Bioinformatics, 38(4), 2022, 908–917
https://doi.org/10.1093/bioinformatics/btab803
Advance Access Publication Date: 2 December 2021
Original Paper

Genome analysis

Uncovering complementary sets of variants for predicting quantitative phenotypes

Serhan Yilmaz[1,*], Mohamad Fakhouri[2], Mehmet Koyutürk[1,3], A. Ercüment Çiçek[2,4,*] and Oznur Tastan [5,*]

[1]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH 44106, USA, [2]Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey, [3]Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, OH 44106, USA, [4]Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA and [5]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey

*To whom correspondence should be addressed.

Associate Editor: Tobias Marschall

Received on April 11, 2021; revised on September 21, 2021; editorial decision on November 11, 2021; accepted on November 24, 2021

Abstract

Motivation: Genome-wide association studies show that variants in individual genomic loci alone are not sufficient to explain the heritability of complex, quantitative phenotypes. Many computational methods have been developed to address this issue by considering subsets of loci that can collectively predict the phenotype. This problem can be considered a challenging instance of feature selection in which the number of dimensions (loci that are screened) is much larger than the number of samples. While currently available methods can achieve decent phenotype prediction performance, they either do not scale to large datasets or have parameters that require extensive tuning.

Results: We propose a fast and simple algorithm, Macarons, to select a small, complementary subset of variants by avoiding redundant pairs that are likely to be in linkage disequilibrium. Our method features two interpretable parameters that control the time/performance trade-off without requiring parameter tuning. In our computational experiments, we show that Macarons consistently achieves similar or better prediction performance than state-of-the-art selection methods while having a simpler premise and being at least two orders of magnitude faster. Overall, Macarons can seamlessly scale to the human genome with $\sim 10^7$ variants in a matter of minutes while taking the dependencies between the variants into account.

Availabilityand implementation: Macarons is available in Matlab and Python at https://github.com/serhan-yilmaz/macarons.

Contact: serhan.yilmaz@case.edu or cicek@cs.bilkent.edu.tr or otastan@sabanciuniv.edu

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

1 Introduction

Genome-wide association studies (GWAS) attempt to find a relation between the genetic variations and a phenotype. Many single-nucleotide polymorphisms (SNPs) have been found to be associated with various diseases or disorders including Type II diabetes, obesity and schizophrenia as well as other quantitative traits like height individually (Goldstein, 2009; Visscher *et al.*, 2017). However, individual SNPs fail to explain complex phenotypes, in which multiple SNPs contribute collectively (Manolio *et al.*, 2009). Thus, as an alternative and more powerful approach, many studies aim at finding a good subset of SNPs that are associated with the phenotype of interest as a group (Cordell, 2009; Phillips, 2008; Wang *et al.*, 2010; Wei *et al.*, 2014). This study is mainly focused on the problem of finding a subset of SNPs that are *collectively* predictive of the phenotype of interest. For the sake of brevity, we will simply refer to it as the SNP subset selection problem throughout this article.

1.1 Approaches investigating combinations of SNPs

Finding combinations of SNPs that are predictive of a phenotype is computationally challenging due to the large number of possible combinations that need to be considered. There are methods that focus on high-order interactions using exhaustive search (Lou *et al.*, 2007; Nelson *et al.*, 2001) or greedy algorithms (Evans *et al.*, 2006; Yosef *et al.*, 2007) on a small, limited pool of SNPs that is usually not more than a few hundreds (Fang *et al.*, 2012; Ritchie *et al.*, 2001). While such pools of 'promising SNP candidates' are typically obtained using *a priori* information sources by limiting the analysis to SNPs residing in the coding regions of the genome, it is also possible to conduct a filtering based on automated searches (Ding *et al.*, 2015; Van Hulse *et al.*, 2012).

Indeed, for combinatorial studies investigating the pair-wise interactions (or as more commonly known as *epistasis*) between the variants in full genome, recent studies show the importance of limiting the search space through prioritization of the tests (Cowman and Koyutürk, 2017; Piriyapongsa *et al.*, 2012), both to alleviate the computational intensity of the task, as well as to improve the overall statistical power. Specifically, Caylak *et al.* (2020) demonstrates the utility of an initial filtering of SNPs based on an automated SNP selection algorithm (Yilmaz *et al.*, 2019) as a powerful approach that can improve the statistical power considerably.

1.2 Approaches for SNP selection problem in quantitative phenotypes

The SNP selection problem in quantitative phenotypes essentially corresponds to a feature subset selection problem for multivariate regression (Miller, 2002). However, due to the high-dimensional nature of typical GWAS data (millions of variants), established methods for feature selection such as linear regression with l_1 (lasso) regularization (Grave *et al.*, 2011; Tibshirani, 1996), spectral-relaxation-based approaches (Das *et al.*, 2012; Zhao and Liu, 2007), graph-constrained feature selection methods like GraphLasso and GroupLasso (Jacob *et al.*, 2009; Meier *et al.*, 2008), as well as various other methods with sparsity constraints known in the bioinformatics community in the context of other problems (e.g. for selecting gene sets) (Jia *et al.*, 2011; Li and Li, 2008; Liu *et al.*, 2017), are computationally too expensive for this task. Thus, a common strategy is to apply a simple threshold-based filtering (e.g. a *P*-value cutoff) based on individual phenotype associations (Van Hulse *et al.*, 2012), for example, using a statistical test like sequence kernel association test (SKAT) (Wu *et al.*, 2011). The downside of this approach is that threshold-based filtering considers each variant independently and does not take into account of the dependencies or interactions between them.

To achieve a scalable solution for all known variants in the genome while considering the dependencies between them, alternative SNP selection algorithms have been proposed (Azencott *et al.*, 2013; Yilmaz *et al.*, 2019). Such algorithms simplify the problem by focusing on a linear combination of individual phenotype associations of SNPs while using some *a priori* information encoded in the form of a biological network to improve the overall predictivity of the selected subset. In particular, SConES (Azencott *et al.*, 2013) uses a minimum-cut solution under sparsity and connectivity constraints on a SNP–SNP network. More recently, SPADIS (Yilmaz *et al.*, 2019) selects a diverse set of SNPs using the SNP–SNP network.

1.3 The drawbacks of existing methods

Linkage disequilibrium (LD) which refers to the non-random association of variants is a common phenomenon for close variants on the same chromosome (Ardlie *et al.*, 2002). While the connectivity constraint of SConES helps to improve the quality of the selected set, it implicitly promotes the selection of SNPs that are in LD impairing the prediction performance. On the other hand, SPADIS seeks to increase the diversity of SNPs by penalizing the selection of close SNPs on the input network. While this diversity helps to avoid redundant SNPs in LD and improves the phenotype predictions, the drawback of SPADIS is that it requires two parameters without any interpretable meanings or default values, that need to be tuned through an external procedure such as cross-validation. The need for such external procedures not only makes the method hard to apply from a user viewpoint, but also considerably exacerbates the run time and reduces the robustness of the selections when there are time and resource constraints.

1.4 Macarons: a fast and simple algorithm to select complementary SNPs

To overcome these limitations, we determined three main objectives a SNP selection algorithm should satisfy: (i) have good prediction performance for quantitative phenotypes (at least as predictive as available methods); (ii) fast enough to consider all variants in the genome; and (iii) easy to use without requiring external parameter tuning procedures like cross-validation. Thus, we propose a new algorithm named Macarons that take into account the correlations between SNPs to avoid the selection of redundant pairs of SNPs in LD. Overall, Macarons features two simple, interpretable parameters to control the time/performance trade-off: the number of SNPs to be selected (k), and maximum intra-chromosomal distance (D, in base pairs) to reduce the search space for redundant SNPs. Note that, since the parameters have interpretable meanings, they can be determined in advance (without requiring an external procedure for parameter tuning) with the available computational resources and the goals of further studies in mind.

2 Materials and methods

2.1 Background

2.1.1 Problem definition

We are given as input a ground set of SNPs V of cardinality n, genotype matrix $\mathbf{X} \in \{0, 1, 2\}^{m\times n}$ decoding the number of alternate alleles for m samples and n SNPs, and a phenotype vector $Y \in \mathbb{R}^{m\times 1}$ containing quantitative values for m samples. The number of SNPs n is much larger than the number of samples $n \gg m$. Thus, we would like a obtain a small subset of SNPs $S = \{s_1, s_2, \ldots, s_k\} \subseteq V$ of size k that maximizes the prediction performance of the given phenotype vector Y based on a regression model $\mathcal{M}$. In this study, we consider a linear model (i.e. without interaction terms modeling epistasis), where each selected SNP $s_i \in S$ has an additive effect on the phenotype:

$$Y \sim \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \cdots + \beta_k s_k + \epsilon \quad (1)$$

where β_i is the regression coefficients to be learned from data and ϵ_i is an error term that is normally distributed with zero mean. Based on this model, the collective effect of the SNP set S can be characterized by the squared multiple correlation coefficient $R^2(Y, S)$ which has the interpretation of the variance explained in Y by S. Thus, the overall SNP selection problem can be defined as a SNP subset search problem that maximize the following function:

$$\max_S R^2(Y, S) \text{ subject to } |S| = k \quad (2)$$

2.1.2 Forward step-wise regression

Generally, solving the regression problem given in Equation (2) is NP-hard (Natarajan, 1995). However, due to near submodularity of R^2, greedy formulations that iteratively grow a set based on a local gain function G (as in Algorithm 1), produce near-optimal results, proving a good approximation for maximizing R^2 under a cardinality constraint (Das and Kempe, 2011; Das *et al.*, 2012). Among such algorithms, a notable one that is commonly used is the forward stepwise regression that maximizes semi-partial squared correlation as its gain function:

$$G(S_t, s_x) = R^2(Y, (s_x|S_t)) \quad (3)$$

where S_t is the subset of selected features at the t iteration of the algorithm, s_x is a candidate feature being considered and $R^2(Y, (s_x|S_t))$ is the semi-partial correlation coefficient between Y and $s_x|S_t$, when s_x is regressed and residualized with every variable in S_t.

The main issue with using this approach for the SNP selection problem is that it requires estimating and inverting the covariance matrix. This requirement not only makes the algorithm computationally intensive with $O(n^3)$ runtime complexity, but also leads to

```
Algorithm 1 Greedy Subset Selection AlgorithM
Input: Gain function $G$, ground set $V$, cardinality constraint $k \leq |V|$.
Output: Set $S \subseteq V$ such that $|S| = k$.
    $S \leftarrow \varnothing$
    while $|S| < k$ do
        $S \leftarrow S \cup \text{argmax}_{s_x \in V \setminus S}\, G(S, s_x)$
    end while
```

the selection of SNP sets that is likely to overfit to the given training data (this is due to the high-dimensionality nature of the problem where $n \gg m$).

2.2 Macarons

Here, we follow an approach similar to the forward step-wise regression where we iteratively grow the selected SNP set based on their estimated contribution G for phenotype prediction as measured by the semi-partial correlation. However, to scale to all SNPs in a typical GWAS study as well as to improve the robustness of the algorithm, we apply some simplifying assumptions that reduce the computational complexity and error in estimation. First, we start by expressing the semi-partial correlation $R^2(Y, (s_x|S_t))$ in an alternate form:

$$R^2(Y, (s_x|S_t)) = R^2(s_x, S_t \cup Y) - R^2(s_x, S_t) \tag{4}$$

where $R^2(s_x, S_t \cup Y)$ and $R^2(s_x, S_t)$ are multiple correlation terms corresponding to linear models predicting s_x using the SNP set S_t with and without the phenotype variable Y, respectively. Here, we can further decompose $R^2(s_x, S_t \cup Y)$ into two parts:

$$R^2(s_x, S_t \cup Y) = 1 - \Big(1 - r^2(s_x, Y)\Big)\Big(1 - R^2(s_x, S_t|Y)\Big) \tag{5}$$

where $r^2(s_x, Y)$ is the squared Pearson's correlation coefficient indicating the individual predictivity of s_x on Y, and $R^2(s_x, S_t|Y)$ is the partial correlation between s_x and S_t given Y. Here, we assume that the portion of variance that overlap between s_x and S_t does not depend on their overlap with Y, which can be expressed as follows:

$$R^2(s_x, S_t|Y) \approx R^2(s_x, S_t|\varnothing) = R^2(s_x, S_t). \tag{6}$$

With this assumption, the gain function $R^2(Y, (s_x|S_t))$ can be simplified as follows:

$$\begin{aligned} R^2(Y, (s_x|S_t)) &\approx 1 - \Big(1 - r^2(s_x, Y)\Big)\Big(1 - R^2(s_x, S_t)\Big) - R^2(s_x, S_t) \\ &= r^2(s_x, Y)\Big(1 - R^2(s_x, S_t)\Big) \end{aligned} \tag{7}$$

Here, since $r^2(s_x, Y)$ quantifies the individual predictivity of the candidate SNP s_x on phenotype Y, which we can also replace with other phenotype association scores (denoted c_x for SNP s_x) such as SKAT. Thus, a more general gain function can be defined as follows:

$$G(S_t, s_x) = c_x\Big(1 - R^2(s_x, S_t)\Big) \tag{8}$$

Overall, the multiple correlation $R^2(s_x, S_t)$ measures the collective redundancy between s_x and S_t, and here used as a penalty function to facilitate the selection of complementary SNPs for the phenotype prediction.

2.2.1 Estimating the penalization function

The main challenge in estimating the multiple correlation is that it requires the computation of high-order interaction terms among the selected SNPs. This makes its estimation for a given data sample both computationally intensive [with $O(mt^2 + t^3)$ runtime complexity], and noisy. To help overcome these issues, we first express the multiple correlation as multiplication of several terms involving partial correlations:

$$\begin{aligned} R^2(S_t, s_x) &= 1 - \Big(1 - R^2(s_1, s_x|\varnothing)\Big) \\ &\quad \Big(1 - R^2(s_2, s_x|s_1)\Big) \ldots \Big(1 - R^2(s_t, s_x|s_1, \ldots, s_{t-1})\Big) \\ &= 1 - \prod_{i=1}^{t}\Big(1 - R^2(s_i, s_x|s_1, \ldots, s_{i-1})\Big) \\ &= 1 - \prod_{i=1}^{t}\Big(1 - R^2(s_i, s_x|S_{i-1})\Big) \end{aligned} \tag{9}$$

where $R^2(s_i, s_j|S')$ denotes the squared partial correlation between SNPs s_i and s_j given the SNPs within the set S'. Note that, these partial correlation calculations also require computing high-order interactions; thus do not simplify the computation of the multiple correlation by themselves. For this purpose, we make the following simplifying assumption:

$$R^2(s_i, s_x|S_{i-1}) \approx R^2(s_i, s_x|\varnothing) = r^2(s_i, s_x) \tag{10}$$

where $r^2(s_i, s_x)$ is the squared zero-order correlation coefficient (i.e. ordinary Pearson's correlation) between SNPs s_i and s_x. Thus, with this assumption, the estimation of the multiple correlation simplifies to:

$$\overline{R}^2(S_t, s_x) = 1 - \prod_{s_i \in S_t}\Big(1 - r^2(s_i, s_x)\Big) \tag{11}$$

This assumption helps with the overfitting problem in the estimation of multiple correlation since it reduces the number of parameters needed to be estimated from the data and reduces the required computation time drastically [from $O(mt^2 + t^3)$ to $O(mt)$].

In the remaining sections of this manuscript, we will refer to the estimation of the squared multiple correlation $\overline{R}^2(S_t, s_x)$ simply as the penalization function, and we will refer to the zero-order correlation $r^2(s_i, s_x)$ as the redundancy function.

Similar to the squared multiple correlation $R^2(S_t, s_x)$, this simplified penalization function has several useful properties such as being bounded in [0,1] region, being monotonic and applying diminishing returns principle (where the increase in penalization decreases proportionally on subsequent iterations as the selected set grows). We explain these properties in more detail in Supplementary Text S1.

2.2.2 Limiting the search space through intra-chromosomal distance

One particular issue for directly using the penalization function given in Equation (11) together with the gain function and algorithm in Equation (8) and Algorithm 1 is that the overall runtime can still be slow for large k (number of SNPs selected) with algorithmic complexity of O(nmk) due to the requirement of computing O(nk) correlation coefficients. For this reason, we make an additional simplifying assumption to limit the search space to intra-chromosomal SNP pairs within a specified distance. Specifically, we assume the following:

$$r^2(s_i, s_j) = \begin{cases} 0 & \text{if } s_i \text{ and } s_j \text{ are } \textit{not} \text{ on the same chromosome} \\ 0 & \text{if } d(s_i, s_j) > D \end{cases} \tag{12}$$

where $d(s_i, s_j)$ is defined as the intra-chromosomal distance between SNPs s_i and s_j (i.e. the distance on the genome) and D is an adjustable parameter (unit in base pairs) to control the time/performance trade-off of the algorithm by limiting the search space for the redundancy estimations. Note that, we consider the $d(s_i, s_j)$ to be infinite for SNP pairs that are on different chromosomes.

2.2.3 Formulation of Macarons algorithm

Overall, with the three assumptions given in Equations (6 , 10 and 12), the penalty function becomes as follows:

$$\begin{aligned} \bar{r}^2(s_i,s_j) &= \left\{ \begin{array}{ll} r^2(s_i,s_j), & \text{if } d(s_i,s_j) \le D \\ 0, & \text{otherwise} \end{array} \right\} \\ \overline{R}^2(S_t,s_x) &= 1 - \prod\nolimits_{s_i \in S_t}\left(1-\bar{r}^2(s_i,s_x)\right) \end{aligned} \quad (13)$$

Thus, the gain function becomes:

$$\begin{aligned} G(S_t,s_x) &= c_x\left(1-\overline{R}^2(S_t,s_x)\right) \\ &= c_x \prod\nolimits_{s_i \in S_t}\left(1-\bar{r}^2(s_i,s_x)\right) \end{aligned} \quad (14)$$

The Macarons algorithm that encodes this gain function for step-wise SNP selection is given in Algorithm 2. Overall, it has O(nk + $\lambda_D$mk) run time complexity where the first term is for maximizing the gain function, and the second term is for computing the gain function, which require the measurement of correlations from data. Here, λ_D is a variable between $[1, n]$ dependent on the D parameter. It represents the average number of SNP pairs that require the computation of correlation for a given D threshold. Thus, the overall complexity for small D is O(nk) when the first term dominates and O(nmk) for large D as the computation of Pearson correlations becomes the bottleneck.

Algorithm 2 Macarons Algorithm

Input: Ground SNP set V, chromosome numbers and positions for all SNPs $s_i \in V$, the phenotype association scores c_i for all $s_i \in V$, cardinality constraint $k \le |V|$, trade-off parameter $D \ge 0$ in base pairs
Output: Set $S \subseteq V$ such that $|S| = k$.

```
    $S \leftarrow \varnothing$
    $G \leftarrow \{c_i \forall i \in V\}$
    while $|S| < k$ do
        $s_i \leftarrow \text{argmax}_{s_x \in V \setminus S}\, G(s_x)$
        $S \leftarrow S \cup \{s_i\}$
        for all $(s_i,\, s_x)$ with $d(s_i, s_x) \le D$ do
            $\bar{r}(s_i, s_x) \leftarrow$ compute Pearson's correlation between $s_i$ and $s_x$
            $G(s_x) \leftarrow G(s_x)\left(1-\bar{r}^2(s_i,s_x)\right)$
        end for
    end while
```

2.2.4 Optimizing Macarons algorithm for runtime

The gain function given in Equation (14) is monotonically non-increasing with respect to the growing set of selected SNPs (i.e. at each iteration, the gain of a SNP either stays the same or decreases). Moreover, we know that the selected SNP set will approximately grow according to their individual association scores (this is particularly true for low k and D parameters since there would be less deviation from individual scores). Here, we leverage these properties to further optimize the runtime of the algorithm. For this purpose, we first sort all SNPs according to their phenotype association scores c_x (such that $c_i \ge c_j$ if $i < j$). Then, we limit the search space of the algorithm to an active region consisting of N_{active} most promising SNPs with highest individual scores (having an initial size of $N_{\text{active}} = \psi$). When the current search space is insufficient (this can be detected by comparing the gain function with the individual scores), we grow the active region by a factor of $\gamma > 1$. Specifically, when the maximum value of gain function is greater than or equal to the minimum individual score in the active region (i.e. when $\max_{x \le N_{\text{active}}}(G(S_t,s_x)) \ge \min_{x \le N_{\text{active}}}(c_x)$), we know that active region is sufficient (since we know $\min_{x \le N_{\text{active}}}(c_x) > c_j \ge G(S_t,s_j) \forall \{j > N_{\text{active}}\}$). Otherwise, the action region might be insufficient, thus, we grow the active region to include the most promising $\lceil \gamma N_{\text{active}} \rceil$ SNPs and repeat this process as necessary. The optimized Macarons algorithm that implements this idea is given in Supplementary Algorithm S1. Note that, the output of this algorithm is always equal to the output of Algorithm 2 regardless of the parameter values (i.e. the parameters ψ and γ does not change the output, only affects the runtime). In our experiments, we use $\psi = 1000$ and $\gamma = 2$ unless otherwise specified.

3 Results

3.1 Experimental setup

3.1.1 Summary of the experiments and the results

First, we investigate the effect of limiting the search of Macarons using intra-chromosomal distance (D parameter) in terms of redundancy and runtime, and whether Macarons can successfully avoid the selection of highly redundant SNPs (Fig. 1). Then, we compare Macarons with other SNP selection methods on a small but comprehensive dataset (*AT* dataset with 17 flowering time phenotypes) in terms of their predictivity, runtime and redundancy characteristics (Figs 2 and 3; Supplementary Fig. S1) and investigate the trade-off between different assumption models in Macarons (Fig. 4). Next, we demonstrate that Macarons can seamlessly scale to large datasets with $\sim 10^7$ variants (in human height dataset). Afterwards, we investigate the utility of avoiding redundancy with Macarons over using a fixed threshold based on individual phenotype association scores on two larger datasets (rice700k and human height) based on two different association scores (Fig. 5) and we inspect the characteristics of Macarons by visualizing the correlation structure of the selected SNPs while marking the ones near coding regions (Fig. 6). Finally, we benchmark the utility of using Macarons in conjuction with various regression models (Fig. 7).

3.1.2 Datasets

For a considerable portion of our analysis (e.g. for the comparisons with other SNP selection methods), we use the *Arabidopsis Thaliana* (AT) dataset (Atwell *et al.*, 2010) which provides data for 17 flowering time phenotypes. The availability of multiple phenotype data helps to estimate the variance in phenotype prediction performance more accurately. Also, this is relatively small dataset where the number of samples is between 119 and 180 (depending on the phenotype), and there are 214 051 SNPs before any filtering. Thus, this dataset allows us to test the performance of some methods that would otherwise not scale to larger datasets. In our analysis, we filter out variants with minor allele frequency (MAF) of < 10%, which remains 173 219 SNPs.

As an additional dataset, we use the rice700k data (McCouch *et al.*, 2016) which contains 1145 samples and 700 000 SNPs before filtering. Here, the phenotype is related to the rice grain-length. In our analysis, after applying a MAF < 5% filter, 463 907 SNPs remain. Note that, this is a medium-sized dataset that is roughly 20 times larger than the *AT* dataset.

As our largest dataset, we consider the human height data (https://zenodo.org/record/1442755) collected from openSNP, which is a crowd-sourced genetic test sharing website (Greshake *et al.*, 2014). It was prepared by researchers from École Polytechnique Fédérale de Lausanne (EPFL) as a part of a machine learning challenge on CrowdAI (https://www.crowdai.org/challenges/opensnp-height-prediction). This dataset contains human height data for 784 individuals and 7 252 636 SNPs. Thus, this dataset is about 10 times larger than the rice700k dataset (and about 200 times larger than AT dataset).

3.1.3 Phenotype association scores

For consistency with the previous results (Azencott *et al.*, 2013; Yilmaz *et al.*, 2019), we use SKAT (Wu *et al.*, 2011) to score the individual phenotype association of each SNP, unless otherwise

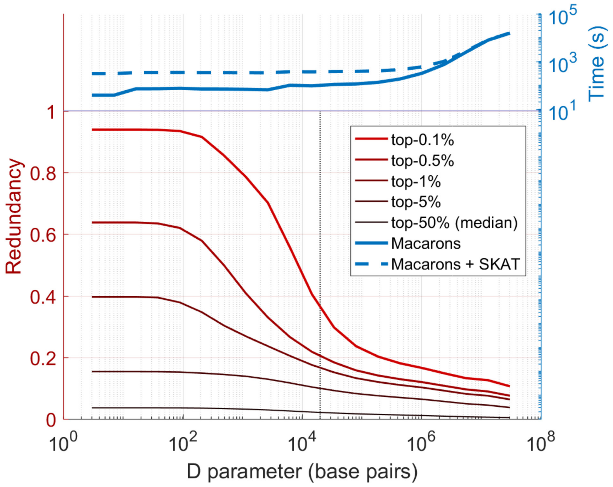

Fig. 1. The characteristics of Macarons algorithm with respect to its D parameter in terms of the redundancy between the selected SNPs and the runtime of the algorithm. Redundancy (measured by squared correlation) between all pairs of selected SNPs in AT data for $k = 1000$. Each line indicates a different percentile for the distribution of redundancy (e.g. top-1% line indicates 1%th most redundant pair). Note that, the lines are averaged across 17 runs corresponding to different flowering time phenotype of AT. The left-most point at the x-axis (for $D = 1$ parameter) corresponds to the baseline method of selecting the highest scoring SNPs (without applying any penalization or regularization). The dotted line indicates *a priori* selected D parameter value of 20 kbp. At the top, the blue lines indicate the total runtime (in seconds) to run Macarons for the corresponding D parameter (the dashed line additionally includes the time to compute SKAT phenotype association scores)

specified (as a part of one of the experiments, we also run our method with another phenotype association measure). While computing the SKAT score, we use the top principal component of the genotype matrix to alleviate the effect of the population stratification (Price *et al.*, 2006).

3.2 Effect of limiting the search space through intra-chromosomal distance

The premise behind Macarons is to select a complementary set of SNPs while avoiding redundant (correlated) SNPs that are in LD. As we discuss in the methods, the process of taking into account of all redundant SNPs overall requires $k \times n$ (number of selected SNPs × number of SNPs) correlation estimations from the data, which is both computationally intensive and superfluous since most highly correlated variants tends to be closely located on the genome. To overcome this issue, we limit the search space for correlated SNPs to close intra-chromosomal pairs with maximum distance of D, where D is an adjustable parameter (unit in base pairs).

Here, we investigate the effect of the parameter D on the SNPs selected by Macarons, particularly to examine its effect on the selection of highly correlated SNPs. For this purpose, we select $k = 1000$ SNPs for each of the 17 flowering time phenotypes of *AT* using Macarons with various D parameter values. For each tested value of the D parameter, we investigate the distribution of the redundancy for $\binom{1000}{2} \times 17$ pairs of selected SNPs, in addition to the overall runtime characteristics of Macarons (Fig. 1). Note that, since the distribution of the redundancy is greatly skewed (i.e. many pairs with low redundancy, a few with high redundancy), we visualize it using percentile lines (similar to a boxplot) starting from the 50th percentile (median) all the way to the 99.9th percentile, denoted as top-0.1% redundancy. As it can be seen in Figure 1, with *a priori* selected D value of 20 kbps [which is the estimated LD range for *AT* according to Atwell *et al.* (2010)], Macarons can considerably reduce the selection of highly redundant SNP pairs without being bottleneck from a runtime perspective (since the association scores needs to be computed regardless of the D parameter or the redundancy calculations). We observe that the redundancy calculations only start to become a bottleneck after around $D = 10^6$ base pairs. Next, we investigate whether avoiding the selection of redundant pairs would translate into an improved phenotype prediction performance by comparing the Macarons with other SNP selection methods.

3.3 Benchmarking SNP selection methods

3.3.1 Compared methods

We compare Macarons with the following methods:

- Baseline: A simple greedy approach that selects the top k SNPs with the highest individual phenotype association scores. This method becomes equivalent to Macarons when the search space (D) parameter of Macarons is set to 0 (since no redundancy calculations are made and phenotype association scores are not updated in that case). This method considers the association of each SNP independently, thus, serves as a baseline for other SNP selection methods that attempt to take into account of interactions or dependencies between selected SNPs in some manner.
- SConES: A SNP selection algorithm that rewards SNPs according to their individual phenotype association scores of SNPs while employing a connectivity constraint on an SNP–SNP network (Azencott *et al.*, 2013). It features two parameters λ and η that controls the connectivity and sparsity constraints respectively.
- SPADIS, our previous work, rewards SNPs according to their individual phenotype association scores of SNPs while applying a diversity penalty based on the shortest-path distances on an input network (Yilmaz *et al.*, 2019). It features three parameters k (for number of SNPs selected), β (for the strength of penalization) and D (for limiting the search range in the network).
- Lasso: A linear regression method with l_1 (lasso) regularization that forces the regression weights of some features (SNPs) to be zero. SNPs with non-zero weights are considered to be selected. It has one parameter λ that determines the strength of regularization and therefore the sparsity (size) of the selected SNP set.

For methods that utilize a SNP–SNP network (i.e. SPADIS and SConES), we use the best performing network. Based on the results of previous benchmarkings (Azencott *et al.*, 2013; Yilmaz *et al.*, 2019): Genomic sequence network (where SNPs that are adjacent on the chromosome are connected) for SPADIS, and Genomic interaction network (where SNPs that are in the same genomic region as well as the SNPs between interacting genes are connected to form cliques).

Since Macarons has interpretable parameters and does not require a parameter optimization procedure, we tested it for two *a priori* selected D values. We choose $D = 20$ kbp as suggested by (Atwell *et al.*, 2010), and we also test $D = \infty$ (which covers the entire chromosome and includes all intra-chromosomal pairs) to see the effect of limiting the search space on phenotype prediction performance.

Note that, to compare phenotype prediction performances of the methods on equal footing, we apply a cardinality constraint k on the selected SNP set and compare the results of the algorithms for different values of k. To control the number of SNPs selected, the baseline method, SPADIS and Macarons already has a parameter k that we can set directly. On the other hand, SConES and lasso features sparsity parameters that indirectly controls the size of the selected SNP subset. For these methods, we apply a binary search and select the sparsity parameters (η for SConES, λ for lasso) that yield the closest number of selected SNPs to the predefined cardinality constraint k.

3.3.2 Evaluating phenotype prediction performance

Our testing scheme consists of using a nested cross-validation scheme (outer for evaluation, inner for parameter selection). First, we use 10 cross-validation folds to split the data into training and test samples. For each of the 10 cross-validation folds, we compute phenotype association scores and run the SNP selection methods

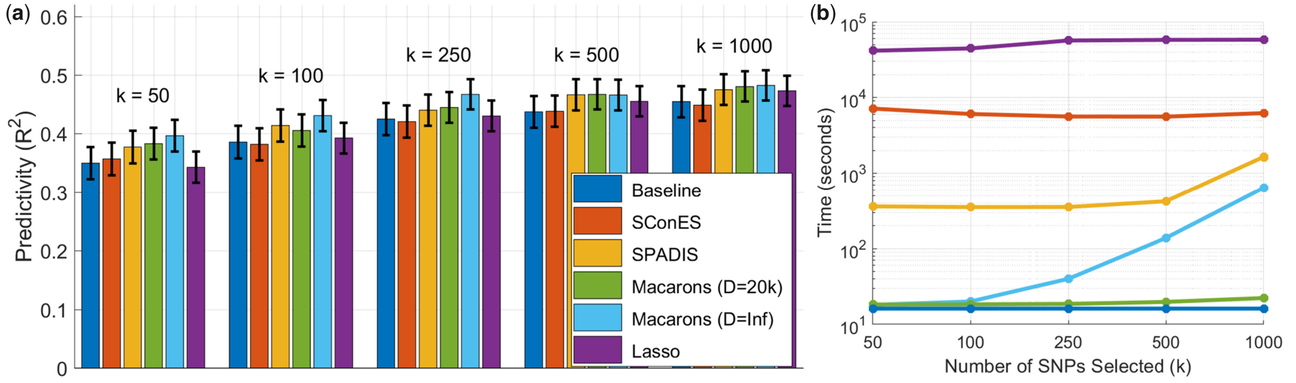

Fig. 2. The phenotype prediction performances and runtimes of the SNP selection methods for different number of selected SNPs (indicated by k values). The methods are tested for $k = 50$, 100, 250 and 1000 selected SNPs. (**A**) Each colored bar represents a different SNP selection method. The *y*-axis shows the averaged Predictivity (measured by Pearson's squared correlation coefficient, R^2) across all 17 flowering time phenotypes. The black lines indicate the 95% confidence interval for the average R^2 performance for the corresponding method and the *k* value. (**B**) Each line indicates time required (in seconds) to run the corresponding method for a flowering time phenotype of AT. Note that, since AT dataset consists of 17 phenotypes, the values shown for runtime are averaged across all phenotypes

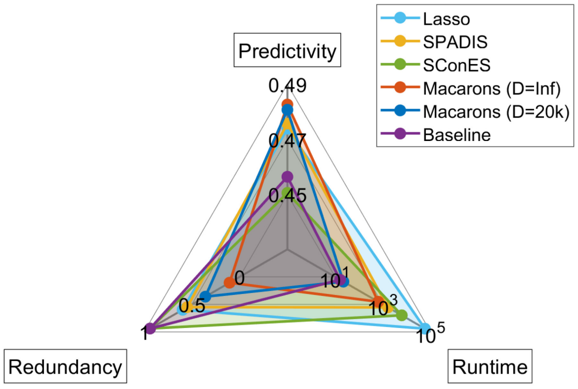

Fig. 3. Top-level overview of the characteristics of the SNP selection methods in terms of their predictivity, redundancy and runtime. The predictivity indicate the average phenotype prediction performance (measured by R^2) of the corresponding method for $k = 1000$ selected SNPs. The redundancy axis indicates the presence of high correlation among the selected SNPs (measured by top-0.1% redundancy: 99.9th percentile of the squared correlation between all pairs of selected SNPs). The time axis (in log-scale) shows the average time required (in seconds) to run the corresponding method in the AT dataset. Note that, since AT dataset consists of 17 phenotypes, the values shown for runtime are averaged across all phenotypes

using training portion of the data, and we predict the phenotype on the test portion using ridge regression. Next, we assess the prediction performance using Pearson's squared correlation coefficient (R^2) between the predicted and observed (actual) phenotype vectors. Note that, some methods (e.g. SPADIS and SConES) require further cross-validation to tune their parameters. For this purpose, we use a nested-5-fold cross-validation where the training portion of the data is further split into five validation folds. On these validation folds, the model's generalizability to unseen samples is measured by using ridge regression with R^2 and the parameters with highest R^2 are selected. Since Macarons's parameters are selected *a priori*, it does not require this nested cross-validation procedure.

3.3.3 Comparison of SNP selection methods

Here, to compare the performances of different SNP selection methods, we use the AT dataset because it has two main advantages: (i) it contains 17 flowering time phenotypes that allows us to more accurately estimate the phenotype prediction performance (by reporting the averages over all flowering time phenotypes) and (ii) it is relatively small dataset (with $\sim 10^2$ samples, $\sim 10^5$ SNPs) which allows us to report results for relatively slow methods (e.g. lasso) that would otherwise not scale to larger datasets. For methods that utilize a phenotype association score (i.e. for all tested methods except lasso), we use SKAT score mainly for consistency with previous benchmarkings that use this dataset (Azencott *et al.*, 2013; Yilmaz *et al.*, 2019).

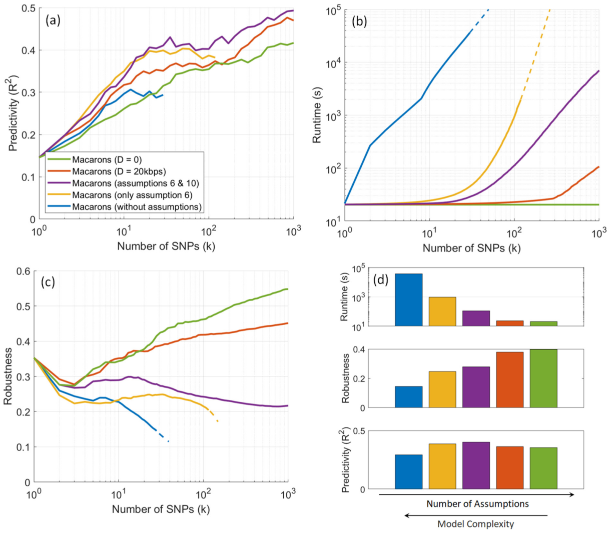

Fig. 4. The predictivity, runtime and robustness characteristics of Macarons under various assumption models on AT data. (**a**) The phenotype prediction performance (measured by R^2) versus the number of selected SNPs (k). Each colored line shows the performance of Macarons under different assumption models. (**b**) The average time required (in seconds) to run the corresponding model in the AT dataset. Note that, since AT dataset consists of 17 phenotypes, the values shown for runtime are averaged across all phenotypes. (**c**) Robustness (measured by the average overlap in the selected sets between different cross-validation folds), (**d**) A snapshot of the predictivity, runtime and robustness characteristics of the models for a fixed k ($k = 100$ for all models except the most complex model without assumptions, due to limiting runtime, that one is given for $k = 30$). The models are ordered such that the ones on the left are more complex models with less simplifying assumptions

First, we run each method on each of the 17 flowering time phenotypes for $k = 1000$ and assess their 10-fold cross-validated phenotype prediction performance (using ridge regression as the prediction model and measuring by R^2). In Supplementary Figure S1, we report the prediction performance of the methods relative to the performance of the baseline method (of selecting the top-k SNPs that are most associated to the phenotype individually). Here, we make the following observations:

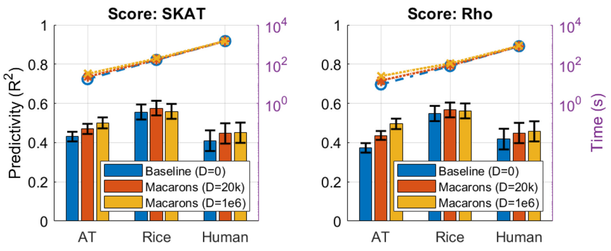

Fig. 5. Contribution of Macarons in improving the prediction performance for various datasets and phenotype association scores. The bars represents the predictivity (measured by R^2) of the selected SNPs by the corresponding method for $k = 1000$. The blue bars indicate the prediction performance of baseline method (filtering based on individual phenotype association scores), whereas, red and orange bars indicate the performance of Macarons for various datasets (AT, Rice, Human height) and association scores (SKAT, Rho). The black error bars indicate the 95% confidence intervals. The dashed lines on the top side indicate the runtime of the corresponding method in seconds. Note that, since AT dataset consists of 17 phenotypes, the values shown for predictivity and runtime are averaged across all phenotypes

- SConES does not seem to perform better than the baseline method for predicting the phenotype. We argue that this may be because the network connectivity constraint in SConES reinforces the selection of highly correlated SNPs that are in LD, which likely pose difficulties for the regression step.
- SPADIS and Macarons (with $D = 20$ kbp) seem to perform quite similarly while both having a higher phenotype performance than the baseline method on most phenotypes.
- Lasso and Macarons (with $D = \infty$, measuring the redundancy of all intra-chromosomal SNP pairs) seem to perform similarly while lasso performs considerably worse than the baseline method on two of the phenotypes.
- For Macarons, using $D = \infty$ to expand the search space over using $D = 20$ kbp does not seem to provide a considerable benefit in phenotype prediction performance for most phenotypes.

Next, in Figure 2, we consider the averaged phenotype prediction performances (R^2, denoted predictivity for brevity) across all 17 phenotypes for various number of selected SNPs ($k = 50, 100, 250, 500$ and 1000). Here, our first observation is that the overall performances of all methods consistently increase as the number of selected SNPs (k) is increased. We argue that this is because ridge regression can provide an adequate amount of regularization and improve the predictivity even for relatively large k (where $k > m$, the number of samples). Secondly, we observe that, for each computational experiment (for different k), the prediction performance of Macarons (for either of the D parameter values) is consistently similar or better than all other methods although there is not sufficient statistical power to conclude that one method has significantly better predictive performance than the others at 95% confidence level for any k experiment. Whereas, when we look at the average performance across the five computational experiments for different k (Supplementary Fig. S2), we observe that Macarons have a significantly higher prediction performance than baseline method, SConES and Lasso, while having a similar performance to SPADIS.

Additionally, in Figure 2 (right panel), we compare the methods in terms of the runtime required to run them on the AT dataset (we perform the time measurement on a 40 core machine with Intel(R) Xeon(R) CPU E5-2650 v3 2.30 GHz, parallelized on 17 threads for phenotypes). For each method, we report the CPU runtime averaged across 17 phenotypes with respect to k. Note that, the reported times include the method runs, 10-fold cross-validation used for evaluation, the calculation of association scores and (if any) the cross-validation for parameter tuning.

As it can be seen on Figure 2, Macarons with $D = 20$ kbp is at least two orders of magnitude faster than other methods (i.e. SPADIS, SConES, lasso), and compared to the baseline method of using individual association scores for subset selection, improves the predictivity and the redundancy characteristics (Fig. 1) of the selected SNP subsets. We also observe that, even though considering all intra-chromosomal pairs (with $D = \infty$) in Macarons does not provide an additional benefit in predictivity over using $D = 20$ kbp for $k = 1000$, the performance of Macarons $D = \infty$ is typically higher than $D = 20$ kbp for lower k values. This indicates that, for target subsets of small size, increasing the depth of the search space through D parameter might be a more optimal choice.

In Figure 3, we summarize the differences and potential trade-offs between different SNP selection methods by considering three metrics: (i) Predictivity (measured by R^2) for phenotype prediction; (ii) Runtime in seconds; and (iii) Redundancy (measured by top-0.1% redundancy, in a similar manner to the results in Fig. 1) that investigates the presence of highly redundant SNP pairs in the selected SNP subset. Overall, Figure 3 suggests that Macarons ($D = 20$ kbp) can offer a good trade-off between different characteristics, with decent predictivity, fast runtime and a moderate level of redundancy.

In Supplementary Text S2, we also investigate the concordance of the selected SNPs by Macarons based on the candidate genes obtained from (Segura *et al.*, 2012) on the AT dataset.

3.4 The impact of the simplifying assumptions in Macarons

Next, we investigate the effect of the simplifying assumptions in Macarons on important characteristics like model predictivity, runtime and robustness. Overall, we utilize three simplifying assumptions in Macarons:

- Assumption in Equation (6) (assuming that the overlap between a candidate SNP and the selected SNP set does not depend on their overlap with the phenotype, Y). This assumption results in a gain function (Equation 7) that is monotonically non-decreasing with respect to the increased set size. This monotonicity allows the optimized algorithm (given in Supplementary Algorithm S1) to be used rather than the straightforward implementation described in Algorithm 1.
- Assumption in Equation (10) (assuming that the partial correlation between two SNPs in the set does not depend on other SNPs in the set, thus, are equal to their zero-order correlation). This assumption eliminates the need for making high-order correlation estimations from data, thus allowing the optimization of SNP sets with cardinality larger than $k>m$ (where m is the number of samples).
- Assumption in Equation (12) (assuming that SNPs that are more than D base pairs apart are not correlated).

Thus, we run Macarons with different versions of these assumptions, where the main difference between these versions is the definition of the gain function that determines which SNP is to be added to the set next. Overall, we consider the following five models (from the most complex to the least complex):

- Macarons (without assumptions): This is a straightforward model implementing Algorithm 1 without any of the assumptions in Equations (6) and (10).
- Macarons (only assumption 6): Here, since we make assumption 6, the gain function becomes monotonic, which allows us to utilize the optimized algorithm to speed-up the computation drastically.
- Macarons (assumptions 6 and 10): Here, the inclusion of assumption 10 allows us to eliminate the high-order estimations from data, thus allowing SNP sets larger than $k > m$ to be

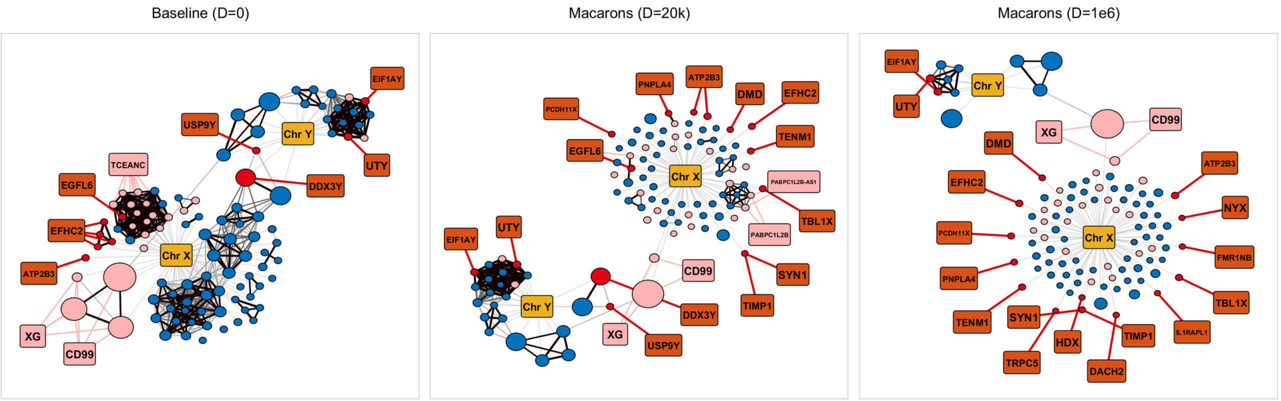

Fig. 6. Visualization of the selected SNPs on human height dataset for $k = 100$. Each panel corresponds to a different SNP selection method (Baseline method of selecting top-k SNPs with highest association, Macarons with $D = 20$ kbp, and Macarons with $D = 10^6$ base pairs). The circles indicate selected SNPs and the rectangles indicate genes (colored red or light red) or chromosomes (colored yellow). Weighted edges between SNPs indicate their redundancy (measured by squared correlation R^2, We include pairs with $R^2 \geq 0.35$ and we highlight highly redundant pairs with $R^2 \geq 0.7$ with thick lines and black color). The red colored SNPs are within coding region, and light red colored SNPs are within ±20 kbp around coding region. Similarly, we use red (light red) color for genes with at least one selected SNP in a coding region (around ±20 kbp of coding region). The sizes of the circles (SNPs) indicate the strength of their individual association with the phenotype (measured by R^2)

considered. Note that this model does not limit the search according to D and corresponds to Macarons with $D = \infty$.

- Macarons ($D = 20$ kbps): Again, with the assumptions 6 and 10, but also assuming that only SNPs that are less than 20kbps apart are correlated. This is the proposed Macarons version that we run.
- Macarons ($D = 0$): This is the simplest model we consider that assumes no SNPs are associated with each other. This is equivalent to the baseline method of using univariate associations (i.e. selecting top k SNPs with highest individual associations with the phenotype).

We investigate the performance of the selected SNP sets by these methods for different k values on the AT dataset (across the 17 phenotypes). For this purpose, we consider three metrics: the predictivity (R^2), the time performance and robustness (the consistency of the selected SNP sets across different cross-validation folds, measured by jaccard index).

As it can be seen in Figure 4, more complex models take more time to run and the formulated assumptions considerably improve the runtime performance as expected. Notably, we also observe that decreasing the complexity has another important benefit of improving the model's robustness to noise. Namely, we observe that models without the simplified assumptions are noticeably less robust compared to simplier models (e.g. Macarons with $D = 0$, or $D = 20$ kbps).

In phenotype prediction, we observe that Macarons (assumption 6) and Macarons (assumption 6 and 10) follows similar performance curves (Fig. 4a), which suggests that assumption 10 does not have a strong effect on the predictive performance of the models and is likely to hold. Here, we also observe that predictive performance typically increases with the selected SNP set size k, and simpler models with larger sets can offer more predictive performance compared to complex models that are limited to smaller sets.

In Figure 4d, we also present a snapshot of the characteristics of these models for a fixed k value ($k = 100$ for all models except the most complex model, which we selected $k = 30$ due to time issues). The models are ordered by their complexity (models with more simplifying assumptions are on the right). Here, we clearly observe the trade-off between predictivity, runtime, robustness and model complexity: More complex models are slower and less robust, but (presumably) better fits/explains the given data. Whereas, the model with the best cross-validated predictivity is in the middle, representing a good trade-off point between the model fit and the model's robustness to noise.

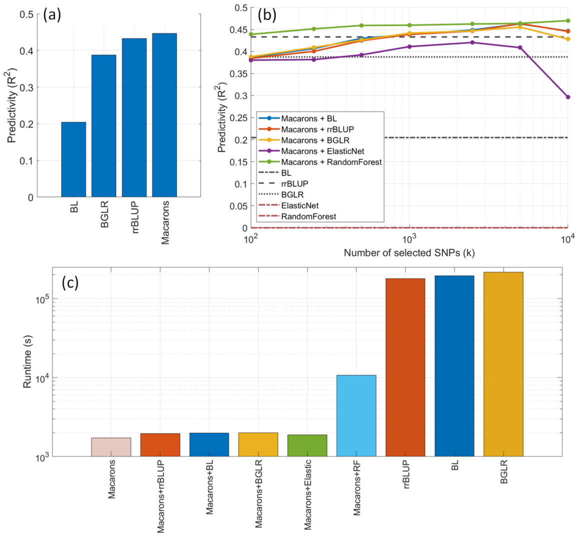

Fig. 7. The prediction performances and runtimes of various regression methods and their Macarons-enhanced versions on the human height dataset. (**a**) The prediction performances (R^2) of three regression methods (BL, BGLR and rrBLUP) that can run using all SNPs, as well as the performance of Macarons (for $k = 1000$ and $D = 20$ kbps, using ridge regression for predictions). (**b**) The prediction performances (R^2) of the Macarons-enhanced regression methods (labeled Macarons + methods) with respect to the number of SNPs (k) parameter of Macarons. The dashed or dotted black lines indicate the standalone R^2 of the corresponding regression method (using all SNPs). Note that, it is not possible to run the random forest and elastic net regression using all SNPs due to time and memory constraints

3.5 Contribution of using Macarons to take dependencies between variants into account

Here, we investigate the effect of Macarons (and avoiding redundancy between the variants) on the characteristics of the selected subsets. First, we compare Macarons with the baseline method (which does not take dependencies into account) in terms of phenotype prediction performance. For this purpose, we benchmark the methods on three datasets (AT, and two larger datasets: rice700k, and human height) and two phenotype association scores: (i) SKAT as done in previous sections and as an alternative measure (ii) absolute Pearson correlation which is denoted as Rho (ρ). For this analysis, we predict the phenotype using ridge regression on $k = 1000$ selected SNPs and report the performance using R^2. For Macarons, we consider two D parameter values: (i) $D = 20$ kbps as previously done and (ii) $D = 10^6$ bp which is approximately the maximum D values before Macarons becomes a bottleneck in terms of runtime (according to our analysis on AT data, Fig. 1).

As it can seen from Figure 5, we observe a consistent increase in phenotype predictivity when using Macarons across different datasets and association scores although the magnitude of the increase depends on the datasets (e.g. rice dataset exhibits minor differences while the differences in AT are more prominent). In addition, we observe that different association scores result in similar prediction performances. In Figure 5, we also report the overall runtime of the methods (total 10 runs for 10-cross validation folds). As it can seen, Macarons can scale to large datasets (e.g. human height data with $\sim 10^7$ variants) without compromising from runtime (i.e. the computation of phenotype association scores becomes the bottleneck rather than the subset selection).

Next, to elucidate the effect of taking into account of dependency on the characteristics of the selected subset, we visualize the selected SNPs (for $k = 100$ subsets on the human height dataset using ρ for phenotype associations) in the form of a correlation network while marking the variants in the coding regions or near ±20 kbp of the coding regions (Fig. 6). Our first observation is that there are some highly correlated clusters of variants in the selections of baseline method (can be considered as Macarons with $D = 0$ bp, which considers the variants to be independent). Whereas, these tightly coupled clusters starts to disappear as a higher portion of the dependencies between the variants are taken into account with higher D parameter (to the point that there are only a few pairs that are highly correlated for $D = 10^6$ bp). Interestingly, we also observe that avoiding the redundancies during the subset selection leads to the selection of more variants in coding regions (and more genes with at least one selected variant in their coding region, Supplementary Table S1). Nevertheless, most of the selected variants are not near coding regions, including some of the highly associated ones (Fig. 6). Note that, for the human height dataset, all $k = 100$ selected SNPs (regardless of the method used) turn out to be either in chromosome X or Y. This is likely because gender is a strong predictor of height, e.g. there is considerable difference in the mean heights of males and females in this dataset (males: 1.79 m and females: 1.65 m).

3.6 Improvement of Macarons over a broad range of regression methods

We investigate the utility of the SNP subsets obtained by Macarons when used in conjuction with various regression methods (other than ridge regression used in the previous analyses) on the human height dataset. For this purpose, we consider five well-established regression models which are: (i) rrBLUP (Endelman, 2011); (ii) Bayesian Lasso (BL) (Park and Casella, 2008); (iii) BGLR (BayesA model) (Pérez and de Los Campos, 2014); (iv) Elastic-Net regression; and (v) Random Forest. Here, the first three methods (rrBLUP, BL and BGLR) are iterative methods that are designed to handle a large number of features. Thus, these can be run on the entire genome (even for a large dataset with high dimensionality like the human height data), while the Random Forest and Elastic Net models are not optimized enough to run on the entire dataset due to runtime and/or memory issues.

First, we benchmark the predictive performance (R^2) of these methods on the human height dataset and compared them with Macarons (followed by ridge regression) for $k = 1000$ and $D = 20$ kbps. The results of this analysis are provided in Figure 7a. Here, we observe that Macarons followed by ridge regression can outperform rrBLUP, BL and BGLR methods while being two magnitudes faster (Macarons framework takes 25 min to run, while the others take 50–60 h to run, Supplementary Fig. S3).

Next, we investigate the performance of Macarons-enhanced regression models (rrBLUP, BL, BGLR, Random Forest, Elastic-Net) that are run using k SNPs selected by Macarons (for $D = 20$ kbps) on the human height dataset. As it can be seen in Figure 7b (for R^2) and Supplementary Figure S4 (for mean squared error), using Macarons in conjunction with rrBLUP, BL and BGLR improves their prediction performance compared to running them alone using all SNPs, while dramatically reducing the runtime as much as 100x (Fig. 7c). Particularly, in the case of BL, we observe that, even though using BL alone has a considerably lower performance, using Macarons together with BL results in a comparable performance to other regression methods. Overall, the results of these experiments suggest that using Macarons to reduce the feature space can benefit various regression methods both from a perspective of prediction performance as well as runtime.

In addition, using Macarons to filter the feature space allows us to run regression methods that would otherwise not be possible to do so, such as Elastic-Net and Random Forest. Most notably, we observe that Macarons + Random Forest has the highest prediction score across all k values compared to all other methods tested, while still being an order of magnitude faster than running rrBLUP, BGLR and BL on the whole dataset. This suggests that running a more sophisticated, non-linear method using a carefully selected subset of features could be a good strategy to improve the predictive performance further.

Finally, we investigated the performance of using the baseline method of univariate selection (i.e. selecting the top k SNPs with highest associations) instead of Macarons to filter the feature space of the regression methods. As shown in Supplementary Figure S5, we observe that selecting using Macarons consistently increases the prediction performance across all regression methods without compromising the runtime (Supplementary Fig. S3).

3.7 Suggested settings for using Macarons

For the maximal chromosomal distance (D) parameter, we recommend an analysis similar to the one in Figure 1 and suggest the use of a default D value of 20 kbps based on our results. Whereas, for selecting the number of SNPs parameter (k), our recommendation is to select the highest possible k based on the available computational and experimental resources in mind (as a general guideline, we suggest $k = 1000$ for use with ridge regression, and $k = 10\ 000$ for use with rrBLUP regression as good initial values to consider) and fine-tune it with the help of a cross-validation analysis as in Figure 7b and Supplementary Figure S6. In Supplementary Text S3, we detail our reasoning and suggestions on the selection of k.

4 Discussion

In order to select a complementary set of SNPs for the prediction of quantitative phenotypes, we develop Macarons, a fast and interpretable model with a simple idea: the joint selection of highly dependent SNPs would be redundant and would not provide complementary information for the prediction of a phenotype.

Overall, this task is known as feature selection in the machine learning literature, and the idea to take redundancy into account is applied extensively. However, most of the established feature selection methods do not scale (from a runtime standpoint) to the SNP selection problem due to the high dimensionality of the GWAS data (e.g. typically up to $\sim 10^7$ variants). Furthermore, such methods suffer from over-fitting since the number of variants is much larger than the number of samples.

To overcome these issues, we make simplifying assumptions (as shown in Equation (6) and Equation (10)) and limit the search space to intra-chromosomal pairs in close proximity (controlled by a parameter D in base pairs, Equation (12).

Our results demonstrate that, with the assumptions and the optimizations in its algorithm, Macarons can seamlessly scale to variant sets as large as $\sim 10^7$ in a matter of minutes. We expect that Macarons (with $D = 20$ kbp, or up to $D = 10^6$ base pairs) can be of practical use in large GWAS studies since it can take into account of the dependencies between the variants without compromising runtime. Overall, it can offer a reasonable trade-off between phenotype predictivity, runtime and redundancy of the selected subsets.

The intra-chromosomal distance idea and D parameter in Macarons can be efficiently generalized to input dependency networks (where the presence of an edge indicates the decision to measure redundancy for that SNP pair, for example, the D parameter can be represented as connecting close SNPs as cliques in the network) to limit the search space of the algorithm. We provide a version of Macarons with input dependency network in our implementation though we leave experimentation with it as future work. We expect that this would be useful to take into account of the dependency between variants through more sophisticated models, for example, by considering the 3D structure of the chromosome through Hi-C data.

Macarons can be used in combination with any metric for individual phenotype association (including for dichotomous phenotypes). We expect that Macarons can be especially useful as a part of a multi-stage analysis for performing the initial filtering to reduce the search space, followed by epistasis tests or other subsequent analyses. Overall, the framework we present can be generalized to various other feature selection problems involving high dimensionality within and beyond biomedical applications.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Ardlie,K.G. *et al.* (2002) Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.*, **3**, 299–309.

Atwell,S. *et al.* (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Azencott,C.-A. *et al.* (2013) Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, **29**, i171–i179.

Caylak,G. *et al.* (2020) Potpourri: an epistasis test prioritization algorithm via diverse SNP selection. *J. Comput. Biol.*, **28**, 365–377.

Cordell,H.J. (2009) Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Cowman,T. and Koyutürk,M. (2017) Prioritizing tests of epistasis through hierarchical representation of genomic redundancies. *Nucleic Acids Res.*, **45**, e131.

Das,A. and Kempe,D. (2011) Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11. Omnipress, Madison, WI, USA, pp. 1057–1064.

Das,A. *et al.* (2012) Selecting diverse features via spectral regularization. *Adv. Neural Inf. Process. Syst.*, **25**, 1583–1591.

Ding,X. *et al.* (2015) Searching high-order SNP combinations for complex diseases based on energy distribution difference. *IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB)*, **12**, 695–704.

Endelman,J.B. (2011) Ridge regression and other kernels for genomic selection with r package rrblup. *Plant Genome*, **4**, 250–255.

Evans,D.M. *et al.* (2006) Two-stage two-locus models in genome-wide association. *PLoS Genet.*, **2**, e157.

Fang,G. *et al.* (2012) High-order SNP combinations associated with complex diseases: efficient discovery, statistical power and functional interactions. *PLoS One*, **7**, e33531.

Goldstein,D.B. (2009) Common genetic variation and human traits. *N. Engl. J. Med.*, **360**, 1696–1698.

Grave,E. *et al.* (2011) Trace lasso: a trace norm regularization for correlated designs. *Adv. Neural Inf. Process. Syst.*, **24**, 2187–2195.

Greshake,B. *et al.* (2014) opensnp—a crowdsourced web resource for personal genomics. *PLoS One*, **9**, e89204.

Jacob,L. *et al.* (2009) Group lasso with overlap and graph lasso. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, pp. 433–440.

Jia,P. *et al.* (2011) dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, **27**, 95–102.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Liu,Y. *et al.* (2017) Sigmod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. *Bioinformatics*, **33**, 1536–1544.

Lou,X.-Y. *et al.* (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.*, **80**, 1125–1137.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

McCouch,S.R. *et al.* (2016) Open access resources for genome-wide association mapping in rice. *Nat. Commun.*, **7**, 10532.

Meier,L. *et al.* (2008) The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **70**, 53–71.

Miller,A. (2002) *Subset Selection in Regression*. CRC Press, New York. p. 256. https://doi.org/10.1201/9781420035933.

Natarajan,B.K. (1995) Sparse approximate solutions to linear systems. *SIAM J. Comput.*, **24**, 227–234.

Nelson,M. *et al.* (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.*, **11**, 458–470.

Park,T. and Casella,G. (2008) The Bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Pérez,P. and de Los Campos,G. (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, **198**, 483–495.

Phillips,P.C. (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, **9**, 855–867.

Piriyapongsa,J. *et al.* (2012) iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. *BMC Genomics*, **13**, S2.

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

Ritchie,M.D. *et al.* (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.

Segura,V. *et al.* (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.*, **44**, 825–830.

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

Van Hulse,J. *et al.* (2012) Threshold-based feature selection techniques for high-dimensional bioinformatics data. *Network Model. Anal. Health Inf. Bioinf.*, **1**, 47–61.

Visscher,P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.

Wang,Z. *et al.* (2010) A general model for multilocus epistatic interactions in case-control studies. *PLoS One*, **5**, e11384.

Wei,W.-H. *et al.* (2014) Detecting epistasis in human complex traits. *Nat. Rev. Genet.*, **15**, 722–733.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Yilmaz,S. *et al.* (2019) Spadis: an algorithm for selecting predictive and diverse SNPs in GWAS. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **18**, 1208–1216. https://doi.org/10.1109/TCBB.2019.2935437.

Yosef,N. *et al.* (2007) A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data. *Bioinformatics*, **23**, e91–e98.

Zhao,Z. and Liu,H. (2007) Spectral feature selection for supervised and unsupervised learning. In: *Proceedings of the 24th International Conference on Machine Learning*, pp. 1151–1157.