RápidoPGS: A rapid polygenic score calculator for summary GWAS data without validation dataset

Guillermo Reales[1,2], Elena Vigorito[3], Martin Kelemen[1,4], Chris Wallace[1,2,3]

[1]Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Puddicombe Way, Cambridge CB2 0AW, UK.
[2]Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, CB2 2QQ, UK.
[3]MRC Biostatistics Unit, University of Cambridge, School of Clinical Medicine, East Forvie Building, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK.
[4]Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK.

Correspondence: Dr. Chris Wallace (cew54@cam.ac.uk)

Category: Genetics and Population Analysis

Abstract

Motivation: Polygenic scores (PGS) aim to genetically predict complex traits at an individual level. PGS are typically trained on genome-wide association summary statistics and require an independent test dataset to tune parameters. More recent methods allow parameters to be tuned on the training data, removing the need for independent test data, but approaches are computationally intensive. Based on fine-mapping principles, we present RápidoPGS, a flexible and fast method to compute PGS requiring only summary-level GWAS datasets and no test data.
Results: We show that RápidoPGS performs well in comparison to a state of the art internally tuned method LDpred2 (median AUC difference = 0.0025), with over 1000-fold improved speed. RápidoPGS is implemented in R, and can work with user-supplied summary statistics or download them from the GWAS catalog.
Availability and implementation: Our method is available with a GPL license as an R package from GitHub.

1. Introduction

Genome-wide association studies (GWAS) have been widely successful at identifying a large number of genetic variants (usually single nucleotide polymorphisms, or SNPs), associated with a wide range of diseases and complex traits (Buniello *et al.*, 2019). Most genetic variants have a small individual effect on the tested traits and thus have low predictive power (Yang *et al.*, 2010; Dudbridge, 2013; International Schizophrenia Consortium *et al.*, 2009). However, simultaneously evaluating the effects of all common SNPs has the potential to explain much of the heritability for complex diseases and phenotypes (Yang *et al.*, 2010; Bulik-Sullivan *et al.*, 2015).

Polygenic scores (PGS) estimate individual propensity to a phenotype (eg. disease) by a summing individuals genotypes (coded 0, 1, 2), weighted by their effect sizes, estimated from GWAS data

$$\sum_{i \in S} \beta_i G_i$$

where β_i is the estimated effect for variant *i*, and G_i the genotype. Challenges relate to the nature of GWAS data, to determine the set of SNPs *S* to use. Linkage disequilibrium (LD) between variants, if ignored, would mean double-counting the effects of causal variants in high LD with multiple other variants. Additionally, error in the estimated effect sizes is most pronounced for small effects, for which true association may not be distinguishable from noise about a null value. Different approaches have been developed to deal with these. Initial approaches selected the most strongly associated variant in each genome wide-significant peak, but it was soon realised that predictive accuracy could be improved by reducing the significance threshold, to capture more truly associated variants even at the cost of including some false associations (which should add noise, but not bias, to any prediction), especially for highly polygenic diseases (Chatterjee *et al.*, 2016; Dudbridge, 2013). Methods were developed to use an external validation set to tune the significance parameter, as well as to use automated LD-pruning algorithms to select independent variants rather than selecting one variant per peak (Privé *et al.*, 2019; Euesden *et al.*, 2015).

More statistically sophisticated approaches have since been developed to average over multiple variants in LD without double counting (which is more accurate than selecting just one), and to shrink effect estimates by a continuous weight, w_i (Privé, Arbel, and Bjarni J Vilhjálmsson, 2020; Vilhjálmsson *et al.*, 2015; Ge *et al.*, 2019)

$$\sum_{i \in S} w_i \beta_i G_i$$

As with the “prune and threshold” approach, such methods initially required external independent validation datasets to tune their parameters, which can present a barrier to practical use. Recently, automated methods have been developed that tune their parameters on the same dataset used for training via hierarchical Bayesian models (Privé, Arbel, and Bjarni J Vilhjálmsson, 2020; Ge *et al.*, 2019) which perform nearly as well as their externally tuned counterparts. However, they require storage and inversion of large LD matrices, which is computationally intensive, and tuning, internally or externally, adds a further layer of iterative computation. Some approaches mitigate this by adding a thinning step, discarding a subset of SNPs to reduce the burden (eg. LDpred2 recommends restricting SNPs to those in the HapMap3 panel, Privé, Arbel, and Bjarni J Vilhjálmsson, 2020), but constructing PGS still takes a long time.

Our framework is based on considering PGS construction as a fine mapping problem. If we knew exactly the causal variants for a trait, an optimal PGS would comprise the estimated effects of just those variants. Fine mapping methods estimate probabilities that a variant is causal given observed data, and so in the absence of knowing the exact set of true causal

variants, a natural PGS might be constructed as above, with w_i set to the probability variant *i* is causal - ie by replacing the estimated effect size of each SNP by the posterior expectation of its causal effect. We present two approaches that can accommodate different assumptions on the number of causal variants in each LD-defined region. The basic fine mapping approach is very fast, because it makes a simplifying assumption that only a single causal variant may exist in any LD-defined region, based on the principles developed by Maller *et al.*, (2012). While unrealistic, it has been shown to perform well for fine-mapping, and of relevance to PGS, allows the definition of posterior probabilities of causality using predefined LD regions and without the burden of processing large LD matrices (Maller *et al.*, 2012). Distinguishing noise around null associations from true associations is dealt with by two parameters: a prior probability that a random SNP is causal, and the variance of the prior distribution of effect sizes at true causal variants. While in a PGS these would be open to tuning, in fine mapping sensible default values can be chosen that reflect existing knowledge from the breadth of GWAS studies already conducted (Wallace, 2020).

It is possible to relax the single causal variant assumption if additional information on LD is available. We propose using an alternative fine mapping method based on the “sum of single effects” model (SuSiE, Wang *et al.*, 2020). SuSiE is a computationally efficient approach to variable selection in linear regression, which uses a multiple causal variant model to compute the posterior inclusion probability for each variant in an LD-defined region, which we can use as w_i.

We present RápidoPGS, a lightweight and fast (*rápido*, in Spanish) method to compute PGS based on fine mapping approaches that only requires a summary statistics dataset, and that can generate weights for millions of SNPs in only a few seconds or minutes. RápidoPGS works fully in R, requiring few dependencies. A PGS can be quickly constructed from any GWAS summary statistic dataset, or any GWAS PubMed ID, if harmonised datasets are available at GWAS catalog.

2. Materials and Methods

2.1. Overview of our approach

We compute posterior probabilities of causality at each variant within LD-blocks defined by ldetect (Berisa and Pickrell, 2016) (publicly available at https://bitbucket.org/nygcresearch/ldetect-data/src).

Under a single causal variant assumption, given summary statistics at each SNP *i*, β_i, V_i (respectively, the estimated effect size and its variance), we can define an approximate Bayes factor summarising the evidence for its association (Wakefield, 2009) as

$$BF_i = \frac{l(\beta_i|\beta_i \sim N(0, W+V_i))}{l(\beta_i|\beta_i \sim N(0, V_i))}$$

where we assume that the true causal effect size at any SNP is sampled from an N(0,W) distribution.

Under the at most one causal variant assumption (often called a single causal variant assumption, but it includes the possibility of no causal variant), the posterior probability that SNP *i* is causal in a region of *n* SNPs, is

$$PP_i = \frac{\pi_i BF_i}{\pi_0 + \sum_{j=1}^{n} \pi_j BF_j}$$

where π_i is the prior SNP *i* is causal (typically set to the same value for all SNPs) and $\pi_0 = 1 - n\pi_i$ is the prior probability there is no causal variant (Maller *et al.*, 2012). Note therefore $0 \leq PP_i < 1$ and $\sum_i PP_i \leq 1$, so that PP_i can serve to shrink SNP effect estimates more or less according to the posterior belief that a SNP is causal for a trait.

We set default parameter values for RápidoPGS according to currently widely adopted values: $\pi_i = 10^{-4}$, $W = 0.2^2$ for case-control traits (Wallace, 2020). There is less consensus on appropriate values of W for quantitative traits, and we propose to estimate W given estimated heritability of the trait under study or a similar trait, which is often available (eg at LDhub, http://ldsc.broadinstitute.org/ldhub/, Zheng *et al.*, 2017):

$$W \simeq h^2 / (\nu \sum_{i=1}^{p} \frac{1}{N_i \sigma_i^2})$$

where, h^2 is the estimated heritability, ν is the prior that a given variant is causal (10^{-4}), *p* is the number of variants in the data, N_i is the number of individuals used in inference for variant *i*, and σ_i^2 is the variance of the estimated effect β_i associated with variant *i*. See Supplementary Note for full derivation of the equation.

Alternatively, we can relax the single causal variant assumption. SuSiE (Wang *et al.*, 2020) is an approach for variable selection in regression, based on previous models for Bayesian variable selection in regression (BVSR), but with a different structure that allows for a faster and simpler model fitting procedure based on fitting multiple 'single-effect regression' (SER) models (multiple-regression models with exactly one variable with non-zero regression coefficient), and then construct the overall effect vector as the sum of the single-effects vectors (hence, sum of single effects model). SuSiE requires the same parameters (π_i W) as the single causal variant approach, although W may also be internally estimated. We call the two approaches to estimating W "informed" or "auto".

Finally, we considered pruning input SNPs according to their p values using 2 parameters in order to increase speed: (1) α_{block}, the minimum p-value at least one variant in a LD block must have for the block to be considered for subsequent analyses (i.e. if no SNPs in a block have p-values below α_{block}, the entire block is skipped), and (2) α_{SNP}, the minimum p-value a variant within a selected block must have to be considered.

2.2. PGS generation using LDpred2-auto

We compared performance of both forms of RápidoPGS to that of a recent method that uses internal optimisation to avoid needing an external validation dataset: LDpred2-auto (Privé, Arbel, and Bjarni J Vilhjálmsson, 2020). We followed the LDpred2 instructions and adapted the code provided in the tutorials and the accompanying code (https://github.com/privefl/paper-ldpred2 and https://privefl.github.io/bigsnpr/articles/LDpred2.html). However, some changes were inevitable, to allow us to analyse a common set of data for all three methods here. We omitted the prediction steps, as our approach does not consider a validation dataset for PGS model generation. We use the 1000 Genomes data to estimate LD rather than the UKBB data, because we wanted to evaluate our method on a fully public dataset. Finally, the LDpred2 authors used the validation dataset to compute standard deviations of SNPs in that dataset (SD_{val}), and remove SNPs with too low (< 0.05, or smaller than SD_{ss} - 0.1) or two big (more than half SD_{ss}) , while we did not, as our tool is intended to be used without the need of validation dataset for PGS model creation.

We ran LDpred2-auto on a per-chromosome basis for all 22 autosomes and appended the results in each dataset. We computed initial h^2 from the data using the snp_ldsc function, and set 15 initial p values, ranging from 1e-4 to 0.9. To ensure convergence, we used 4,000 iterations (3,000 iterations + 1,000 burn-in). We skipped computation for chromosomes for which estimated h^2 was below 1e-4, as was done in the LDpred2 paper (Privé, Arbel, and Bjarni J Vilhjálmsson, 2020). Note that LDpred2 authors recommend running LDpred2 genome-wide rather than per-chromosome, due to better performance in their tests. For memory and speed reasons, we run LDpred2 per-chromosome, but provide LDpred2-auto-genomewide AUC results from LDpred2 paper, for comparison purposes.

2.3. GWAS datasets

We downloaded ten publicly available GWAS summary statistics datasets. We used the same eight case-control datasets that were used in the LDpred2 paper (Privé, Arbel, and Bjarni J. Vilhjálmsson, 2020) to ensure a fair comparison: Asthma, rheumatoid arthritis [RA], Type 1 diabetes [T1D], Type 2 diabetes [T2D], breast cancer [BRCA], prostate cancer [PRCA], coronary artery disease [CAD], and major depression disorder [MDD](Demenais *et al.*, 2018; Okada *et al.*, 2014; Cooper *et al.*, 2017; Scott *et al.*, 2017; Michailidou *et al.*, 2017; Schumacher *et al.*, 2018; Nikpay *et al.*, 2015; Wray *et al.*, 2018). We also added two quantitative traits: body mass index [BMI] and height (Locke *et al.*, 2015; Wood *et al.*, 2014).

We applied common quality control to all datasets prior to PGS generation, by removing unneeded rows with missing data at required columns (genomic coordinates, alleles, allele frequencies, log(OR), standard error, and case/control sample size if provided). Following LDpred2 quality control guidance for summary statistic datasets, we computed effective sample size (Neff = 4 / (1 / N controls + 1 / N cases)), and removed SNPs for which summary statistic standard deviation (SD_{ss}) $\sqrt{\frac{Neff}{4\,SE^2}}^{-1} < 0.1$.

For case-control traits, we set our informed prior for RápidoPGS as $W=0.2^2$. For quantitative traits, we estimated the informed W using estimates of heritability from LDhub: "BMI" h^2 = 0.246297646, and "Standing Height" h^2 = 0.4622973 which led to estimates of W of $(8.246 \times 10^{-2})^2$ for BMI, and $(9.135 \times 10^{-2})^2$ for height. We considered thinning the input SNPs for RápidoPGS-multi using $\alpha_{SNP} = 0.1$, or $\alpha_{SNP} = 0.01$. As only 10% of truly null SNPs achieve p values < 0.1 and only 1% < 0.01, this should remove 90% or 99% of null SNPs respectively. Assuming most blocks have 1000 SNPs or more, nearly all blocks containing only null SNPs should achieve a minimum p value of 10^{-3}, and only 10% a minimum p value of 10^{-4}, so we chose block-thinning parameters of $\alpha_{block} = 10^{-3}$ or $\alpha_{block} = 10^{-4}$.

Both LDpred2 and RápidoPGS-multi methods require LD matrices, which can be computed using a reference panel including individuals from the same (or similar) population as the GWAS dataset. We used the 1000 Genomes Phase III as a reference panel, from which we removed SNPs with MAF < 0.01 or with HWE p-value < 1^{-10}. We chose to use only 503 individuals from European ancestry to match the populations on which our GWAS datasets were generated. We used the variants in the resulting panel to filter all datasets to ensure that RápidoPGS-single, which by default uses all available SNPs, could be compared to RápidoPGS-multi. LDpred2 tutorial recommends further filtering by HapMap3 variants (available at https://github.com/privefl/bigsnpr/raw/master/data-raw/hm3_variants.rds), which we performed at LDpred2 runtime. Note that the authors now recommend using a bigger panel for better results, and they provide UK Biobank-based LD matrices, but the published results were very similar for the two approaches.

2.4. Model evaluation using UK Biobank individual data

We evaluated the predictive performance of our PGS models generated with RápidoPGS and LDpred2 using individual genotypes from the UK Biobank cohort. For all traits we excluded SNPs with low imputation quality (info score <0.3) or multi-allelic SNPs. Moreover, we removed related individuals and restricted the analysis to individuals of European ancestry (UKBB field 22006, genetic ethnic group). The binary traits selected (Table S1) were the same used in the evaluation of LDpred2 (Privé, Arbel, and Bjarni J. Vilhjálmsson, 2020) and we applied the same selection criteria for cases and controls as previously described (Privé *et al.*, 2019) following the code for each of the traits as in https://github.com/privefl/simus-PRS/tree/master/paper3-SCT/code_real). For all disease traits we included as cases those individuals who self-reported the condition or were diagnosed by a medical doctor or the condition was included in their death record. For breast and prostate cancer we excluded individuals with other cancer diagnosis. Moreover, for breast cancer we restricted the analysis to females, for prostate cancer to males. For rheumatoid arthritis we excluded individuals with any other musculoskeletal system and connective tissue condition. For type 1 diabetes we excluded individuals with type 2 diabetes and vice versa. For coronary artery disease we excluded individuals with other heart conditions. For asthma we excluded individuals with additional respiratory conditions. For major depression disorders we excluded individuals with additional mental and behavioral disorders as well as individuals which were included in the GWAS used to construct the score. For BMI and height we used the UK Biobank codes "21001" and "50", respectively.

After computing the scores, for case control phenotypes we calculated the area under the curve (AUC) using the R package “pROC” whereas for continuous traits we assessed the squared correlation between the PGS and the measured trait (r^2).

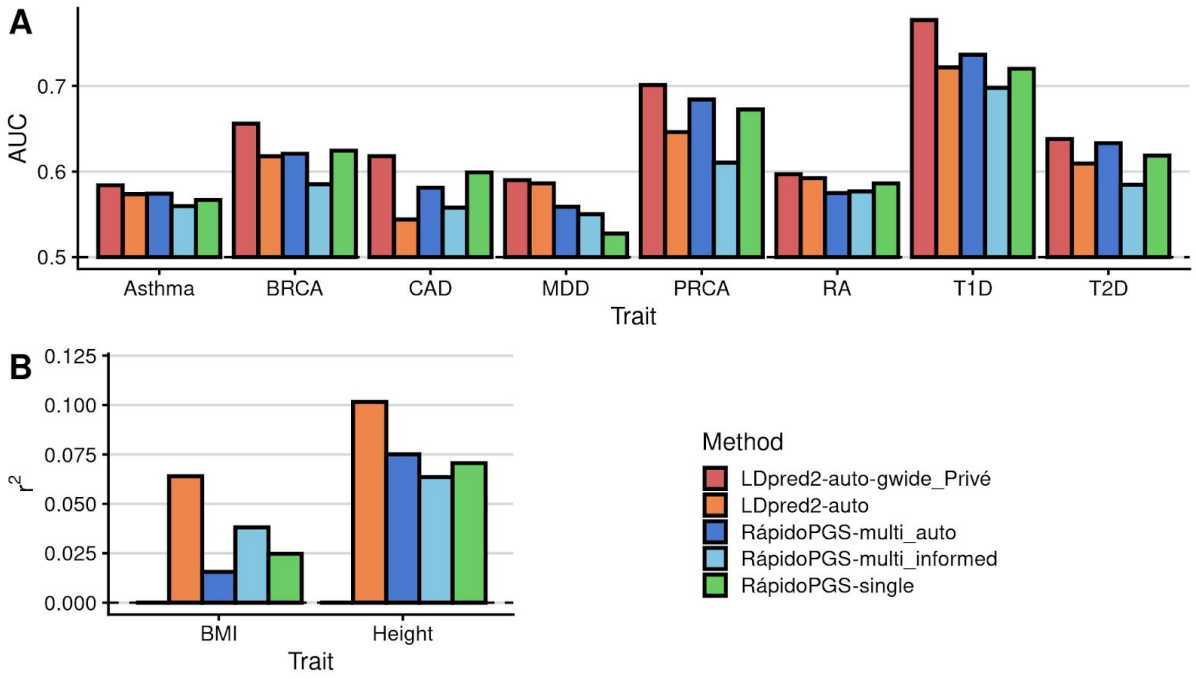
2.5. Benchmarking

We used “microbenchmark” R package to time full PGS model generation in the asthma dataset between RápidoPGS (single, multi [$\alpha_{block} = 10^{-3}$, $\alpha_{SNP} = 0.1$] and multi [$\alpha_{block} = 10^{-4}$, $\alpha_{SNP} = 0.01$]) and LDpred2 with 3,000 iterations. Since LDpred2 and RápidoPGS-multi allow parallelisation, we used 15 nodes for benchmarking at our computing cluster. LDpred2-auto benchmark showed very variable finishing times, so we ran the benchmark 10 independent times, and showed the mean and standard deviation. In addition, for comparison of time and performance of RápidoPGS-single and RápidoPGS-multi with the above parameters, we benchmarked them for all 10 datasets using 15 cores (Figure S1).

3. Results

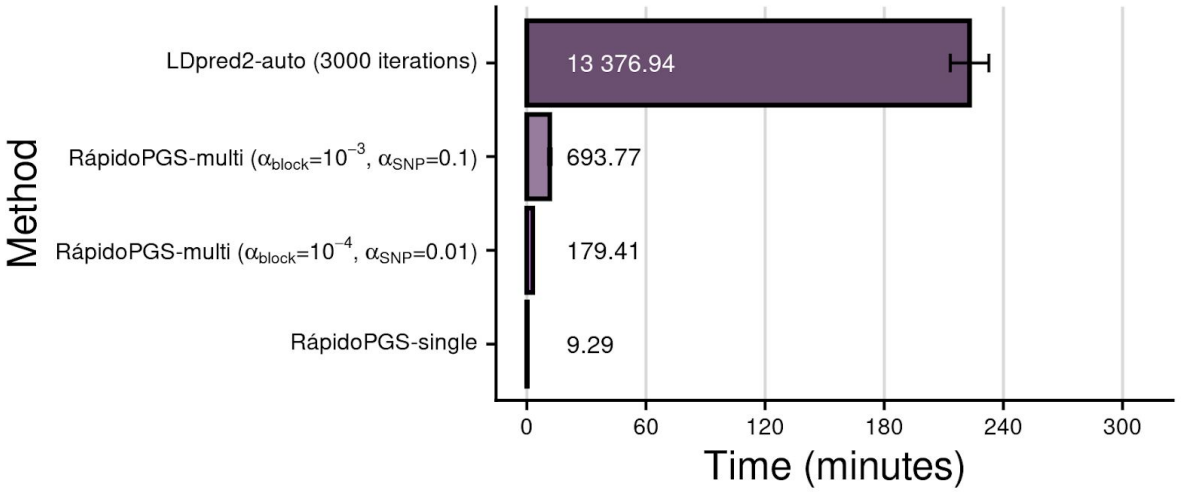
For case-control datasets (Fig. 1A, Supplementary Table 2), AUC were slightly lower in our runs of the LDpred2-auto algorithm compared to those published, presumably due to our difference choices outlined in Methods, but the general pattern, with T1D performing best, was generally preserved. For RápidoPGS-multi, several parameter choices are allowed, to thin SNPs according to significance, and to use informed or auto-estimates of W. We first considered the milder thinning (α_{SNP}=0.1, α_{block}=10^{-3}). We found that for RápidoPGS-multi auto-estimation of W did generally increase performance compared to informed priors. However, RápidoPGS-multi with auto-W outperformed RápidoPGS-single in only 5 cases out of 8 even though RápidoPGS-single can only use informed W.Comparing to our runs of LDpred-2, we found performance was similar overall (median AUC difference [RápidoPGS-single - LDpred2-auto] = 0.0025, range = [-0.0738, 0.0621]; RápidoPGS-multi-auto - LDpred2-auto] = 0.009, range = [-0.0273, 0.0380]; RápidoPGS-multi-informed - LDpred2-auto] = -0.024, range = [-0.036, 0.0138]). For both quantitative traits (Fig. 1B), LDpred2-auto performed slightly better than all options for RápidoPGS (r^2 difference for height, BMI [RápidoPGS-single - LDpred2-auto] = -0.0310, -0.0393; [RápidoPGS-multi-informed - LDpred2-auto] = -0.0381, -0.0259). Whilst using the auto option in RápidoPGS-multi for BMI resulted in reduced r^2 for BMI compared to the informed prior setting (-0.0225), it translated into slightly higher r^2 (+0.0116) for height.

Figure 1. RápidoPGS has comparable performance to LDpred2-auto. A. AUC for case-control traits using LDpred2-auto and RápidoPGS (single and multi with $\alpha_{block} = 10^{-3}$, $\alpha_{SNP} = 0.1$ parameters). For comparison purposes, we added AUC for LDpred2-auto genomewide, reported by (Privé, Arbel, and Bjarni J Vilhjálmsson, 2020). **B.** r^2 results for BMI and height. No LDpred2 paper results were available for these traits.



We benchmarked the speed of RápidoPGS and LDpred2-auto, using 10 independent runs on the asthma GWAS dataset (Demenais *et al.*, 2018), using automatic variance prior selection for RápidoPGS-multi, 3,000 iterations for LDpred2-auto, and 15 cores for all models (Figure 2). RápidoPGS-single does not require parallelisation nor high computational power and can be run on a regular laptop, although for comparison we ran it on the same high performance computing system as LDpred2. On the other hand, RápidoPGS-multi can benefit from parallelisation. We found that RápidoPGS-single took seconds to finish, while RápidoPGS-multi took minutes and LDpred2-auto hours (see Fig. 2 for detailed comparison). For RápidoPGS-multi, changing the SNP thinning parameters α_{block} and α_{SNP} from (10^{-3}, 0.1) to (10^{-4},0.01) and retain more on average 5-10 fold fewer SNPs decreased computation time (~4x faster on average) without any considerable decrease in predictive performance (Supp Fig. 1). On the other hand, internal estimation of W, shown to increase performance above, had only a small effect on run times (at most 1.5x longer).

Figure 2. Timings of RápidoPGS (single and multi approaches, the latter with two combinations of parameters) and LDpred2-auto (3,000 iterations + burn-in) calculating a PGS model for Asthma, in minutes. Numbers by the bars represent the time in seconds. All computations were run using 15 cores, although RápidoPGS-single does not benefit from parallel computation.



4. Discussion

Having PGS scores computed easily and quickly can be advantageous in a context in which there is a need for rapid assessment of many traits. The main downside of most sophisticated PGS methods is their computational cost and running time, taking many hours to finish even when using multiple cores in a high performance computing context. This is due to their need to optimise parameters either internally or by external validation. While fine mapping methods do require setting parameters (typically describing the prior frequency and distribution of true causal effect sizes), sensible values for these have been learnt from analysis of the many GWAS datasets to date, or can be back-calculated from public estimates of trait heritability as we suggest for quantitative traits. This avoids the need to tune for individual datasets, which improves speed, and which may be an insurmountable barrier for rare traits where a pair of training and test GWAS do not exist. By using fine mapping principles and few assumptions, RápidoPGS can generate reasonably accurate PGS models in a fraction of the time, going from only a few seconds to about an hour, depending on the approach and dataset size.

Our method for improving the speed of the SuSiE, by filtering SNPs according to p value, is approximate and was not recommended by the method authors. We use it to allow this fine mapping approach to run in a reasonable time for our purposes, and it appears to perform well in this situation, but it is unlikely to be optimal and it is probable a more accurate PGS could be constructed if all SNPs were supplied, but this would be at the cost of more than an order of magnitude slower speed.

Like LDpred2, RápidoPGS-multi requires LD matrices, constructed on the same or similar population to the dataset on which the PGS is trained. However, since individual data is often not available due to privacy concerns, this can be done using a publicly available reference panel. Despite its relatively small size (2504 individuals of worldwide origin), 1000 Genomes Project Phase III is publicly available, and as we have shown, can be used for LD matrix computation and obtain good results. We provide a function to download and pre-process a 1000 Genomes-based reference panel from scratch, although users are free to use their own panel.

On the basis of datasets considered here, we recommend using RápidoPGS-multi if suitable LD information is available, because although the number of cases when one outperformed the other was very similar in the datasets considered here (6 to 4), when it outperformed RápidoPGS-single it did so by a greater margin than it underperformed in the other cases. For RápidoPGS-multi, we recommend allowing internal estimation of W for binary traits, and choosing W according to estimates of h^2 where possible for quantitative traits. However, the performance of fine-mapping methods can be substantially reduced when LD information from different ancestry to the study sample is used. We therefore recommend that when suitable LD information is not available, which is a particular concern in broad-ancestry studies or meta analyses, or when speed is particularly important, that RápidoPGS-single is chosen.

We are not the first to suggest that fine mapping approaches can be helpful for PGS construction. Newcombe and colleagues (2019) used reversible jump MCMC to fit a fine mapping model to GWAS summary statistics, parallelising across LD blocks as we do here. However, rather than estimating posterior inclusion probabilities and using these to shrink frequentist effect estimates, they averaged samples from the posterior distribution of causal effect estimates in their fine mapping model. Our work shows that fine mapping approaches can also be adapted to the standard PGS framework, offering an alternative method to generate the weights w_i.

We conclude that RápidoPGS is orders of magnitude times faster than a competitive automated PGS method with only a small loss in accuracy, and is thus suited for rapid screening of many potentially related traits, before maximising the accuracy of scores for specific traits of interest using a method where parameters can be tuned for optimality. In addition, RápidoPGS is meant to be easy to use. We wrote it entirely in R, with few dependencies, and we supply helper functions to download summary statistics straight from GWAS catalog, to build the required reference panel for RápidoPGS-multi from scratch, and to compute variance priors using external heritability estimates. We also provide step-by-step tutorials in the dedicated https://grealesm.github.io/RapidoPGS/.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

We thank Florian Privé and Bjarni Vilhjálmsson for useful guidance on running LDpred2. We thank Chris Q. Eijsbouts for valuable feedback on early versions of the RápidoPGS package.

Funding information

This work was supported by the MRC (MC_UU_00002/4). This research was funded in whole, or in part, by the Wellcome Trust [WT107881, 203950/Z/16/A]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

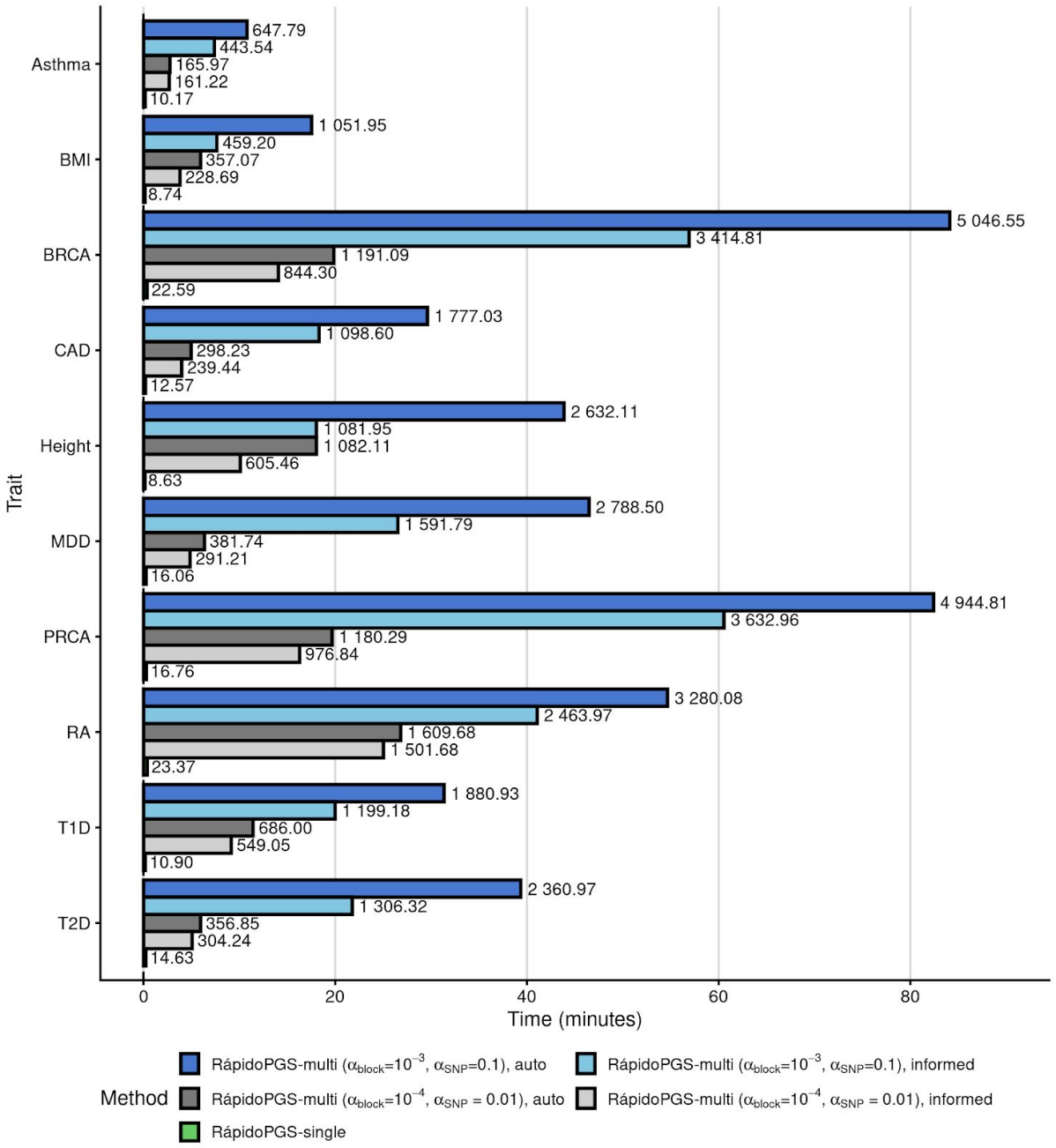
This research has been conducted using the UK Biobank Resource under Application Number 30931.

Web resources
R - https://cran.r-project.org/
RápidoPGS - https://github.com/GRealesM/RapidoPGS
LDpred2 - https://privefl.github.io/bigsnpr/articles/LDpred2.html
LD detect datasets - https://bitbucket.org/nygcresearch/ldetect-data/src
GWAS catalog - https://www.ebi.ac.uk/gwas/
1000 Genomes Project - https://www.internationalgenome.org/
LDHub - http://ldsc.broadinstitute.org/ldhub/

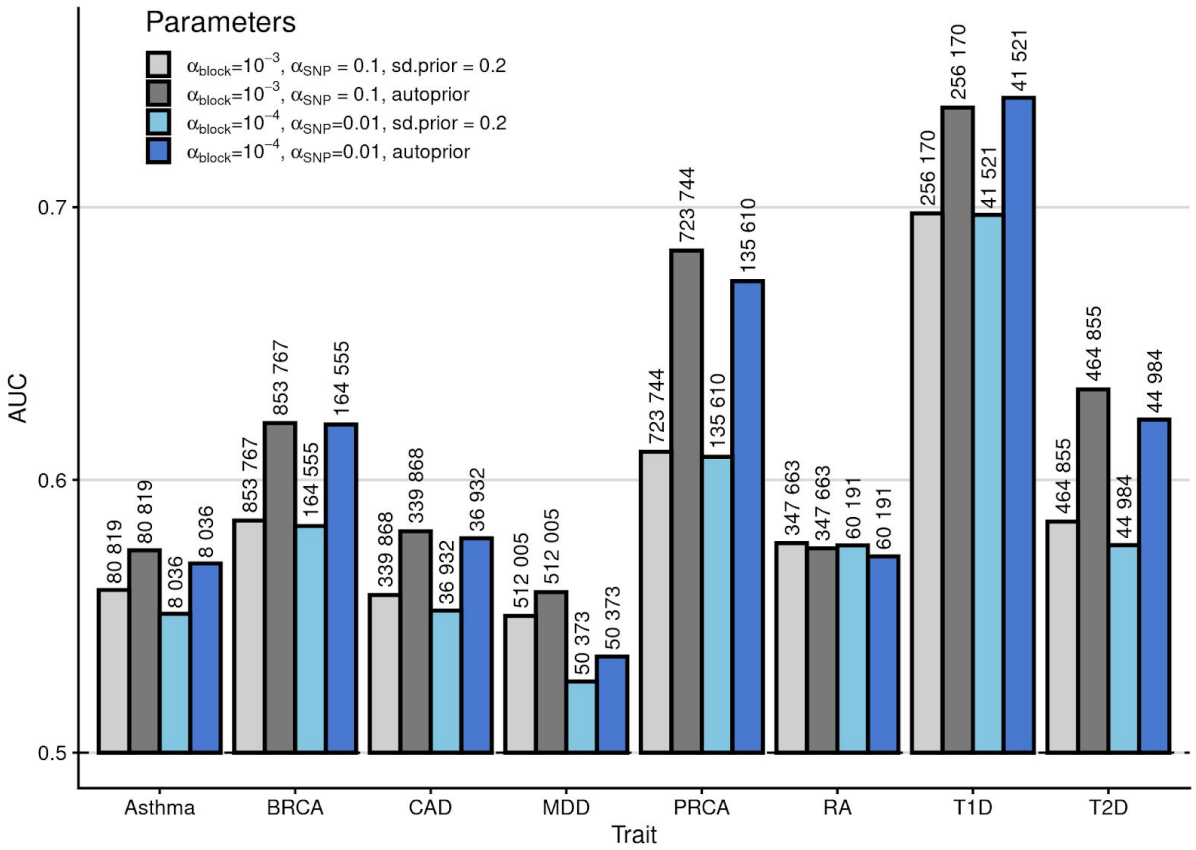
Data and code availability
Summary statistics datasets used are publicly available in their respective publications (see Suppl. Table 1). Code used in the analyses is available at GitHub (https://github.com/GRealesM/RapidoPGS_paper) and code used for model evaluation is available at GitLab (https://gitlab.com/evigorito/applyrapidopgs).
We have extensively used tools in the *bigsnpr* (Privé *et al.*, 2018) and *data.table* for large dataset handling and analysis.

Description of Supplemental Data
Supplemental data contains Supplementary Materials Methods, 2 figures, and 2 tables.

Supplementary Figure 1. RápidoPGS benchmark for all selected traits, measured in minutes and seconds (right of the bars). RápidoPGS-single is the fastest approach, followed by RápidoPGS-multi with $\alpha_{block} = 10^{-4}$, $\alpha_{SNP} = 0.01$ parameters. RápidoPGS-multi approaches differ in the number of SNPs being considered, with $\alpha_{block} = 10^{-3}$, $\alpha_{SNP} = 0.1$ throwing away less information, but being slower overall. For all traits, the automatic prior variance option takes longer. For quantitative traits (BMI and height) we used custom priors, and only those runs are shown here. All computations were performed using 15 cores.



Supplementary Figure 2. Performance comparison between RápidoPGS-multi parameters for case-control traits, measured in AUC. We tested two pairs of parameters for the RápidoPGS-multi approach, to find a balance between performance and computational speed. Both sets of parameters differ in their threshold to exclude LD blocks and SNPs by p-value (see Methods). SNPs included in each model at evaluation time are displayed vertically on top of the bars. For most traits, $\alpha_{block} = 10^{-3}$, $\alpha_{SNP} = 0.1$ (in blue), which includes ~5-10 fold more SNPs, performs better than when using $\alpha_{block} = 10^{-4}$, $\alpha_{SNP} = 0.01$, although these differences are not very big for most traits. We also compared model performance automatic and informed (0.2^2) prior variance computation, being the former slower but in general better.



Supplementary Table legends

Table S1. GWAS summary statistic datasets. Training sample sizes correspond to the dataset on which PGS were built, test sample sizes correspond to the UKBB cohort on which PGS were evaluated. For T1D dataset, we used Cooper et al., 2017, instead of Censin et al., 2017 dataset. These should be equivalent so we provide a link to the available Censin et al. dataset.

Table S2. Results of RápidoPGS and LDpred2-auto model evaluations, in AUC and r^2.

References

Berisa,T. and Pickrell,J.K. (2016) Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**, 283–285.

Bulik-Sullivan,B.K. *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

Chatterjee,N. *et al.* (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.*, **17**, 392–406.

Cooper,N.J. *et al.* (2017) Type 1 diabetes genome-wide association analysis with imputation identifies five new risk regions. *bioRxiv*, 120022.

Demenais,F. *et al.* (2018) Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.*, **50**, 42–53.
Dudbridge,F. (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, **9**, e1003348.
Euesden,J. *et al.* (2015) PRSice: Polygenic Risk Score software. *Bioinformatics*, **31**, 1466–1468.
Ge,T. *et al.* (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.*, **10**, 1776.
International Schizophrenia Consortium *et al.* (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.
Locke,A.E. *et al.* (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197–206.
Maller,J.B. *et al.* (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.*, **44**, 1294–1301.
Michailidou,K. *et al.* (2017) Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**, 92–94.
Newcombe,P.J. *et al.* (2019) A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genet. Epidemiol.*, **43**, 730–741.
Nikpay,M. *et al.* (2015) A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.*, **47**, 1121–1130.
Okada,Y. *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**, 376–381.
Privé,F. *et al.* (2018) Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinforma. Oxf. Engl.*, **34**, 2781–2787.
Privé,F., Arbel,J., and Vilhjálmsson,Bjarni J (2020) LDpred2: better, faster, stronger. *Bioinformatics*, btaa1029.
Privé,F. *et al.* (2019) Making the Most of Clumping and Thresholding for Polygenic Scores. *Am. J. Hum. Genet.*, **105**, 1213–1221.
Schumacher,F.R. *et al.* (2018) Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.*, **50**, 928–936.
Scott,R.A. *et al.* (2017) An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes*, **66**, 2888–2902.
Vilhjálmsson,B.J. *et al.* (2015) Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.*, **97**, 576–592.
Wakefield,J. (2009) Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.*, **33**, 79–86.
Wallace,C. (2020) Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genet.*, **16**, e1008720.
Wang,G. *et al.* (2020) A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.*
Wood,A.R. *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, **46**, 1173–1186.
Wray,N.R. *et al.* (2018) Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.*, **50**, 668–681.
Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.
Zheng,J. *et al.* (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, **33**, 272–279.