (123200 / 22) C+T scores corresponding to all different sets of parameters tested. Note that stdCT and maxCT use the same set of parameters for all chromosomes, i.e. for one set of the four hyper-parameters, they are defined as $V^{(1)} + \cdots + V^{(22)}$. In contrast, SCT uses the whole matrix of 123,200 vectors.

2.2 Simulations

We use variants from the UK Biobank (UKBB) imputed dataset that have a minor allele frequency larger than 1% and an imputation INFO score larger than 0.3. There are almost 10M such variants, of which we randomly choose 1M. To limit population structure and family structure, we restrict individuals to the ones identified by the UK Biobank as British with only subtle structure and exclude all second individuals in each pair reported by the UK Biobank as being related (Bycroft *et al.* 2018). This results in a total of 335,609 individuals that we split into three sets: a set of 315,609 individuals for computing summary statistics (GWAS), a set of 10,000 individuals for training hyper-parameters and lastly a test set of 10,000 individuals for evaluating models.

We read the UKBB BGEN files using function `snp_readBGEN` from package bigsnpr (Privé *et al.* 2018). For simulating phenotypes and computing summary statistics, we read UKBB data as hard calls by randomly sampling hard calls according to reported imputation probabilities. For the training and test sets, we read these probabilities as dosages (expected values). This procedure enables us to simulate phenotypes using hard calls and then to use the INFO score (imputation accuracies) reported by the UK Biobank to assess the quality of the imputed data used for the training and test sets.

We simulate binary phenotypes with a heritability $h^2 = 0.5$ using a Liability Threshold Model (LTM) with a prevalence of 10% (Falconer 1965). We vary the number of causal variants (100, 10K, or 1M) in order to match a range of genetic architectures from low to high polygenicity. Liability scores are computed from a model with additive effects only: we compute the liability score of the i-th individual as $y_i = \sum_{j \in S_{\text{causal}}} w_j \widetilde{G_{i,j}} + \epsilon_i$, where S_{causal} is the set of causal variants, w_j are weights generated from a Gaussian distribution $N(0, h^2/|S_{\text{causal}}|)$, $G_{i,j}$ is the allele count of individual i for variant j, $\widetilde{G_{i,j}}$ corresponds to its standardized version (zero mean and unit variance), and ϵ follows a Gaussian distribution $N(0, 1 - h^2)$.

We explore three additional scenarios with more complex architectures. In scenario "2chr", 100 variants of chromosome 1 and all variants of chromosome 2 are causal with half of the heritability for both chromosomes; it aims at assessing predictive performance when disease architectures are different for different chromosomes. In scenario "err", we sample 10,000 random causal variants but 10% of the GWAS effects are reported with an opposite effect in the summary statistics; it aims at assessing if methods are able to partially correct for errors or mere differences in effect sizes between GWAS and the target data. In scenario "HLA", 7105 causal variants are chosen in one long-range LD region of chromosome 6; it aims at assessing if methods can handle strong correlation between causal variants.

To compute summary statistics, we use Cochran-Armitage additive test (Zheng *et al.* 2012). Given that we restricted the data to have minimal population structure, this test based on contingency tables is much faster than using a logistic regression with 10 principal components as covariates (a few minutes vs several hours) while providing similar effect sizes and Z-scores (Figure S1).

In simulations, we compare four methods: stdCT, maxCT, SCT (defined in section 2.1) and lassosum (Mak *et al.* 2017). Each simulation scenario is repeated 10 times and the average AUC is reported. We prefer to use AUC over Nagelkerke's R^2 because AUC has a desirable property of being independent of the proportion of cases in the validation sample; one definition of AUC is the probability that the score of a randomly selected case is larger than the score of a randomly selected control (Wray *et al.* 2013). An alternative to AUC would be to use a better R^2 on the liability scale (Lee *et al.* 2012; Allegrini *et al.* 2019).

2.3 Real summary statistics

We also investigate predictive performance of C+T and SCT in the UK Biobank using external summary statistics from published GWAS of real diseases, for which we summarize the number of individuals and variants in table 1 (Buniello *et al.* 2018). As in simulations, we restrict individuals to the ones identified by the UK Biobank as British with only subtle structure and exclude all second individuals in each pair reported by the UK Biobank as being related (Bycroft *et al.* 2018). Table 1 also summarizes the number of cases and controls in the UKBB, after this filtering and for each phenotype analyzed. For details on how we define phenotypes in the UKBB, please refer to our R code (Section 2.4). Briefly, we use self-reported illness codes (field #20001 for cancers and #20002 otherwise) and ICD10 codes (fields #40001, #40002, #41202 and #41204 for all diseases, and field #40006 specifically for cancers).

Table 1: Number of cases and controls in UK Biobank (UKBB) for several disease phenotypes, along with corresponding published GWAS summary statistics. Summary statistics are chosen from GWAS that did not include individuals from UKBB. For depression, we remove UKBB individuals from the pilot release since they were included in the GWAS from which we use summary statistics.

Trait	UKBB size	GWAS size	GWAS #variants	GWAS citation
Breast cancer (BRCA)	11,578 / 158,391	137,045 / 119,078	11,792,542	Michailidou *et al.* (2017)
Rheumatoid arthritis (RA)	5615 / 226,327	29,880 / 73,758	9,739,303	Okada *et al.* (2014)
Type 1 diabetes (T1D)	771 / 314,547	5913 / 8828	8,996,866	Censin *et al.* (2017)
Type 2 diabetes (T2D)	14,176 / 314,547	26,676 / 132,532	12,056,346	Scott *et al.* (2017)
Prostate cancer (PRCA)	6643 / 141,321	79,148 / 61,106	20,370,946	Schumacher *et al.* (2018)
Depression (MDD)	22,287 / 255,317	59,851 / 113,154	13,554,550	Wray *et al.* (2018)
Coronary artery disease (CAD)	12,263 / 225,927	60,801 / 123,504	9,455,778	Nikpay *et al.* (2015)
Asthma	43,787 / 261,985	19,954 / 107,715	2,001,280	Demenais *et al.* (2018)

We keep all variants with a GWAS p-value lower than 0.1 except for prostate cancer (0.05) and asthma (0.5). This way, we keep around 1M variants for each phenotype, deriving all C+T scores and stacking them in SCT in less than one day for each phenotype, even when using 300K individuals in the training set. To match remaining summary statistics with data from the UK Biobank, we first remove ambiguous alleles [A/T] and [C/G]. We then augment the summary statistics twice: first by

duplicating each variant with the complementary alleles, then by duplicating variants with reverse alleles and effects. Finally, we include only variants that we match with UKBB based on the combination of chromosome, position and the two alleles. Note that, when no or very few alleles are flipped, we disable the strand flipping option and therefore do not remove ambiguous alleles; this is the case for all phenotypes analyzed here. For example, for type 2 diabetes, there are 1,408,672 variants in summary statistics ($p < 0.1$), of which 215,821 are ambiguous SNPs. If we remove these ambiguous SNPs, 1,145,260 variants are matched with UKBB, of which only 38 are actually flipped. So, instead, we do not allow for flipping and do not remove ambiguous alleles, then 1,350,844 variants are matched with UKBB.

Training SCT and choosing optimal hyper-parameters for C+T (stdCT and maxCT) use 63%-90% of the UK Biobank data reported in table 1. The training set can therefore contain as many as 300K individuals. To assess how sample size affects predictive performance of methods, we also compare these methods using a much smaller training set of 500 cases and 2000 controls.

2.4 Reproducibility

The code to reproduce the analyses and figures of this paper is available as R scripts at `https://github.com/privefl/simus-PRS/tree/master/paper3-SCT` (R Core Team 2018). To execute these scripts, you need to have access to the UK Biobank data that we are not allowed to share (`http://www.ukbiobank.ac.uk/`). A quick introduction to SCT is also available at `https://privefl.github.io/bigsnpr/articles/SCT.html`.

3 Results

3.1 Simulations

We test 6 different simulations scenarios. In all these scenarios, maxCT –that tests a much larger grid of hyper-parameters values for C+T on the training set– consistently provides higher AUC values on the test set as compared to stdCT that tests only several p-value thresholds while using default values for the other parameters (Figure 1). The absolute improvement in AUC of maxCT over stdCT is particularly large in the cases of 100 and 10,000 causal variants, where causal effects are mostly independent of one another. In these cases, using a very stringent $r_c^2 = 0.01$ threshold of clumping provides higher predictive performance than using a standard default of $r_c^2 = 0.2$ (Figures S4a and S4b). However, $r_c^2 = 0.2$ provides best predictive performance when simulating 1M causal variants. Still, using a large window size w_c of 2500 kb increases AUC as compared to using default values of either 250 or 500 kb (Figure S4c).

As for SCT, it provides equal or higher predictive performance than maxCT in the different simulation scenarios (Figure 1). In the first three simple scenarios simulating 100, 10K or 1M causal variants anywhere on the genome, predictive performance of SCT are similar to maxCT. In the "2chr" scenario where there are large effects on chromosome 1, small effects on chromosome 2 and no effect

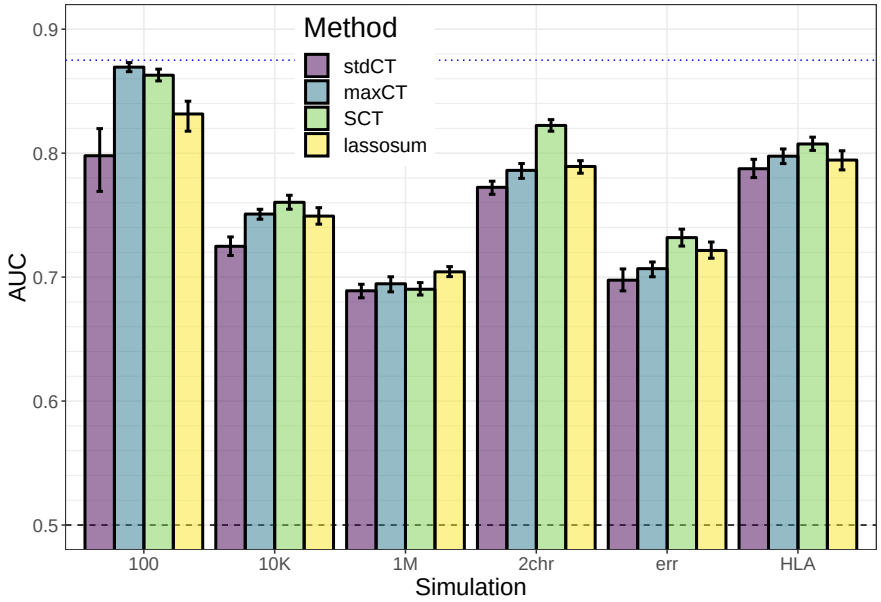


Figure 1: Results of the 6 simulation scenarios: (100) 100 random causal variants; (10K) 10,000 random causal variants; (1M) all 1M variants are causal variants; (2chr) 100 variants of chromosome 1 are causal and all variants of chromosome 2, with half of the heritability for both chromosomes; (err) 10,000 random causal variants, but 10% of the GWAS effects are reported with an opposite effect; (HLA) 7105 causal variants in a long-range LD region of chromosome 6. Mean and 95% CI of 10^4 non-parametric bootstrap replicates of the mean AUC of 10 simulations for each scenario. The blue dotted line represents the maximum achievable AUC for these simulations (87.5% for a prevalence of 10% and an heritability of 50% – see equation (3) of Wray *et al.* (2010)). See corresponding values in table S1.

on other chromosomes, mean AUC is 78.7% for maxCT and 82.2% for SCT. In the "err" scenario where we report GWAS summary statistics with 10% reversed effects (errors), mean AUC is 70.2% for maxCT and 73.2% for SCT. SCT also provides higher AUC than lassosum, expect when simulating all variants as causal (1M).

Effects resulting from SCT (Figure S3) are mostly comprised between the GWAS effects and 0. For the simulation with only 100 causal variants, resulting effects are either nearly the same as in the GWAS, or near 0 (or exactly 0). When there are some correlation between causal predictors (Scenarios "1M" and "HLA") or when reporting GWAS effects with some opposite effect ("err"), some effects resulting from SCT are in the opposite direction as compared to the GWAS effects.

3.2 Real summary statistics

In terms of AUC, maxCT outperfoms stdCT for all 8 diseases consireded with a mean absolute increase of 1.3% (Figures 2 and S2). A particularly large increase can be noted when predicting depression status (MDD), from an AUC of 55.7% (95% CI: [54.4-56.9]) with stdCT to an AUC of

59.2% (95% CI: [58.0-60.4]) with maxCT. For MDD, a liberal inclusion in clumping (r_c^2 = 0.8) and a stringent threshold on imputation accuracy (INFO$_T$ = 0.95) provides the best predictive performance (Figure S6f). For all 8 diseases, predictions were optimized when choosing a threshold on imputation accuracy of at least 0.9, whereas optimal values for r_c^2 where very different depending on the architecture of diseases, with optimal selected values over the whole range of tested values for r_c^2 (Table S3).

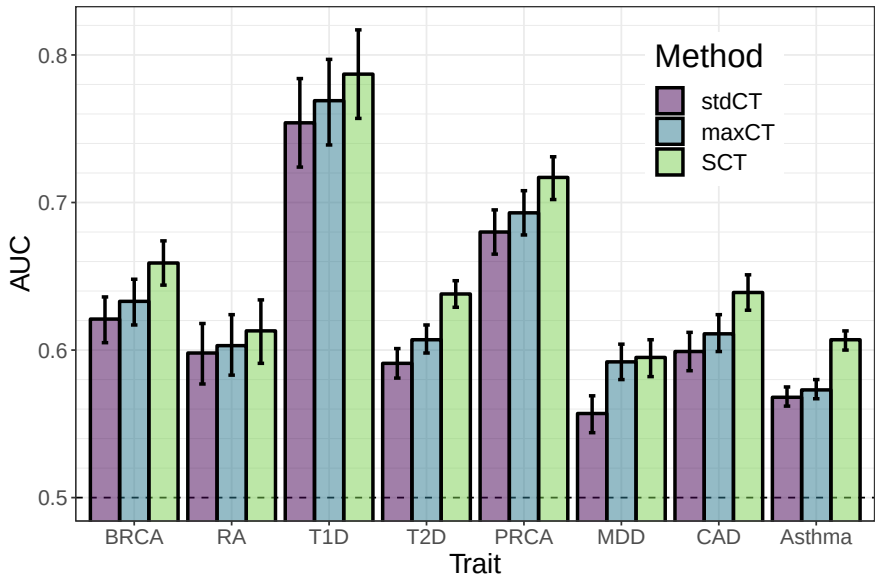


Figure 2: AUC values on the test set of UKBB (mean and 95% CI from 10^4 bootstrap samples). Training SCT and choosing optimal hyper-parameters for C+T use 63%-90% of the data reported in table 1. See corresponding values in table S2.

Furthermore, when training size uses a large proportion of the UK Biobank data, SCT outperforms maxCT for all 8 diseases considered with an additional mean absolute increase of AUC of 2.2%, making it 3.5% as compared to stdCT (Figure 2 and table S2). Predictive performance improvement of SCT versus maxCT is particularly notable for coronary artery disease (2.8%), type 2 diabetes (3.1%) and asthma (3.4%).

Effects resulting from SCT have mostly the same sign as initial effects from GWAS, with few effects being largely unchanged, and others having an effect that is shrunk to 0 or equals to 0, i.e. variants not included in the final model (Figure S5).

When training size is smaller (500 cases and 2000 controls only instead of 200K-300K individuals), SCT is not as good as when training size is large, yet SCT remains a competitive method expect for depression for which maxCT performs much better than SCT (Figure S2).

4 Discussion

4.1 Predictive performance improvement of C+T

C+T has the advantage that it is intuitive and an easily applicable method for obtaining polygenic scores trained on GWAS summary statistics. Two popular pieces of software that implement C+T, PLINK and PRSice, have further contributed to the prevalence of C+T (Purcell *et al.* 2007; Euesden *et al.* 2014; Chang *et al.* 2015). Usually, C+T scores for different p-value thresholds are derived, using some default values for the other 3 hyper-parameters. In R package bigsnpr, we extend C+T to efficiently consider more hyper-parameters (4 by default) and enable the user to define their own qualitative variant annotations to filter on (e.g. minor allele frequency could be used as a fifth parameter). Using simulated and real data, we show that choosing different values rather than default ones for these hyper-parameters can substantially improve the performance of C+T, making C+T a very competitive method. Indeed, in our simulations (Figure 1), we found that optimizing C+T (maxCT) performed on par with more sophisticated methods such as lassosum. Moreover, it is possible to rerun the method using a finer grid in a particular range of these hyper-parameters. For example, it might be useful to include variants with p-values larger than 0.1 for predicting rheumatoid arthritis and depression (Figures S6b and S6f). Another example would be to focus on a finer grid of large values of r_c^2 for coronary artery disease (Figure S6g), or to focus on a finer grid of stringent imputation thresholds only (Table S3).

Using a large grid of C+T scores for different hyper-parameters, we show that stacking these scores instead of choosing the best one improves prediction further (Breiman 1996). Combining multiple PRS is not a new idea (Krapohl *et al.* 2018; Inouye *et al.* 2018), but we push this idea to the limit by combining 123,200 polygenic scores. This makes SCT more flexible than any C+T model, but it of course also requires a larger training dataset with individual-level genotypes and phenotypes to learn the weights in stacking.

Normally, cross-validation should be used to prevent overfitting when using stacking and it is also suggested to use positivity constraints in stacking (Breiman 1996). However, cross-validation is not necessary here since building C+T scores does not make use of the phenotype of the training set that is later used in the stacking; the training set is only used to choose the best set of hyper-parameters for C+T. Moreover, we allow C+T scores to have negative weights in the final model for three reasons. First, because C+T scores are overlapping in the variants they use, using some negative weights allows to weight groups of variants that correspond to the difference of two sets of variants. Second, because of LD, variants may have different effects when learned jointly with others (Figures S3c and S3f). Third, if reported GWAS effects are heterogenous between the GWAS dataset and the validation or target dataset, then having variants with opposite effects can help adjust the effects learned during GWAS.

4.2 Limitations of the study

In this study, we limited the analysis to 8 common diseases and disorders, as these all had substantial number of cases and publicly available GWAS summary statistics based on substantial sample sizes. For example, for psychiatric disease, we include only depression (MDD) because diseases such as schizophrenia and bipolar disorder have very few cases in the UK Biobank; dedicated datasets should be used to assess effectiveness of maxCT and SCT for such diseases. We also do not analyze many automimmune diseases because summary statistics are often outdated (2010-2011[1]) and, because there are usually large effects in regions of chromosome 6 with high LD, methods that use individual-level data instead of summary statistics are likely to provide better predictive models (Privé *et al.* 2019). We also chose not to analyze any continuous trait such as height or BMI because there are many individual-level data available in UKBB for such phenotypes and methods directly using individual-level data are likely to provide better predictive models for predicting in UKBB than the ones using summary statistics (Privé *et al.* 2019; Chung *et al.* 2019). Phenotypes with tiny effects such as educational attainment for which huge GWAS summary statistics are available might be an exception (Lee *et al.* 2018).

The principal aim of this work is to study and improve the widely used C+T method. The idea behind C+T is simple as it directly uses weights learned from GWAS; it further removes variants as one often does when reporting hits from GWAS, i.e. only variants that pass the genome-wide threshold (p-value thresholding) and that are independent association findings (clumping) are reported. Yet, there are two other established methods based on summary statistics, LDpred and lassosum (Vilhjálmsson *et al.* 2015; Mak *et al.* 2017; Allegrini *et al.* 2019). Several other promising and more complex methods such as NPS, PRS-CS and SBayesR are currently being developed (Chun *et al.* 2019; Ge *et al.* 2019; Lloyd-Jones *et al.* 2019). Here, we include lassosum in the simulations since no other method yet shown that they provide some improvement over lassosum. In addition, we found lassosum to be easy to set up and use. However, lassosum requires substantial computation time when there are too many samples or too many variants. Therefore, we did not apply lassosum to the full UK Biobank data. A full comparison of methods (including individual-level data methods), including binary and continuous traits with different architectures, using different sizes of summary statistics and individual-level data for training, and in possibly different populations would be of great interest, but is out of scope for this paper. Indeed, we believe that different methods may perform very differently in different settings and that understanding what method is appropriate for each case is of paramount interest if the aim is to maximize prediction accuracy to make PRS clinically useful.

4.3 Extending SCT

The stacking step of SCT can be used for either binary or continuous phenotypes. Yet, for some diseases, it makes sense to include age in the models, using for example Cox proportional-hazards model to predict age of disease onset, with possibly censored data (Cox 1972). Cox regression has already proven useful for increasing power in GWAS (Hughey *et al.* 2019). Currently, we support linear and

[1] https://www.immunobase.org/downloads/protected_data/GWAS_Data/

logistic regressions in our efficient implementation of package bigstatsr, but not Cox regression. This is an area of future development; for now, if sample size is not too large, one could use R package glmnet to implement stacking based on Cox model (Tibshirani *et al.* 2012).

One might also want to use other information such as sex or ancestry (using principal components). Indeed, it is easy to add covariates in the stacking step as (possibly unpenalized) variables in the penalized regression. Yet, adding covariates should be done with caution (see the end of supplementary materials).

Finally, note that we added an extra parameter in the SCT pipeline that makes possible for an user to define their own groups of variants. This allows to refine the grid of computed C+T scores and opens many possibilities for SCT. For example, we could derive and stack C+T scores for two (or more) different GWAS summary statistics, e.g. for different ancestries or for different phenotypes. This would effectively extend SCT as a multivariate method. We could also learn to differentiate between two genetically different phenotypes with similar symptoms such as type 1 and type 2 diabetes, which is in our research interests.

4.4 Conclusion

In this paper, we focused on understanding and improving the widely-used C+T method by testing a wide range of hyper-parameters values. More broadly, we believe that any implementation of statistical methods should come with an easy and effective way to choose hyper-parameters of these methods well. We believe that C+T will continue to be used for many years as it is both simple to use and intuitive. Moreover, as we show, when C+T is optimized using a larger grid of hyper-parameters, it remains a competitive method since it can adapt to many different disease architectures by tuning all hyper-parameters.

Moreover, instead of choosing one set of hyper-parameters, we show that stacking C+T predictions improves predictive performance further. SCT has many advantages over any single C+T prediction: first, it can learn different architecture models for different chromosomes, it can learn a mixture of large and small effects and it can more generally adapt initial weights of the GWAS in order to maximize prediction. Moreover, SCT remains a linear model with one vector of coefficients as it is a linear combination (stacking) of linear combinations (C+T scores).

Acknowledgements

Authors acknowledge LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and ANR project FROGH (ANR-16-CE12-0033). Authors also acknowledge the Grenoble Alpes Data Institute that is supported by the French National Research Agency under the “Investissements d’avenir” program (ANR-15-IDEX-02). This research has been conducted using the UK Biobank Resource under Application Number 25589.

References

Allegrini, A. G., Selzam, S., Rimfeld, K., von Stumm, S., Pingault, J.-B., and Plomin, R. (2019). Genomic prediction of cognitive traits in childhood and adolescence. *Molecular Psychiatry*, page 1.

Breiman, L. (1996). Stacked regressions. *Machine learning*, **24**(1), 49–64.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., *et al.* (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, **47**(D1), D1005–D1012.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203.

Censin, J., Nowak, C., Cooper, N., Bergsten, P., Todd, J. A., and Fall, T. (2017). Childhood adiposity and risk of type 1 diabetes: A mendelian randomization study. *PLoS medicine*, **14**(8), e1002362.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**(1), 7.

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, **17**(7), 392.

Chun, S., Imakaev, M., Hui, D., Patsopoulos, N. A., Neale, B. M., Kathiresan, S., Stitziel, N. O., and Sunyaev, S. R. (2019). Non-parametric polygenic risk prediction using partitioned gwas summary statistics. *BioRxiv*, page 370064.

Chung, W., Chen, J., Turman, C., Lindstrom, S., Zhu, Z., Loh, P.-R., Kraft, P., and Liang, L. (2019). Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. *Nature communications*, **10**(1), 569.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202.

Demenais, F., Margaritte-Jeannin, P., Barnes, K. C., Cookson, W. O., Altmüller, J., Ang, W., Barr, R. G., Beaty, T. H., Becker, A. B., Beilby, J., *et al.* (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature genetics*, **50**(1), 42.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, **9**(3), e1003348.

Euesden, J., Lewis, C. M., and Oŕeilly, P. F. (2014). PRSice: polygenic risk score software. *Bioinformatics*, **31**(9), 1466–1468.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, **29**(1), 51–76.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via bayesian regression and continuous shrinkage priors. *bioRxiv*, page 416859.

Hughey, J. J., Rhoades, S. D., Fu, D. Y., Bastarache, L., Denny, J. C., and Chen, Q. (2019). Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BioRxiv*, page 599910.

Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., Lai, F. Y., Kaptoge, S., Brozynska, M., Wang, T., *et al.* (2018). Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *Journal of the American College of Cardiology*, **72**(16), 1883–1893.

Krapohl, E., Patel, H., Newhouse, S., Curtis, C. J., von Stumm, S., Dale, P. S., Zabaneh, D., Breen, G., O'Reilly, P. F., and Plomin, R. (2018). Multi-polygenic score approach to trait prediction. *Molecular psychiatry*, **23**(5), 1368.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Linnér, R. K., *et al.* (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, **50**(8), 1112.

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012). A better coefficient of determination for genetic profile analysis. *Genetic epidemiology*, **36**(3), 214–224.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., *et al.* (2019). Improved polygenic prediction by bayesian multiple regression on summary statistics. *bioRxiv*, page 522961.

Maier, R. M., Zhu, Z., Lee, S. H., Trzaskowski, M., Ruderfer, D. M., Stahl, E. A., Ripke, S., Wray, N. R., Yang, J., Visscher, P. M., *et al.* (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature communications*, **9**(1), 989.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, **41**(6), 469–480.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., *et al.* (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**(7678), 92.

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., *et al.* (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, **47**(10), 1121.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., *et al.* (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**(7488), 376.

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, **69**(1), 1–14.

Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, **34**(16), 2781–2787.

Privé, F., Aschard, H., and Blum, M. G. (2019). Efficient implementation of penalized regression for genetic risk prediction. *Genetics*, pages genetics–302019.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., *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.

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.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., *et al.* (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature genetics*, **50**(7), 928.

Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., Pervjakova, N., Pers, T. H., Johnson, A. D., Eicher, J. D., *et al.* (2017). An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes*, **66**(11), 2888–2902.

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.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., *et al.* (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, **97**(4), 576–592.

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.

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

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., and Visscher, P. M. (2013). Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, **14**(7), 507.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., and Middeldorp, C. M. (2014). Research review: polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, **55**(10), 1068–1087.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M., *et al.* (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics*, **50**(5), 668.

Zheng, G., Yang, Y., Zhu, X., and Elston, R. C. (2012). *Analysis of genetic association studies*. Springer Science & Business Media.

Supplementary Materials

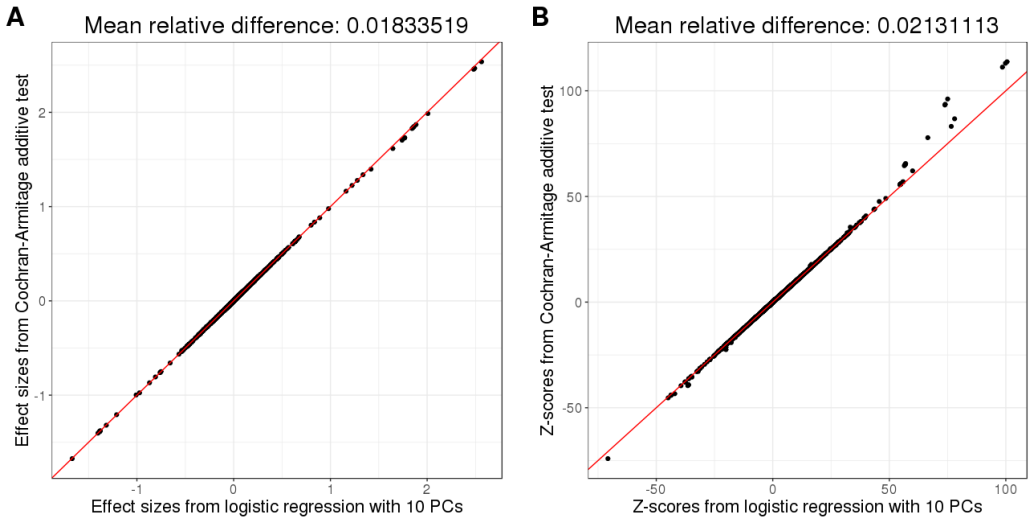


Figure S1: Comparison of estimated effect sizes (**A**) and Z-scores (**B**) if computed using a logistic regression with 10 principal components as covariates, or with a simple Cochran-Armitage additive test. Phenotypes were simulated using 100 causal variants only, allowing for large effects.

Table S1: AUC values on the test set for simulations (mean [95% CI] from 10^4 bootstrap samples).

Scenario	stdCT	maxCT	SCT	lassosum
100	79.8 [77.0-82.0]	86.9 [86.6-87.3]	86.3 [85.8-86.8]	83.2 [81.8-84.2]
10K	72.5 [71.8-73.3]	75.1 [74.7-75.5]	76.0 [75.5-76.6]	74.9 [74.3-75.6]
1M	68.9 [68.3-69.4]	69.5 [68.8-70.0]	69.0 [68.5-69.6]	70.4 [70.0-70.9]
2chr	77.2 [76.7-77.7]	78.6 [78.0-79.2]	82.2 [81.8-82.7]	78.9 [78.4-79.4]
err	69.8 [68.9-70.7]	70.7 [70.1-71.2]	73.2 [72.5-73.9]	72.1 [71.5-72.8]
HLA	78.7 [78.0-79.5]	79.8 [79.1-80.4]	80.7 [80.2-81.3]	79.4 [78.7-80.2]

Table S2: AUC values on the test set of UKBB (mean [95% CI] from 10^4 bootstrap samples) and the number of variants used in the final model. Training SCT and choosing optimal hyper-parameters for C+T use 63%-90% of the data reported in table 1.

Trait	stdCT	maxCT	SCT
Breast cancer (BRCA)	62.1 [60.5-63.6]	63.3 [61.7-64.8]	65.9 [64.4-67.4]
	6256	2572	670,050
Rheumatoid arthritis (RA)	59.8 [57.7-61.8]	60.3 [58.3-62.4]	61.3 [59.1-63.4]
	12,220	88,556	317,456
Type 1 diabetes (T1D)	75.4 [72.4-78.4]	76.9 [73.9-79.7]	78.7 [75.7-81.7]
	1112	267	135,991
Type 2 diabetes (T2D)	59.1 [58.1-60.1]	60.7 [59.8-61.7]	63.8 [62.9-64.7]
	177	33,235	548,343
Prostate cancer (PRCA)	68.0 [66.5-69.5]	69.3 [67.8-70.8]	71.7 [70.2-73.1]
	1035	356	696,575
Depression (MDD)	55.7 [54.4-56.9]	59.2 [58.0-60.4]	59.5 [58.2-60.7]
	165,584	222,912	524,099
Coronary artery disease (CAD)	59.9 [58.6-61.2]	61.1 [59.9-62.4]	63.9 [62.7-65.1]
	1182	87,577	315,165
Asthma	56.8 [56.2-57.5]	57.3 [56.7-58.0]	60.7 [60.0-61.3]
	3034	360	446,120

Table S3: Choice of C+T parameters based on the maximum AUC in the training set. Choosing optimal hyper-parameters for C+T use 63%-90% of the data reported in table 1.

Trait	w_c	r_c^2	INFO$_T$	p_T
Breast cancer (BRCA)	2500	0.2	0.95	2.2e-04
Rheumatoid arthritis (RA)	200	0.5	0.95	7.5e-02
Type 1 diabetes (T1D)	10K-50K	0.01	0.90	2.6e-05
Type 2 diabetes (T2D)	625	0.8	0.95	1.1e-02
Prostate cancer (PRCA)	10K-50K	0.01	0.90	4.2e-06
Depression (MDD)	625	0.8	0.95	1.0e-01
Coronary artery disease (CAD)	526	0.95	0.95	3.5e-02
Asthma	2500	0.2	0.90	2.2e-04

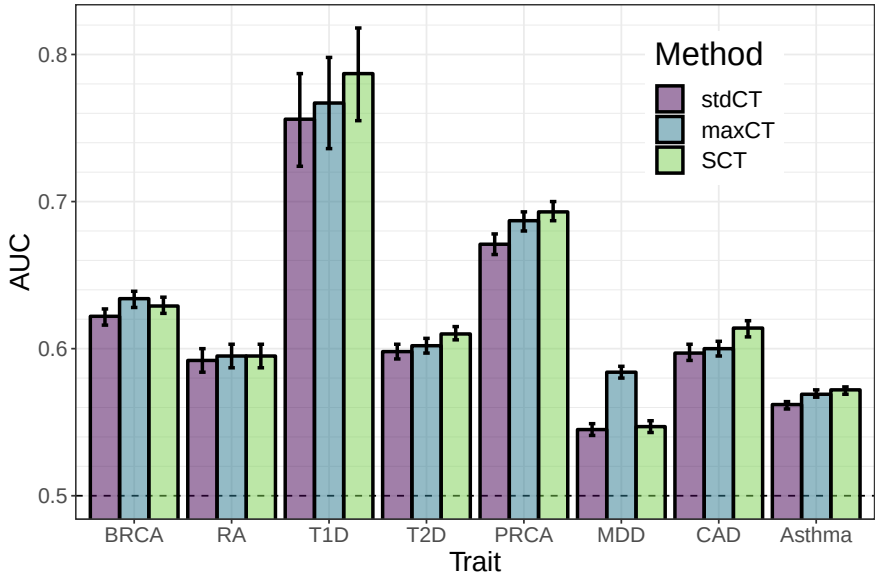
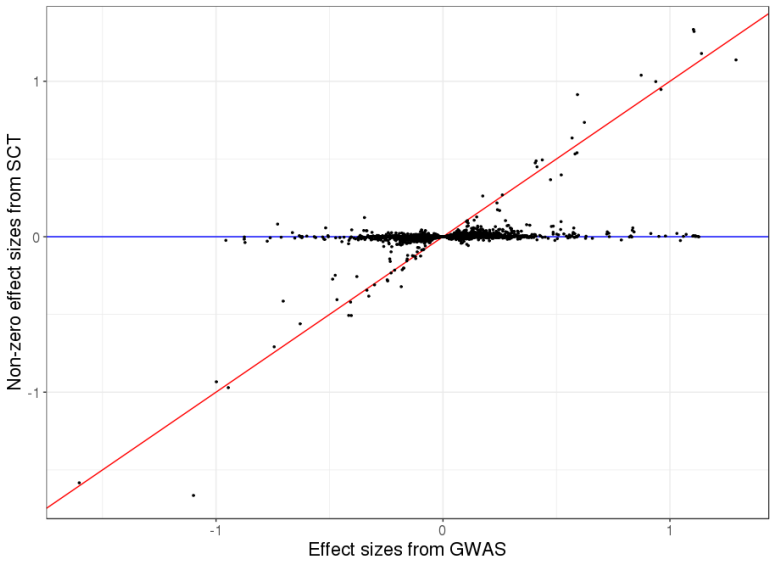


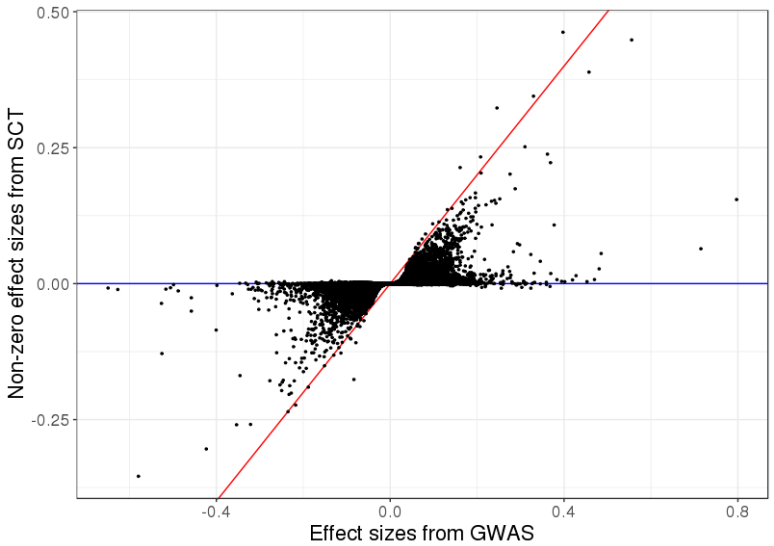
Figure S2: AUC values on the test set of UKBB (mean and 95% CI from 10^4 bootstrap samples). Training SCT and choosing optimal hyper-parameters for C+T use 500 cases and 2000 controls only. See corresponding values in table S4.

Table S4: AUC values on the test set of UKBB (mean [95% CI] from 10^4 bootstrap samples) and the number of variants used in the final model. Training SCT and choosing optimal hyper-parameters for C+T use 500 cases and 2000 controls only.

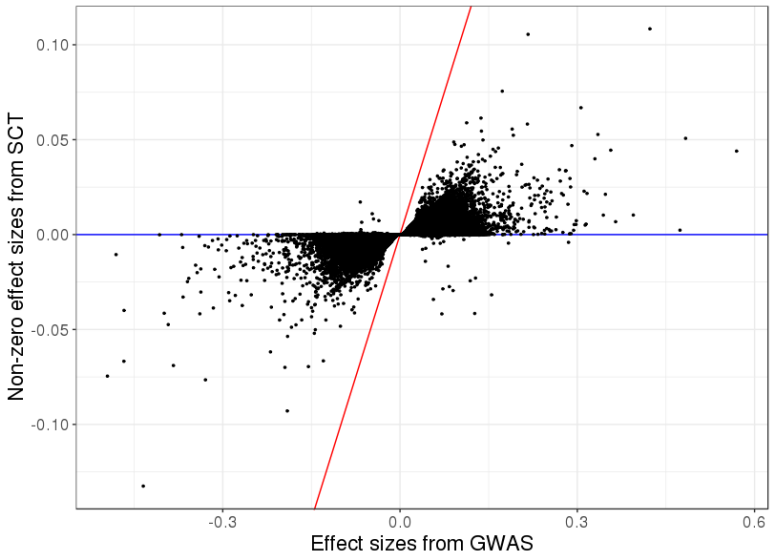
Trait	stdCT	maxCT	SCT
Breast cancer (BRCA)	62.2 [61.6-62.7]	63.4 [62.8-63.9]	62.9 [62.4-63.5]
Rheumatoid arthritis (RA)	59.2 [58.4-60.0]	59.5 [58.7-60.3]	59.5 [58.7-60.3]
Type 1 diabetes (T1D)	75.6 [72.4-78.7]	76.7 [73.6-79.8]	78.7 [75.5-81.8]
Type 2 diabetes (T2D)	59.8 [59.3-60.3]	60.2 [59.7-60.7]	61.0 [60.6-61.5]
Prostate cancer (PRCA)	67.1 [66.4-67.8]	68.7 [68.0-69.3]	69.3 [68.7-70.0]
Depression (MDD)	54.5 [54.1-54.9]	58.4 [58.0-58.8]	54.7 [54.3-55.1]
Coronary artery disease (CAD)	59.7 [59.2-60.3]	60.0 [59.5-60.5]	61.4 [60.8-61.9]
Asthma	56.2 [55.9-56.4]	56.9 [56.7-57.2]	57.2 [56.9-57.4]



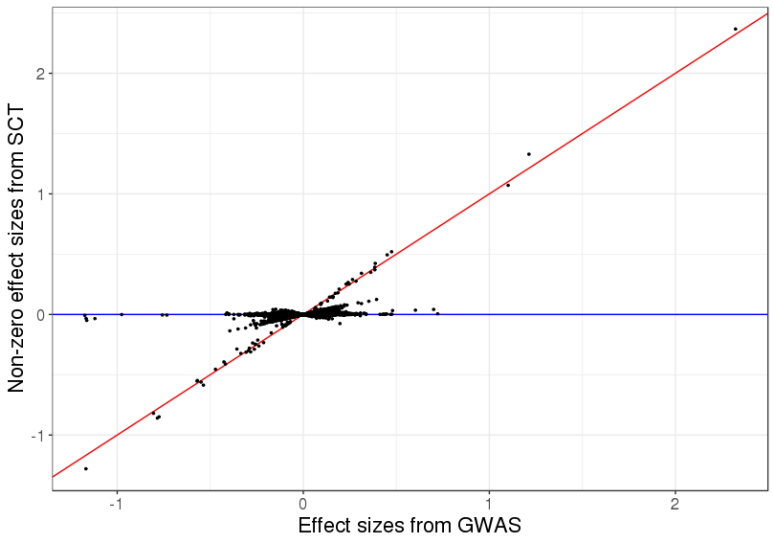
(a) "100": 100 random causal variants



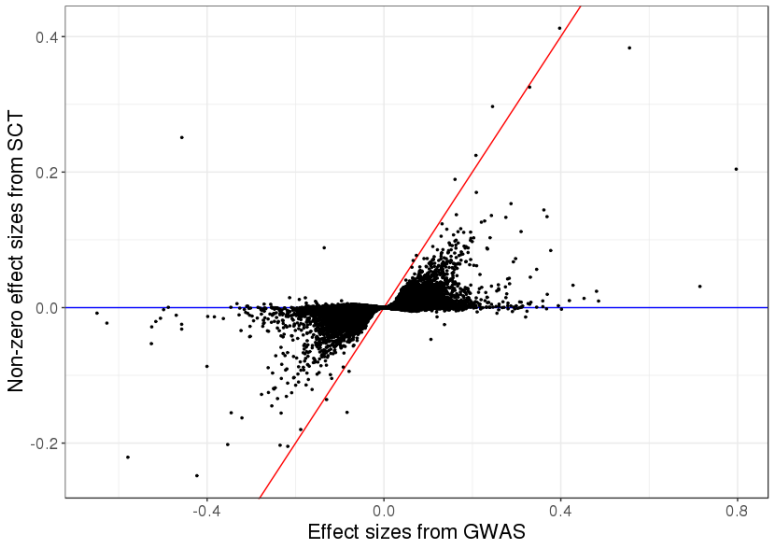
(b) "10K": 10,000 random causal variants



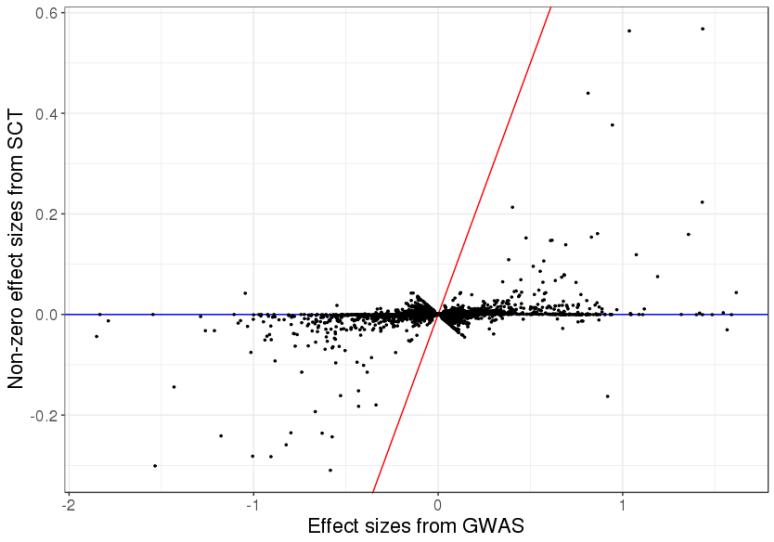
(c) "1M": all 1M variants are causal variants



(d) "2chr": Causal variants on chromosomes 1 & 2

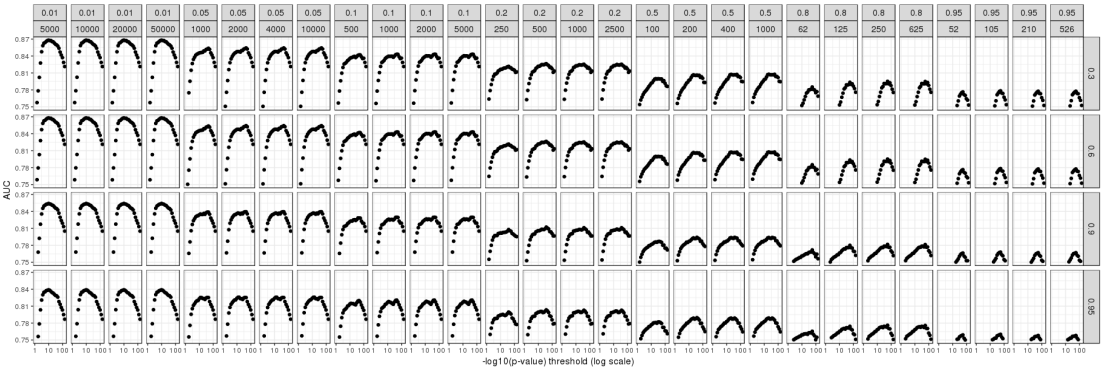


(e) "err": 10,000 random causal variants, but 10% of the GWAS effects are reported with an opposite effect

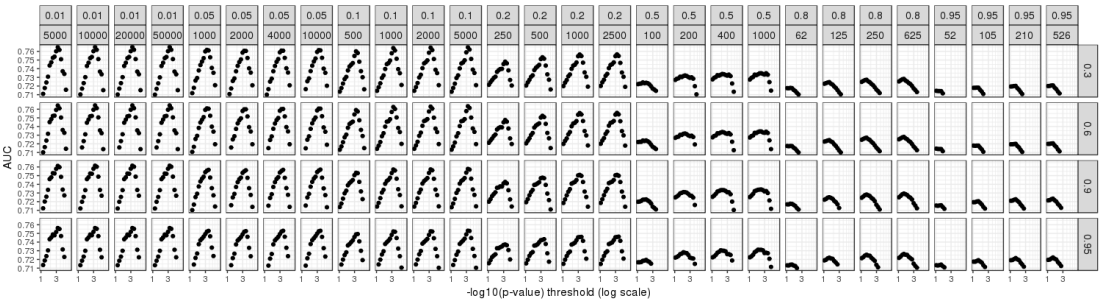


(f) "HLA": 7105 causal variants in a long-range LD region of chromosome 6

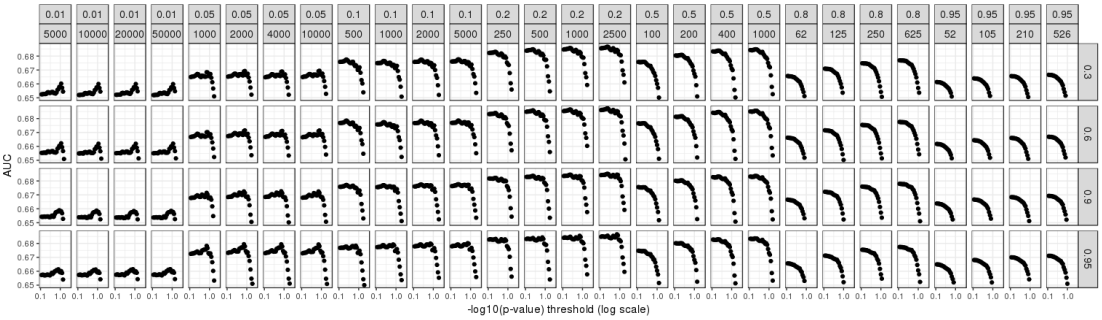
Figure S3: New effect sizes resulting from SCT versus initial effect sizes of GWAS in the first simulation of each simulation scenario. Only non-zero effects are represented. Red line corresponds to the 1:1 line.



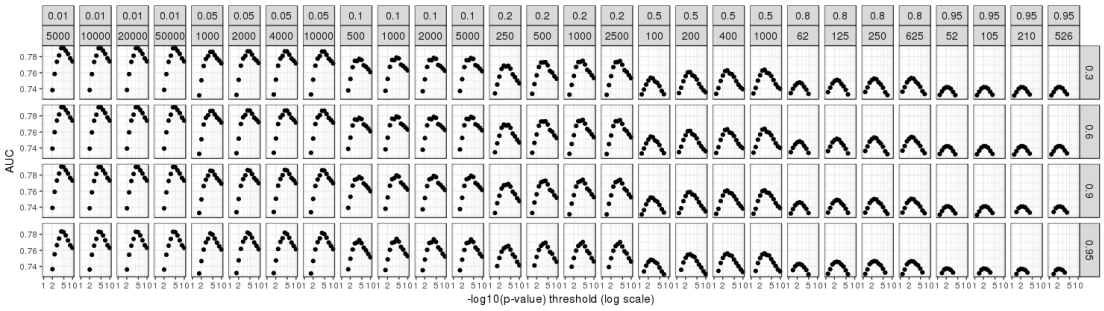
(a) "100": 100 random causal variants



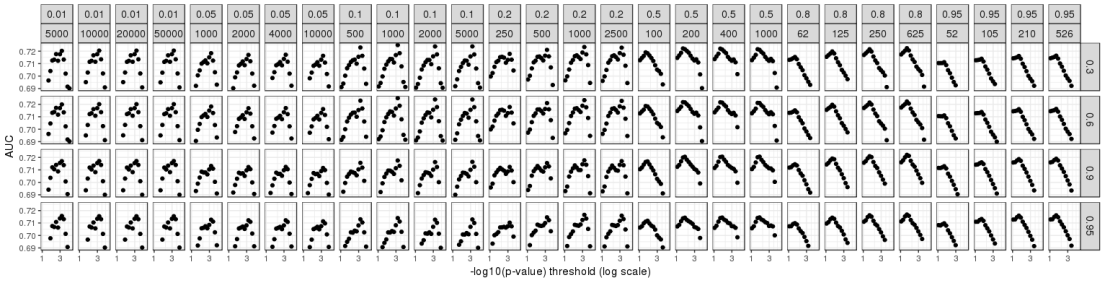
(b) "10K": 10,000 random causal variants



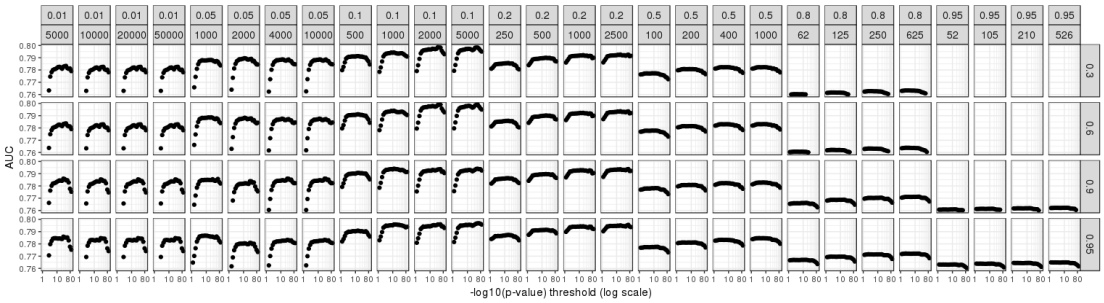
(c) "1M": all 1M variants are causal variants



(d) "2chr": Causal variants on chromosomes 1 & 2

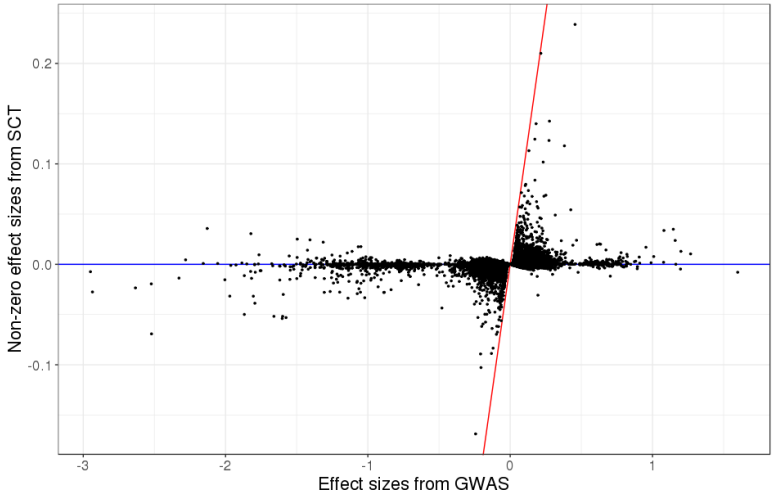


(e) "err": 10,000 random causal variants, but 10% of the GWAS effects are reported with an opposite effect

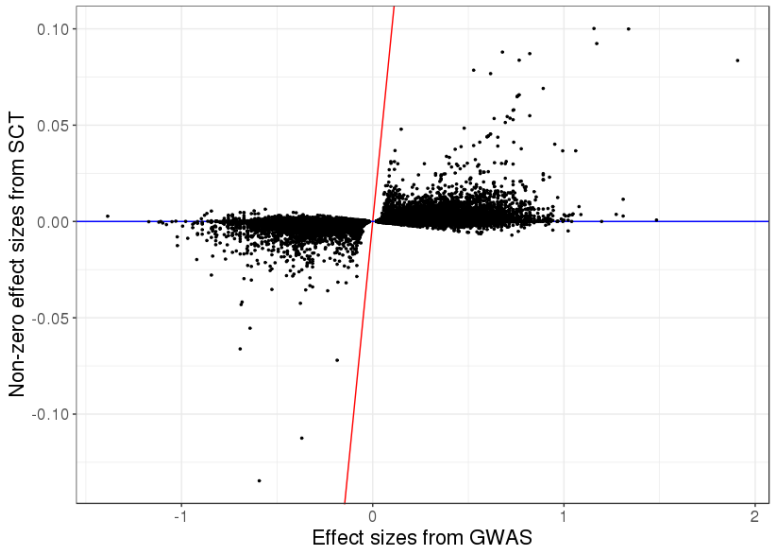


(f) "HLA": 7105 causal variants in a long-range LD region of chromosome 6

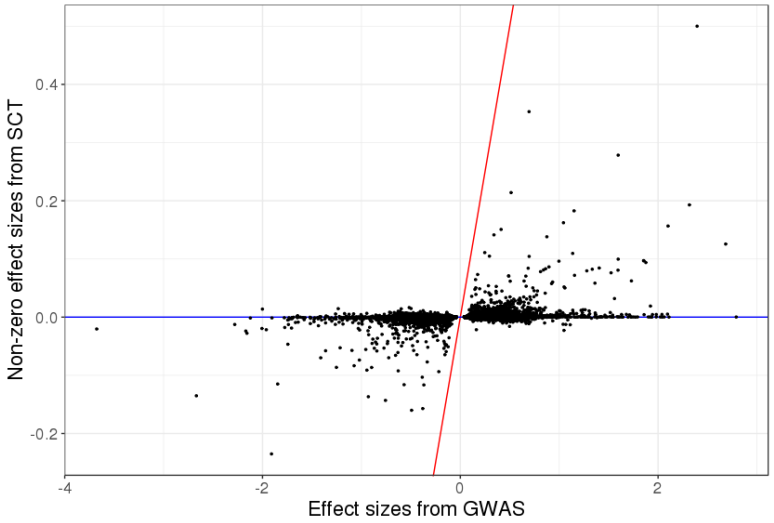
Figure S4: AUC values (for the training set) when predicting disease status for many parameters of C+T in the first simulation of each simulation scenario. Facets are presenting different clumping thresholds r_c^2 from 0.01 to 0.95, window sizes w_c from 52 to 50,000 kb, and imputation thresholds from 0.3 to 0.95. The x-axis corresponds to the remaining hyper-parameter, the p-value threshold p_T; here, -log10(p-values) are represented using a logarithmic scale.



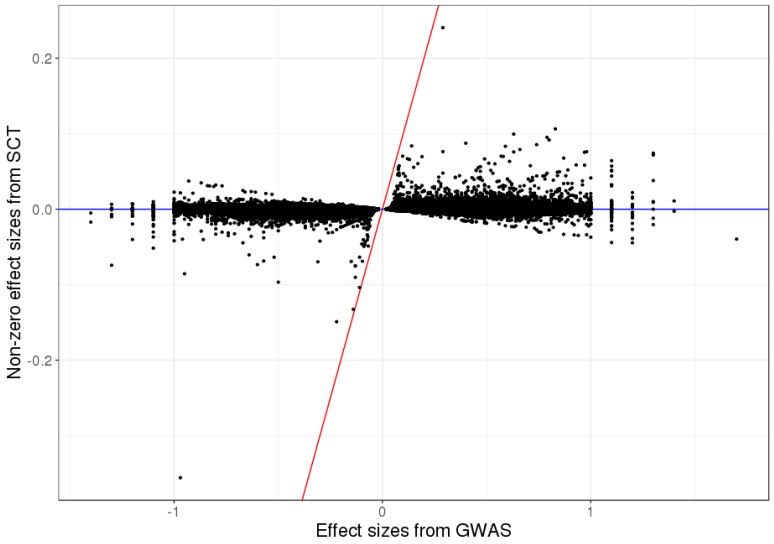
(a) Breast cancer



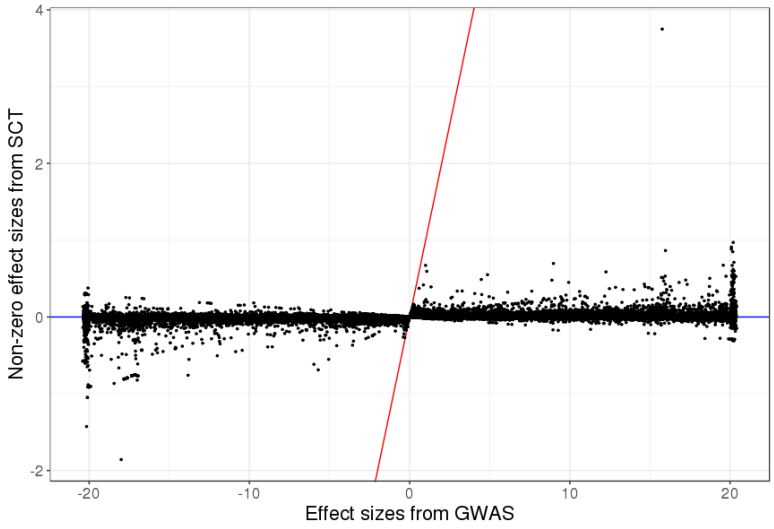
(b) Rheumatoid arthritis



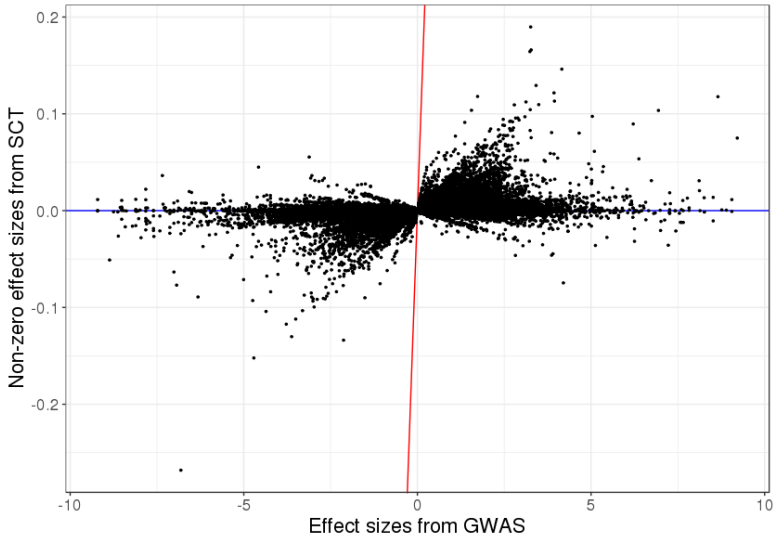
(c) Type 1 diabetes



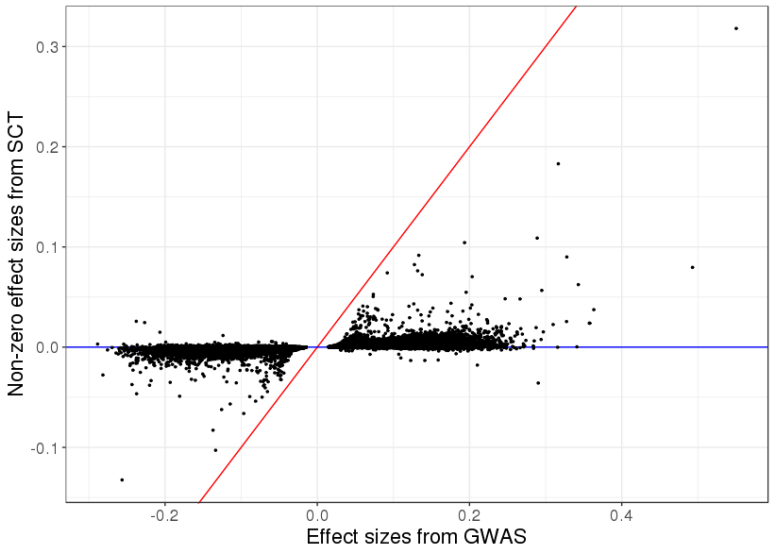
(d) Type 2 diabetes



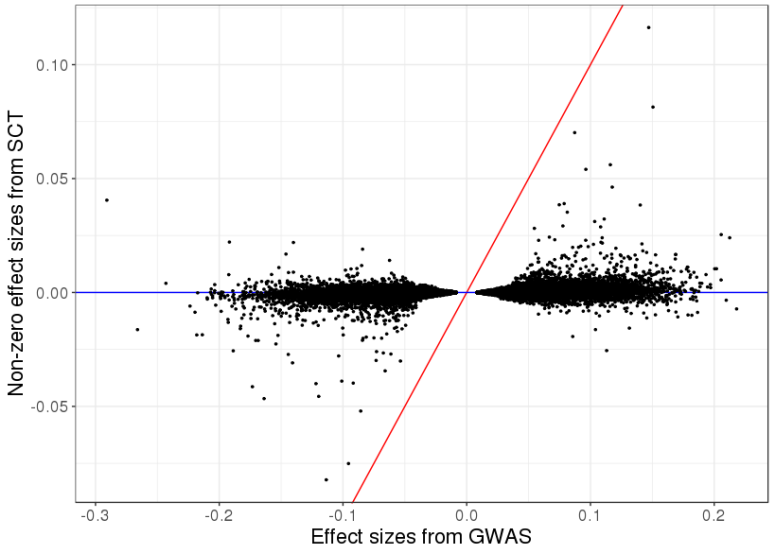
(e) Prostate cancer



(f) Depression

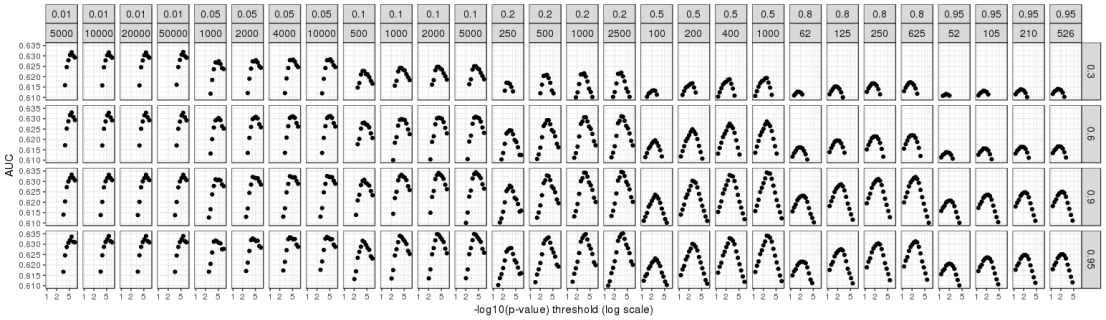


(g) Coronary artery disease

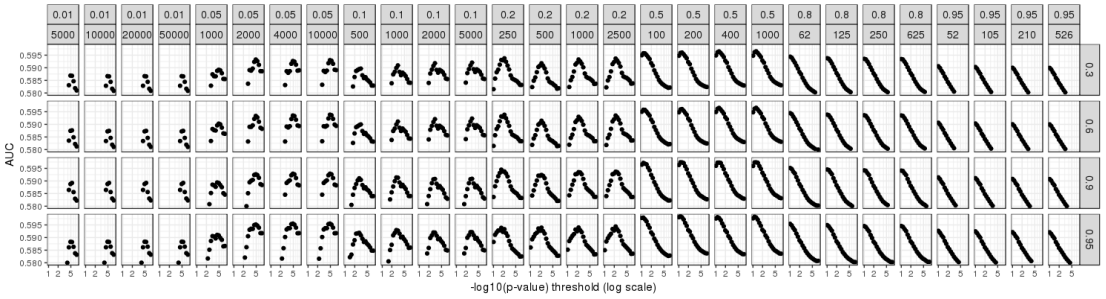


(h) Asthma

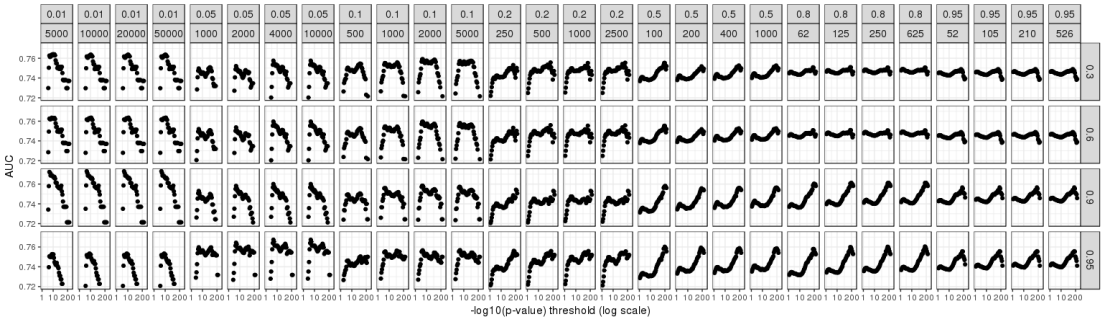
Figure S5: New effect sizes resulting from SCT versus initial effect sizes of GWAS in real data applications. Only non-zero effects are represented. Red line corresponds to the 1:1 line.



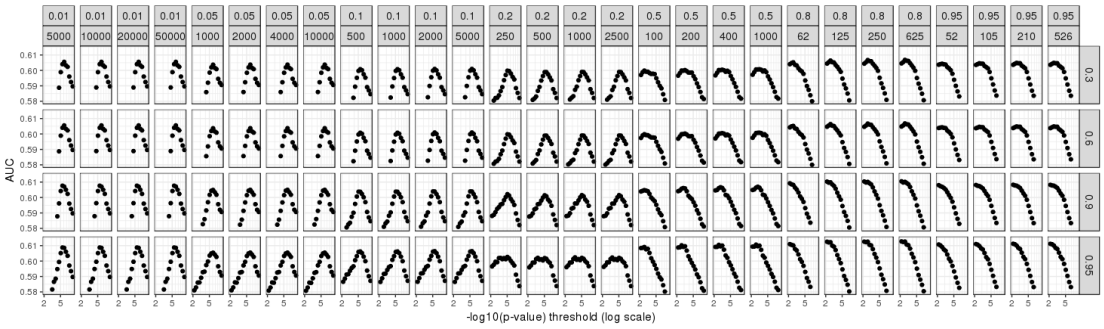
(a) Breast cancer



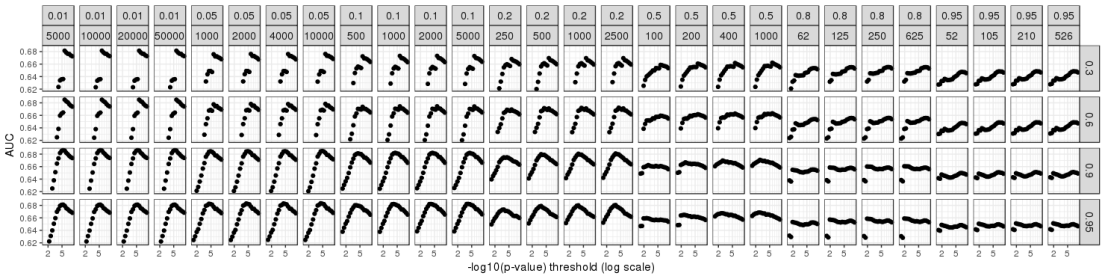
(b) Rheumatoid arthritis



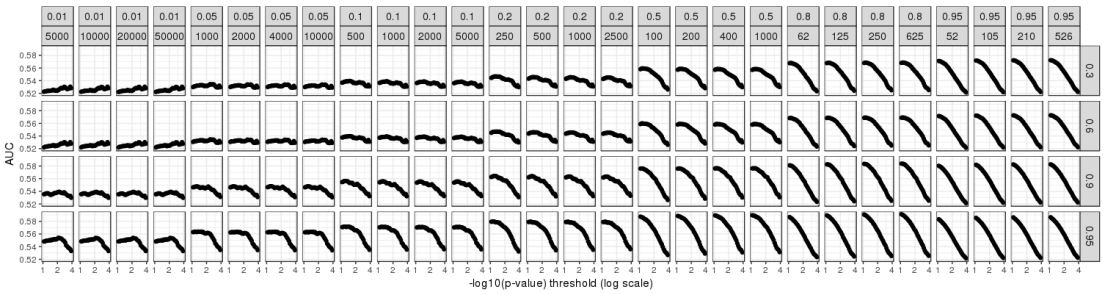
(c) Type 1 diabetes



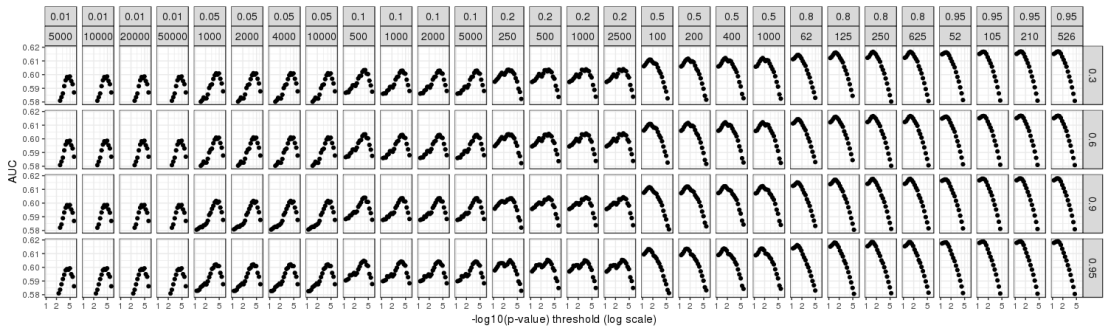
(d) Type 2 diabetes



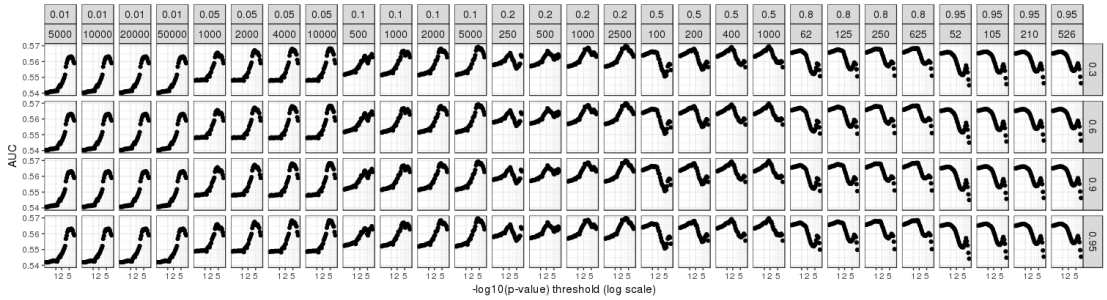
(e) Prostate cancer



(f) Depression



(g) Coronary artery disease



(h) Asthma

Figure S6: AUC values (for the training set) when predicting disease status for many parameters of C+T in real data applications. Facets are presenting different clumping thresholds r_c^2 from 0.01 to 0.95, window sizes w_c from 52 to 50,000 kb, and imputation thresholds from 0.3 to 0.95. The x-axis corresponds to the remaining hyper-parameter, the p-value threshold p_T; here, -log10(p-values) are represented using a logarithmic scale.

Caution on using covariates

For example, because prevalence of CAD is much higher in men than in women in the UKBB (8-9% vs 2%), adding sex in the model amount to fitting two different intercepts, centering distributions of fitted probabilities around disease prevalence (Figure S7). This increases the AUC from 63.9% to 74.4% but results in a model that would classify all women as healthy. A possible solution would be to report AUC figures for each gender separately, or even to fit a model for each gender separately (in the stacking step). Fitting models separately would enable the use of sex chromosomes without introducing bias. As for ancestry concerns, fitting different models for different ancestries might be a way to get more calibrated results and to account for differences in effect sizes and LD. However, here for CAD, fitting two separate models for each gender results in a slight loss of predictive performance, while using variable 'sex' does not change results when they are reported for each gender separately, with an AUC of 64.9% [63.5-66.3] for men and 62.5% [59.8-65.2] for women. Thus, adding 'sex' as a covariate in the model may provide a model with similar discrimination and with better calibration of probabilities (if prevalence in the data is representative of prevalence in the population). Yet, we would like to emphasize again that reporting one AUC figure for all individuals would be misleading in the case of using variable 'sex' in the model.

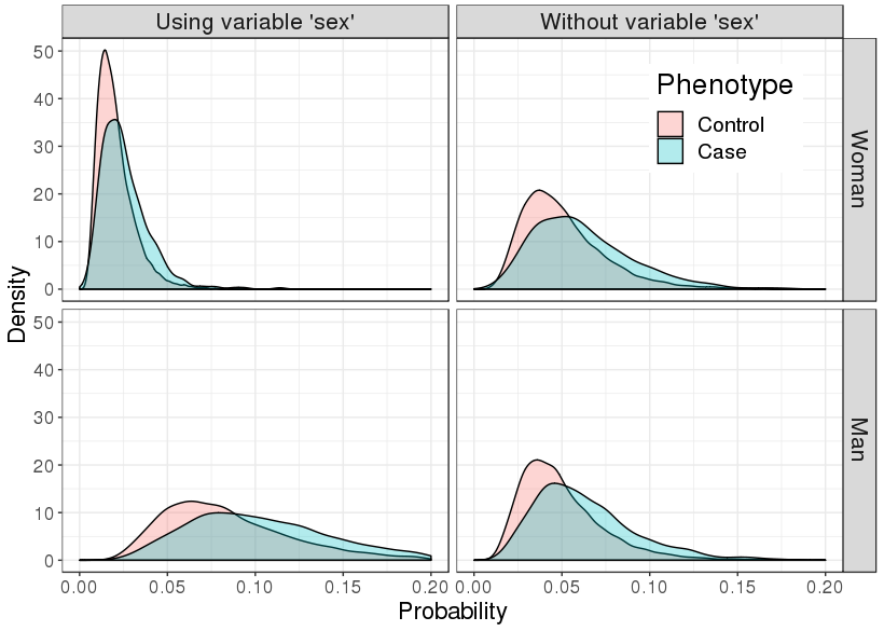


Figure S7: Distribution of predicted probabilities of Coronary Artery Disease (CAD) in the UK Biobank using SCT. Upper / lower panels corresponds to women / men. Left panels correspond to a model using C+T scores and variable 'sex' when fitting penalized logistic regression in the stacking step. Right panels correspond to performing stacking of C+T scores without using variable 'sex'.

Chapter 5

Conclusion and Discussion

5.1 Summary of my work

The first part of my work has consisted in developing tools to easily analyze genotype matrices. There are different software using different input formats so that they are sometimes difficult to use in the same analysis. These software are of tremendous utility for the research community because they efficiently implement some of the validated analyses that are used in genetics, such as performing GWAS or heritability estimation. Yet, if you want to do some exploratory analysis, looking at new ideas, modifying the code a little, it is practically impossible to do so. I understood that I would need some kind of standard matrix format if I wanted to use simple code for explanatory analyses and to develop new ideas. So, I started to develop R package bigsnpr. There is no better way for understanding methods than to implement them. At some point, I realized that lots of methods I was using and reimplementing for the on-disk matrix format was just some standard statistical tools that would be useful for other fields too. So, I put all these functions (PCA, multiple association tests, numerical summaries, matrix products, etc.) in another R package called bigstatsr that can be used by other people outside the field of genetics. Hopefully, this package will become useful for many people as data are getting larger in other fields too. I have spent a lot of time documenting, testing and optimizing the code in these two R packages. For example, you can now do an association analysis for a continuous outcome in no time thanks to the use of some linear algebra tricks (see the Appendix). I have also spent some time exploring some genotype datasets

and found for example that standard software for computing PCA such as PLINK and FastPCA sometimes are not accurate enough, i.e. that the approximation they use does not give the same results as when using an exact PCA implementation. Moreover, I have found that one should be extra careful about the SNPs that are used in PCA to avoid capturing something else than population structure, such as LD structure[1]; I developed the "autoSVD" procedure to detect those SNPs automatically and remove them.

Then, I developed some efficient implementation of penalized (linear and logistic) regressions. Two efficient implementations were already available for these models, but those implementations did not scale well with the very large datasets we have in the field. For example, they could not be used to analyze the UK Biobank data. It is now possible to do so with the implementation we provide in package bigstatsr. The difference in computation time resides mainly in the use of some early stopping criterion in our implementation. We also provide a way to choose the two hyper-parameters of elastic net regularization so that the user does not have to choose them arbitrarily or to implement a cross-validation framework. We extensively compared the predictive performance of our implementation of penalized regressions with standard methods such as C+T, where SNP effects are learned independently before being combined using heuristics. We showed that for large sample sizes, penalized regressions are able to capture very small effects and that prediction is improved as compared to C+T. For example, we are able to predict 43% of the variance in height, which represents almost all heritability of height that can be captured by standard genotyping chips (Yang *et al.*, 2010; Lello *et al.*, 2018).

Finally, we focused on developing a predictive method that uses summary statistics. We first made it possible to derive the widely used C+T method for many hyper-parameters, using an efficient implementation. We showed that choosing over a wider range of hyper-parameter values as compared to the current practice of using C+T could substantially improve predictive performance of C+T. We then proposed to stack all those C+T predictors instead of choosing the best one. Stacking corresponds to finding an optimal combination of different predictors in order to get higher predictive performance than any single of these predictors. We called this method SCT, which stands for Stacked Clumping and Thresholding. We showed that when using external summary statistics and the UK Biobank data, we could substantially improve prediction over any

[1] `https://privefl.github.io/bigsnpr/articles/how-to-PCA.html`

C+T model.

Thus, overall we developed tools to analyze large matrices, especially genotype matrices, possibly in dosage format. We then proposed two methods for building polygenic predictive models, one based on individual-level data (that could use summary statistics to prioritize SNPs in the model), and one based on large summary statistics and individual-level data. These two methods provide ones of the currently best predictive performance for many diseases and traits.

5.2 Problem of generalization

Polygenic Risk Scores (PRS) might become a central part in precision medicine. For now, predictive performance for most complex diseases are not good enough to be used in clinical settings. A major concern with PRS at the moment is their problem of generalization / transferability in different populations. Indeed, most GWAS have included European people only (Figure 5.1). In 2009, 96% of individuals included in GWAS datasets were of European ancestry (Need and Goldstein, 2009). In 2016, still more than 80% of those individuals were of European descent, with an increase of the inclusion of non-European participants, mostly constituted of Asian people (Popejoy and Fullerton, 2016). People from Hispanic or African ancestry are still poorly represented (Martin *et al.*, 2019). This poor heterogeneity in inclusion can be explained by the fact that the more diverse are the population in the data we analyze, the more possible confounders there are to account for in order to avoid spurious results (Popejoy and Fullerton, 2016).

This lack of heterogeneity in inclusion of diverse populations results in several problems. First, there are some SNP ascertainment bias because SNPs that are more common are more likely to be discovered in GWAS so that associated SNPs tend to have larger frequencies in European than in other populations, due to the winner's curse. If alleles have a frequency that is different between populations, using the corresponding effects on disease naturally introduces some shift in PRS distributions for different populations. Second, rare variants are missed in GWAS if they are specific to some population that is not included in the association study (Martin *et al.*, 2019). Thus, this limit the predictive ability of PRS in different populations to the one(s) included in the GWAS. Third, it is accepted that genotyped SNPs, or even imputed SNPs, that are discovered in GWAS may not be true functional SNPs (fSNPs) having an effect on disease susceptibility. In-

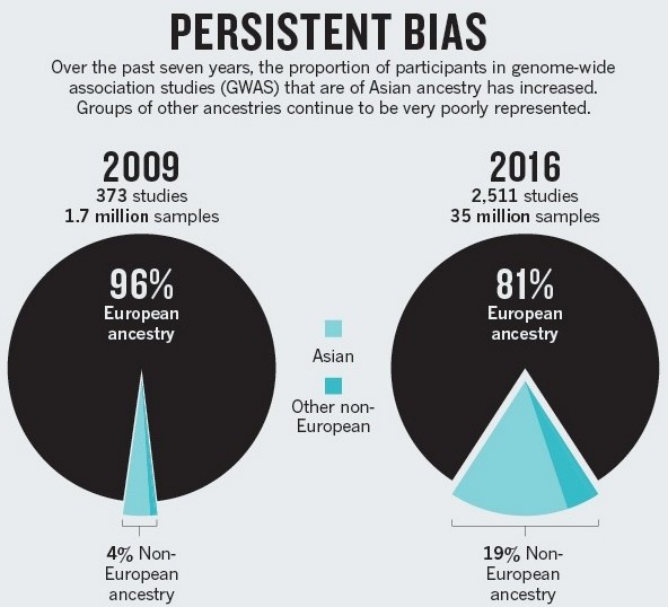


Figure 5.1: Proportion of GWAS participants by ancestry. Most GWAS include mainly European people, some now include Asian people, but other ethnicities are still poorly represented. Source: Popejoy and Fullerton (2016).

stead, GWAS are assumed to discover SNPs that tag fSNPs (tagSNPs), i.e. are correlated with fSNPs. Yet, LD may be different between populations so that a tagSNP can have a different correlation with the corresponding fSNP. Thus, effects of these SNPs can be different and are often diluted toward zero for populations not included in the GWAS (Carlson *et al.*, 2013). In conclusion, for many reasons, magnitude and frequency of effects can vary considerably between populations, and these differences are larger when populations are more genetically distant such as African population with either European or Asian populations. These differences in prediction between populations are two-fold (Figures 5.2 and 5.3): distributions of PRS are shifted and prediction within each distribution is also reduced (Vilhjálmsson *et al.*, 2015; Martin *et al.*, 2019).

Several solutions have been proposed to partially correct for the differences of prediction between populations. First, Martin *et al.* (2017) proposed to mean-center PRS for each population, yet this would require an accurate way to assess ancestry and would not work for admixed people, e.g. one person with a father of African ancestry and a mother of European ancestry (Reisberg *et al.*, 2017). Second, it has been suggested to include more diverse population in GWAS (Pulit *et al.*, 2010). Indeed, new associations

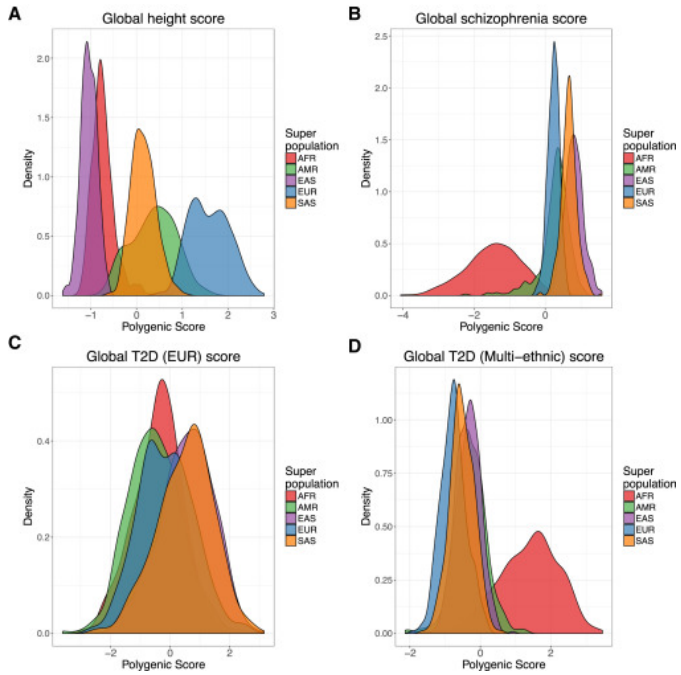


Figure 5.2: Distributions of Polygenic Risk Scores (PRS) for many populations and phenotypes (T2D: type 2 diabetes). Source: Martin *et al.* (2017).

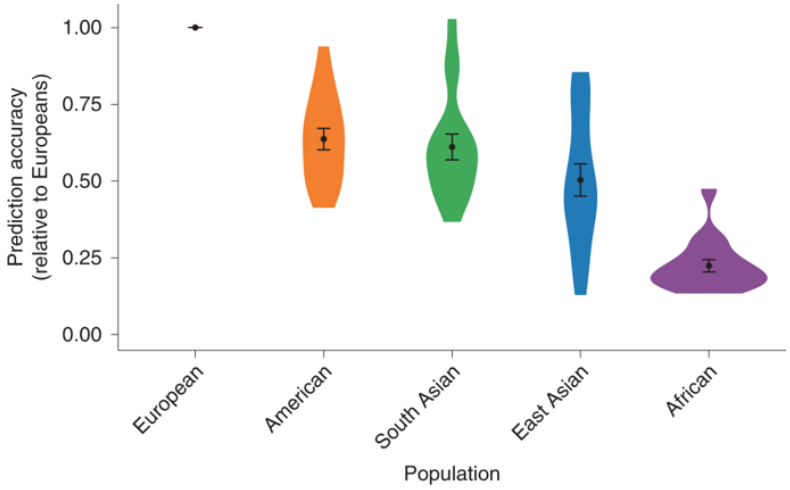


Figure 5.3: Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UK Biobank data (Bycroft *et al.*, 2018). Source: Martin *et al.* (2019).

can be found if the frequency is higher in an under-represented population. It would be also possible to fine-map fSNPs in common for multiple populations so that their effects generalize better to any population, irrespective of LD (Carlson *et al.*, 2013; Mägi *et al.*, 2017; Wojcik *et al.*, 2018). Finally, statistical methods are being developed to use large European GWAS in conjunction with smaller data from another population in order to leverage both the discoveries from the large dataset and the specificities of the smaller dataset (Márquez-Luna *et al.*, 2017; Coram *et al.*, 2017).

5.3 Looking for missing heritability in rare variants

"Still missing heritability", i.e. the gap between heritability estimations from current GWAS studies and from family studies, could reside in rare variants. Indeed, for height and colorectal cancer, it has been shown that estimations of heritability from GWAS data could recover almost all heritability when a large proportion of low-frequency variants was present in the data (Yang *et al.*, 2015; Huyghe *et al.*, 2019; Wainschtein *et al.*, 2019). However, actual findings of significantly associated variants of low-frequency are scarce. For example, a GWAS of height including more than 700K individuals found 83 associated variants with allele frequencies between 0.1% and 4.8%, with effects up to 2 cm per allele (Marouli *et al.*, 2017). Yet, these 83 variants together accounts for only 1.7% of the total heritability of height. In other large studies, one for coronary artery disease and one for type 2 diabetes, there was little evidence of low-frequency variants with large effects (Nikpay *et al.*, 2015; Fuchsberger *et al.*, 2016).

Associations of rare variants with traits are difficult to find for two reasons. First, it is very difficult to impute low-frequency variants with a good quality if using for example a small reference panel such as the 1000 genomes (Nikpay *et al.*, 2015). There is now a reference panel of 32,000 individuals that is used to accurately impute variants with allele frequencies as low as 0.1% (McCarthy *et al.*, 2016). This large reference panel is European specific, which means that imputing data from other ancestries is more difficult. This is a problem because, one way to discover and accurately estimate the effect of a rare variant is to look for it in a population in which its allele frequency is larger (Moltke *et al.*, 2014; Minster *et al.*, 2016). Indeed, the power of association studies is dependent on the variance explained by a locus and its frequency; for example, for a disease that affects 1% of the population, we have the same power to detect a risk

locus of 50% frequency and odds ratio of 1.1 as we do for a risk locus of 0.1% frequency and odds ratio of 2.9 (Wray *et al.*, 2018). Fortunately, ongoing large-scale projects, such as the Trans-Omics for Precision Medicine (TOPMed) program, are expected to produce reference panels of more than 100,000 individuals including diverse populations (Taliun *et al.*, 2019).

The second reason for which rare variant associations are difficult to find is that sequencing technologies are more expensive than genotyping and imputation. Currently, studies have mostly focused on whole exome sequencing (WES) because it is cheaper than whole genome sequencing (WGS). Indeed, the exome is where the effect sizes of variants are expected to be larger and where discoveries are likely to be more immediately actionable (Zuk *et al.*, 2014). Yet, sample sizes of sequencing studies remain small and special considerations and challenges arise when testing rare frequency variants from these studies (Auer and Lettre, 2015). Thus, sample size is the limiting factor in variant discovery, not genotyping technology (Wray *et al.*, 2018). It is probably the limiting factor in prediction too.

5.4 Looking for missing heritability in non-additive effects

Knowledge about biological pathways and gene networks implies that epistasis (gene interactions) might be important to consider (Hill *et al.*, 2008). Apart from explaining missing heritability, genetic interactions could also create phantom heritability, i.e. could make current estimation of heritability upward biased (Zuk *et al.*, 2012). There have been some findings of interaction between loci, but mainly for autoimmune diseases for which there are strong effects in regions of chromosome 6 that have an effect on the autoimmune system (Lenz *et al.*, 2015; Goudey *et al.*, 2017). Yet, these interaction effects explain little to phenotypic variance as compared to additive effects (Lenz *et al.*, 2015). In general, data and theory point to mainly additive genetic variance (Hill *et al.*, 2008).

Moreover, interactions are challenging to find for two reasons, and dedicated methods to epistasis detection have been implemented (Niel *et al.*, 2015). First, it is analytically impractical to search for such interaction effects because it would require testing

more than 100 billion pairs of variants, even for a small genotyping array. Second, because of this huge number of tests, correction for multiple testing allows the detection of highly significant interactions only.

Finally, even if we find such interaction effects, they are unlikely to dramatically improve risk prediction for complex diseases, but could still provide insights into their ethiology (Aschard *et al.*, 2012). Moreover, due to differences in effect sizes and LD between populations, epistatic effects are even more unlikely than additive effects to replicate to different populations (Hill *et al.*, 2008; Visscher *et al.*, 2017).

5.5 Integration of multiple data sources

There are many genetic data out there. Some large individual-level data such as the UK biobank are available (Bycroft *et al.*, 2018). When GWAS data is not publicly available, summary statistics are often publicly shared instead. Usually, predictive models are based on either individual-level data (e.g. penalized regression) or summary statistics (e.g. C+T). Building models that combine both individual-level data and summary statistics, possibly including different populations, is necessary to increase predictive power. We started to do this by implementing the SCT method in our latest paper, where we combine several summary statistics based predictors using large individual-level data. Alike with the adaptive lasso (Zou, 2006), one could also think of penalizing SNPs differently in individual-level data methods, applying a penalization factor to each SNP based on their significance or effect sizes in external summary statistics.

Human diseases are inherently complex and governed by the complicated interplay of several underlying factors (Dey *et al.*, 2013). For a trait or a disease, prediction based on genetic data only is ultimately capped by heritability. Therefore, prediction must integrate other types of data if we want to predict beyond the limit of heritability (Figure 5.4). For example, DNA methylation data can accurately predict age of any tissue across the entire life course (Horvath, 2013; Horvath and Raj, 2018), gene expression profiles enable to gain a broad picture of the genomic response to environmental perturbation (Gibson, 2008) and microbiota can also be an important "environmental" factor to take into account (Bäckhed *et al.*, 2004). Yet, integrating variables with different formats, types, structure, dimensionality and missing values is a challenging problem (Dey *et al.*, 2013). One could integrate genetic data with clinical data. For example,

Inouye *et al.* (2018) designed a polygenic risk score (PRS) with higher discriminative ability for coronary artery disease than any of 6 conventional risk factors (smoking, diabetes, hypertension, body mass index, high cholesterol and family history). Using this PRS with all 6 conventional risk factors increases discriminative ability as compared to using the PRS only or the 6 factors only. Moreover, electronic health records (EHR) make possible to integrate large biobank datasets with large clinical, environmental and phenotypic information (Roden and Denny, 2016).

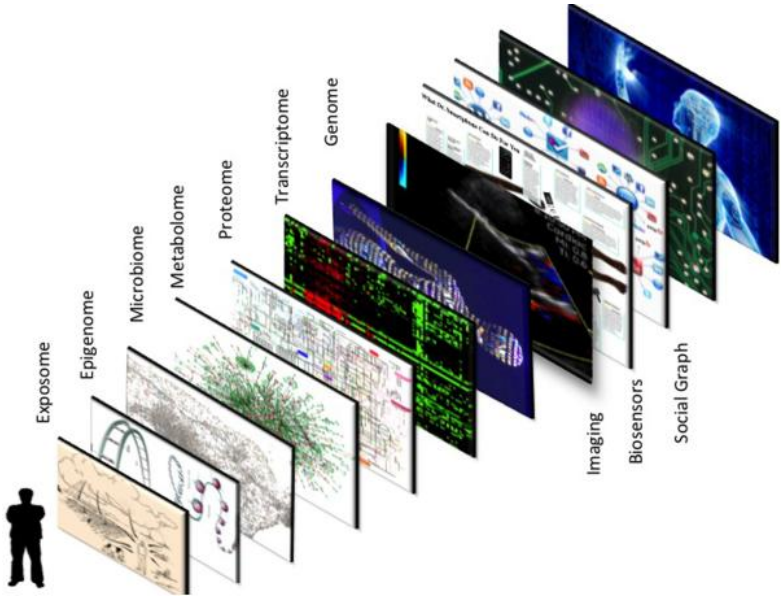


Figure 5.4: Geographic information system of a human being. The different layers of data available for an individual. Source: Topol (2014).

5.6 Future work

I will probably continue to work in the field of predictive human genetics. I am currently visiting the National Center for Register-based Research (NCRR) in Aarhus, Denmark. Researchers there are mostly epidemiologists using national registers where they have information on all Danes over decades. Most of their work is funded to look at psychiatric disorders and they are now interested in how genetics influence psychiatric conditions. It is a good opportunity to work on a large national biobank dataset with Bjarni Vilhjálmsson.

I would be interested in looking at many things. First, I think investigating which method works best in which scenario is of great interest for the field. Many scenarios could involve different sample sizes of summary statistics and individual-level data, but also training and prediction in different populations. Such work could be useful to make some guidelines about which method to use in which situation. For example, individual-level data methods often work best when large individual-level data are available, but what about predicting in a different population where only smaller datasets are available?

Second, it would be interesting to account for age in the prediction, for example extending with Cox regression the methods I implemented. Many diseases such as cancer, heart diseases and Alzheimer's disease have an age component; modeling this age component and accounting for right censoring (people who might develop disease later) should increase predictive performance and usefulness of models.

Third, I would like to investigate more about imputation. At the moment, imputed data is taken for granted. How to properly account for imputation accuracy in association testing (using e.g. multiple imputation) and in predictive models?

Finally, other ideas could be to investigate how we can integrate many sources of information such as functional annotations, looking at many phenotypes at once, or to distinguish between two diseases with similar symptoms (e.g. diabetes) using polygenic risk scores.

Bibliography

Abraham, G. and Inouye, M. (2015). Genomic risk prediction of complex human disease and its clinical application. *Current opinion in genetics & development*, **33**, 10–16.

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.

Anglian Breast Cancer Study Group *et al.* (2000). Prevalence and penetrance of brca1 and brca2 mutations in a population-based series of breast cancer cases. *British Journal of Cancer*, **83**(10), 1301.

Aschard, H., Chen, J., Cornelis, M. C., Chibnik, L. B., Karlson, E. W., and Kraft, P. (2012). Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *The American Journal of Human Genetics*, **90**(6), 962–972.

Astle, W., Balding, D. J., *et al.* (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, **24**(4), 451–471.

Auer, P. L. and Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome medicine*, **7**(1), 16.

Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., Semenkovich, C. F., and Gordon, J. I. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences*, **101**(44), 15718–15723.

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. (1996). Stacked regressions. *Machine learning*, **24**(1), 49–64.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., of the Psychiatric Genomics Consortium, S. W. G., *et al.* (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, **47**(3), 291.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203.

Carlson, C. S., Matise, T. C., North, K. E., Haiman, C. A., Fesinmeyer, M. D., Buyske, S., Schumacher, F. R., Peters, U., Franceschini, N., Ritchie, M. D., *et al.* (2013). Generalization and dilution of association results from european gwas in populations of non-european ancestry: the page study. *PLoS biology*, **11**(9), e1001661.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**(1), 7.

Chasioti, D., Yan, J., Nho, K., and Saykin, A. J. (2019). Progress in polygenic composite scores in alzheimer's and other complex diseases. *Trends in Genetics*.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Choi, S. W., Mak, T. S. H., and O'reilly, P. (2018). A guide to performing polygenic risk score analyses. *BioRxiv*, page 416545.

Chun, S., Imakaev, M., Hui, D., Patsopoulos, N. A., Neale, B. M., Kathiresan, S., Stitziel, N. O., and Sunyaev, S. R. (2019). Non-parametric polygenic risk prediction using partitioned gwas summary statistics. *BioRxiv*, page 370064.

Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L., and Tang, H. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *The American Journal of Human Genetics*, **101**(2), 218–226.

DeSantis, C. E., Fedewa, S. A., Goding Sauer, A., Kramer, J. L., Smith, R. A., and Jemal, A. (2016). Breast cancer statistics, 2015: Convergence of incidence rates between black and white women. *CA: a cancer journal for clinicians*, **66**(1), 31–42.

Dey, S., Gupta, R., Steinbach, M., and Kumar, V. (2013). Integration of clinical and genomic data: a methodological survey. *Briefings in Bioinformatics*.

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

Editors of Nature Genetics (2012). Asking for more. *Nature Genetics*, **44**(7), 733–733.

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, **31**(9), 1466–1468.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., *et al.* (2016). The genetic architecture of type 2 diabetes. *Nature*, **536**(7614), 41.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via bayesian regression and continuous shrinkage priors. *bioRxiv*, page 416859.

Gibson, G. (2008). The environmental contribution to gene expression profiles. *Nature reviews genetics*, **9**(8), 575.

Golan, D. and Rosset, S. (2014). Effective genetic-risk prediction using mixed models. *The American Journal of Human Genetics*, **95**(4), 383–393.

Goudey, B., Abraham, G., Kikianty, E., Wang, Q., Rawlinson, D., Shi, F., Haviv, I., Stern, L., Kowalczyk, A., and Inouye, M. (2017). Interactions within the mhc contribute to the genetic architecture of celiac disease. *PloS one*, **12**(3), e0172826.

Grande, B. M., Baghela, A., Cavalla, A., Privé, F., Zhang, P., and Zhen, Y. (2018). Hackathon-driven tutorial development. *F1000Research*, **7**.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, **33**(suppl_1), D514–D517.

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics*, **4**(2), e1000008.

Horvath, S. (2013). Dna methylation age of human tissues and cell types. *Genome biology*, **14**(10), 3156.

Horvath, S. and Raj, K. (2018). Dna methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*, page 1.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, **159**(4), 1805–1817.

Huyghe, J. R., Bien, S. A., Harrison, T. A., Kang, H. M., Chen, S., Schmit, S. L., Conti, D. V., Qu, C., Jeon, J., Edlund, C. K., *et al.* (2019). Discovery of common and rare genetic risk variants for colorectal cancer. *Nature genetics*, **51**(1), 76.

Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., Lai, F. Y., Kaptoge, S., Brozynska, M., Wang, T., *et al.* (2018). Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *Journal of the American College of Cardiology*, **72**(16), 1883–1893.

Jakobsdottir, J., Gorin, M. B., Conley, Y. P., Ferrell, R. E., and Weeks, D. E. (2009). Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS genetics*, **5**(2), e1000337.

Janssens, A. C. J. and Joyner, M. J. (2019). Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: Is more, better? *Clinical chemistry*, pages clinchem–2018.

Janssens, A. C. J., Aulchenko, Y. S., Elefante, S., Borsboom, G. J., Steyerberg, E. W., and van Duijn, C. M. (2006). Predictive testing for complex diseases using multiple genes: fact or fiction? *Genetics in medicine*, **8**(7), 395.

Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., Jervis, S., Van Leeuwen, F. E., Milne, R. L., Andrieu, N., *et al.* (2017). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *Jama*, **317**(23), 2402–2416.

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de los Campos, G., and Hsu, S. D. (2018). Accurate genomic prediction of human height. *Genetics*, **210**(2), 477–497.

Lenz, T. L., Deutsch, A. J., Han, B., Hu, X., Okada, Y., Eyre, S., Knapp, M., Zhernakova, A., Huizinga, T. W., Abecasis, G., *et al.* (2015). Widespread non-additive and interaction effects within hla loci modulate the risk of autoimmune diseases. *Nature genetics*, **47**(9), 1085.

Lin, D. and Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, **34**(1), 60–66.

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature genetics*, **50**(7), 906.

Luu, K., Bazin, E., and Blum, M. G. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular ecology resources*, **17**(1), 67–77.

Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M. I., COGENT-Kidney Consortium, T.-G. C., and Morris, A. P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Human Molecular Genetics*, **26**(18), 3639–3650.

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.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, **41**(6), 469–480.

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

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, **11**(7), 499.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., *et al.* (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, **542**(7640), 186.

Márquez-Luna, C., Loh, P.-R., Consortium, S. A. T. . D. S., Consortium, S. T. . D., and Price, A. L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic epidemiology*, **41**(8), 811–823.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., and Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, **100**(4), 635–649.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, **51**(4), 584.

Mavaddat, N., Pharoah, P. D., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., Wang, Q., Dennis, J., Dunning, A. M., Shah, M., *et al.* (2015). Prediction of breast cancer risk based on profiling with common genetic variants. *JNCI: Journal of the National Cancer Institute*, **107**(5).

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., *et al.* (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, **48**(10), 1279.

Minster, R. L., Hawley, N. L., Su, C.-T., Sun, G., Kershaw, E. E., Cheng, H., Buhule, O. D., Lin, J., Tuitele, J., Naseri, T., *et al.* (2016). A thrifty variant in crebrf strongly influences body mass index in samoans. *Nature genetics*, **48**(9), 1049.

Moltke, I., Grarup, N., Jørgensen, M. E., Bjerregaard, P., Treebak, J. T., Fumagalli, M., Korneliussen, T. S., Andersen, M. A., Nielsen, T. S., Krarup, N. T., *et al.* (2014). A common greenlandic tbc1d4 variant confers muscle insulin resistance and type 2 diabetes. *Nature*, **512**(7513), 190.

Need, A. C. and Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, **25**(11), 489–494.

Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G., *et al.* (2008). The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, **83**(3), 347–358.

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in genetics*, **6**, 285.

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., *et al.* (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, **47**(10), 1121.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., and Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. *PLoS genetics*, **10**(11), e1004754.

Pasaniuc, B. and Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, **18**(2), 117.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., *et al.* (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature genetics*, **44**(6), 631.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, **30**(20), 2906–2914.

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.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology*, **32**(4), 381–385.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology*, **159**(9), 882–890.

Popejoy, A. B. and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature News*, **538**(7624), 161.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**(8), 904.

Pritchard, J. K. and Cox, N. J. (2002). The allelic architecture of human disease genes: common disease–common variant...or not? *Human molecular genetics*, **11**(20), 2417–2423.

Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, **34**(16), 2781–2787.

Privé, F., Aschard, H., and Blum, M. G. (2019). Efficient implementation of penalized regression for genetic risk prediction. *Genetics*, pages genetics–302019.

Pulit, S. L., Voight, B. F., and De Bakker, P. I. (2010). Multiethnic genetic association studies improve power for locus discovery. *PloS one*, **5**(9), e12600.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., *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.

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.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reisberg, S., Iljasenko, T., Läll, K., Fischer, K., and Vilo, J. (2017). Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PloS one*, **12**(7), e0179238.

Roden, D. M. and Denny, J. C. (2016). Integrating electronic health record genotype and phenotype datasets to transform patient care. *Clinical Pharmacology & Therapeutics*, **99**(3), 298–305.

Salari, K., Watkins, H., and Ashley, E. A. (2012). Personalized medicine: hope or hype? *European heart journal*, **33**(13), 1564–1570.

Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., Pervjakova, N., Pers, T. H., Johnson, A. D., Eicher, J. D., *et al.* (2017). An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes*, **66**(11), 2888–2902.

Sikorska, K., Lesaffre, E., Groenen, P. F., and Eilers, P. H. (2013). Gwas on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC bioinformatics*, **14**(1), 166.

Silventoinen, K. (2003). Determinants of variation in adult body height. *Journal of biosocial science*, **35**(2), 263–285.

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., Chiang, C. W., Hirschhorn, J., Daly, M. J., Patterson, N., *et al.* (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, **8**, e39702.

Speed, D. and Balding, D. J. (2014). Multiblup: improved snp-based prediction for complex traits. *Genome research*, **24**(9), 1550–1557.

Speed, D. and Balding, D. J. (2018). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature genetics*, page 1.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., *et al.* (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *BioRxiv*, page 563866.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, page 1.

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.

Topol, E. J. (2014). Individualized medicine from prewomb to tomb. *Cell*, **157**(1), 241–253.

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, page 1.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., *et al.* (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, **97**(4), 576–592.

Visscher, P. M., Medland, S. E., Ferreira, M. A., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., and Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, **2**(3), e41.

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.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, **101**(1), 5–22.

Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., Psaty, B. M., *et al.* (2019). Recovery of trait heritability from whole genome sequence data. *bioRxiv*, page 588020.

Wald, N. J. and Old, R. (2019). The illusion of polygenic disease risk prediction. *Genetics in Medicine*, page 1.

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.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., *et al.* (2013). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, **42**(D1), D1001–D1006.

Wojcik, G., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., *et al.* (2018). The page study: How genetic diversity improves our understanding of the architecture of complex traits. *bioRxiv*, page 188094.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., *et al.* (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, **46**(11), 1173.

Wray, N. R., Goddard, M. E., and Visscher, P. M. (2008). Prediction of individual genetic risk of complex disease. *Current opinion in genetics & development*, **18**(3), 257–263.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., and Middeldorp, C. M. (2014). Research review: polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, **55**(10), 1068–1087.

Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J., and Visscher, P. M. (2018). Common disease is more complex than implied by the core gene omnigenic model. *Cell*, **173**(7), 1573–1580.

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.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, **46**(2), 100.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., Robinson, M. R., Perry, J. R., Nolte, I. M., van Vliet-Ostaptchouk, J. V., *et al.* (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, **47**(10), 1114.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., *et al.* (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, **50**(9), 1335.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, **109**(4), 1193–1198.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, **111**(4), E455–E464.

Appendix A

Code optimization based on linear algebra

A.1 Lightning fast multiple association testing

.

Here, I describe how to quickly test many variables for an association with a continuous outcome of interest. For example, let us make a Genome-Wide Association Study (GWAS) of height, i.e. we want to determine which genome variants are associated with height.

The model we want to test is

$$y = \beta s + X\gamma + \epsilon \,,$$

where s is one variant (we want to do this for each variant, separately), X are some covariates to adjust for some possible confounding factors (a matrix of N samples over K columns, including a column of 1s to account for an intercept in the model). We are only interested in estimating $\hat{\beta}$ and computing a p-value corresponding to the significance of the alternative hypothesis that $\beta \neq 0$.

Sikorska *et al.* (2013) show that we can rewrite this problem as

$$y^* = \beta s^* + \epsilon \,,$$

where $y^* = y - X(X^TX)^{-1}X^Ty$ and $s^* = s - X(X^TX)^{-1}X^Ts$. Thus, this becomes a simple linear problem which is easy and fast to solve. We have

$$\begin{aligned}
\hat{\beta} &= \frac{s^{*T}y^*}{s^{*T}s^*} \,, \\
\widehat{\text{var}}(\hat{\beta}) &= \frac{(y^* - \hat{\beta}s^*)^T(y^* - \hat{\beta}s^*)}{(N-K-1)\; s^{*T}s^*} \,, \\
\frac{\hat{\beta}}{\sqrt{\widehat{\text{var}}(\hat{\beta})}} &\sim T(N-K-1) \,.
\end{aligned}$$

We extend this idea further by computing the singular value decomposition $X = U\Delta V^T$ ($N \times K$ matrix). As $N \gg K$, we have $U^TU = I_K$, $V^TV = I_K$ and $VV^T = I_K$. Thus $X(X^TX)^{-1}X^T = U\Delta V^T(V\Delta U^TU\Delta V^T)^{-1}V\Delta U^T = U\Delta V^T(V\Delta^2V^T)^{-1}V\Delta U^T = U\Delta V^T(V\Delta^{-2}V^T)V\Delta U^T = UU^T$. Then, we can simplify $s^{*T}y^* = (s - UU^Ts)^Ty^* = s^Ty^* - s^T\underbrace{UU^Ty^*}_{0} = s^Ty^*$, $s^{*T}s^* = (s - UU^Ts)^T(s - UU^Ts) = s^Ts - 2s^TUU^Ts + s^TUU^TUU^Ts = s^Ts - s^TUU^Ts = s^Ts - z^Tz$, where $z = U^Ts$, and $(y^* - \hat{\beta}s^*)^T(y^* - \hat{\beta}s^*) = y^{*T}y^* - 2\hat{\beta}s^{*T}y^* + \hat{\beta}^2s^{*T}s^* = y^{*T}y^* - 2\hat{\beta}s^{*T}y^* + \hat{\beta}s^{*T}y^* = y^{*T}y^* - \hat{\beta}s^Ty^*$. So, we only need to compute

$$\begin{aligned}
z &= U^Ts \,, \\
\hat{\beta}_{\text{num}} &= s^Ty^* \,, \\
\hat{\beta}_{\text{deno}} &= s^Ts - z^Tz \,, \\
\hat{\beta} &= \hat{\beta}_{\text{num}}/\hat{\beta}_{\text{deno}} \,, \\
\widehat{\text{var}}(\hat{\beta}) &= \frac{y^{*T}y^* - \hat{\beta}\;\hat{\beta}_{\text{num}}}{(N-K-1)\;\hat{\beta}_{\text{deno}}} \,.
\end{aligned}$$

Since U and y^* are computed only once for all variants, you can apply those formulas to compute these statistics for 1,000,000 variants and N=500,000 samples and K=11 covariates in one hour only (Privé *et al.*, 2018). This is implemented in function `big_univLinReg()` of package bigstatsr.

A.2 Implicit scaling of a matrix

The matrix formulation of column scaling is $\tilde{X} = C_n X S$, where $C_n = I_n - \frac{1}{n} 1_n 1_n^T$ is the centering matrix[1] and S is a diagonal matrix with the scaling coefficients (typically, $S_{j,j} = 1/\text{sd}_j$).

In algorithms such as Principal Component Analysis (PCA) or multiple linear regression, we must compute e.g. $\tilde{X}V$ and $\tilde{X}^T\tilde{X}$, where V is another matrix. We can show how to compute these products without explicitly scaling the matrix X. This is really useful when working with on-disk matrices such as in R package bigstatsr, because you do not need to compute (and store) an intermediate scaled matrix.

For example, for computing products, $\tilde{X}V = C_n X S V = C_n(X(SV))$. So, you can compute $\tilde{X}V$ without explicitly scaling X. Another example, for computing self cross-products, $\tilde{X}^T\tilde{X} = (C_n X S)^T \cdot C_n X S = S^T X^T C_n X S$ ($C_n^2 = C_n$ is intuitive because centering an already centered matrix does not change it). Then, $\tilde{X}^T\tilde{X} = S^T X^T (I_n - \frac{1}{n} 1_n 1_n^T) X S = S^T (X^T X - X^T (\frac{1}{n} 1_n 1_n^T) X) S = S^T (X^T X - \frac{1}{n} s_X * s_X^T) S$ where s_X is the vector of column sums of X.

This implicit scaling can be quite useful if you manipulate very large matrices because you are not copying the matrix nor making useless computation. For example, this can be used to compute a correlation matrix 20 times as fast as base R function `cor()`[2].

[1] https://en.wikipedia.org/wiki/Centering_matrix

[2] https://privefl.github.io/blog/(Linear-Algebra)-Do-not-scale-your-matrix/