The Role of Different Linkage Disequilibrium Patterns in Genomic Prediction: The gBULP Based Exploratory Method in Tehran Cardiometabolic Genetic Study

Mahdei Akbarzadeh
Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran

Saeid Dehkordi
Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran

Mahmoud Roudbar
Safiabad-Dezful Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education & Extension Organization (AREEO), Dezful

Parisa Riahi
Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran

Mehdi Sargolzaei
Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, Canada, and Select Sires Inc.

Kamran Guity
Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran

Bahareh Sedaghati-khayat
Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran

Hossein Lanjanian
Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran

Fereidoun Azizi
Endocrine research center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran

Maryam Daneshpour (✉ daneshpour@sbmu.ac.ir)
Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran

Research Article

Keywords: TCGS, Genomic Prediction, Lipid Profile Traits, gBLUP, linkage disequilibrium, SNP-based Heritability

Posted Date: December 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-127117/v1

Title:

The role of different linkage disequilibrium patterns in genomic prediction: The gBULP based exploratory method in Tehran cardiometabolic genetic study

Running Title: Improve Genomic Prediction using Mutually Uncorrelated SNPs

Authors' name:

1. Mahdi Akbarzadeh, Ph.D., Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: akbarzadehms@sbmu.ac.ir
2. Saeid Rasekhi Dehkordi, MSc, Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: dehkordi.s.r@gmail.com
3. Mahmoud Amiri Roudbar, Department of Animal Science, Safiabad-Dezful Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education & Extension Organization (AREEO), Dezful, Iran, Email: mahmood.amiri225@gmail.com
4. Parisa Riahi, MSc, Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: parisaariyahii@gmail.com
5. Mehdi Sargolzaei, Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, Canada, and Select Sires Inc., Plain City, USA, Email: msargol@uoguelph.ca
6. Kamran Guity, MSc, Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: kamran.guity@gmail.com
7. Bahareh Sedaghati-khayat, MSc, Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: b.s.khayat@gmail.com
8. Hossein Lanjanian, Ph.D., Endocrine research center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran Email: H.Lanjanian@ut.ac.ir

9. Fereidoun Azizi, MD, Ph.D., Endocrine research center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: azizi@endocrine.ac.ir
10. Maryam S Daneshpour, Ph.D., Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: daneshpour@sbmu.ac.ir

Correspondence Author:

Maryam S Daneshpour, Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran; POBox: 19195-4763, Tel: +982122432500, Fax: +982122416264, Email: daneshpour@sbmu.ac.ir

Abstract:

Background:

Current GWAS discoveries have discovered novel clinical improvements in recent decades, such as estimating whole-genome risk. Genetic prediction of traits has substantial impacts on public health care and disease prevention. This study aimed to investigate the effects of different linkage disequilibrium (LD) patterns on genomic prediction accuracy and SNP-based heritability estimation for four lipid profile traits.

Results:

This family-based study included 11,798 individuals ranging from 3 to 80 ys, extracted from Tehran Cardiometabolic Genetic Study (TCGS). LD patterns were considered on different R^2 thresholds (0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3, 0.5, 0.6, 0.7, 0.8, and 0.9) to create subsets of SNPs. We have compared the prediction accuracy and SNP-based heritability estimation of the selected SNPs within these patterns as well as randomly selected SNPs with equal sizes. Subsets of SNPs selected based on LD patterns had a higher prediction accuracy level than subsets of SNPs selected randomly, and when the LD threshold increases, the difference tends to zero. The results were consistent when the prediction accuracy of subsets were adjusted for their SNP numbers in all traits. For all traits, when the number of SNPs was adjusted, between LD threshold 0.01 and 0.2, both prediction accuracy and SNP-based heritability have a dramatic rise. After substantial growth, there was a steady decline, and they reach a peak at an LD threshold between 0.2 and 0.3.

Conclusions:

This research indicated that having selected subsets of SNPs based on the LD threshold always outperform randomly selected SNPs for prediction objectives. However, determining the specific LD threshold for prediction purposes might be controversial since achieving the highest level of prediction accuracy, when the number of SNPs is adjusted, prompts different results (in our case, 0.3 when the SNP number was adjusted and 0.9 when the SNP number is not adjusted). Finally, we concluded that choosing the LD threshold as a tool to boost genetic prediction accuracy should be used with intense care.

Key Words: TCGS, Genomic Prediction, Lipid Profile Traits, gBLUP, linkage disequilibrium, SNP-based Heritability

1- Introduction:

Most human complex traits and diseases are heritable, which have increased the importance of research in genetic variations inheritance. The remarkable advanced genotyping technique's achievements have brought about the emergence of genomic medicine in prevention, diagnosis, and treatment (1,2). However, due to the low power of genome-wide association analysis (GWAS) in detecting markers with small effect sizes, only a small proportion of phenotypic variance could be explained by the significantly associated SNPs with complex traits and diseases (3,4). Researchers put forth to increase the predictive power of genetic variants to the simultaneous inclusion of hundreds of thousands of markers and using Whole-Genome Regression (WGR) methods (5).

Numerous statistical approaches have been proposed for carrying out WGR. The simplest one, single marker linear regression, commonly referred to as GWAS mostly applied to obtain significant regions on the genome for finding quantitative trait loci (QTL). However, a large number of SNP markers in comparison to the number of individuals give rise to using alternative approaches to control false positive error(6)**.** Considering SNP effects as random effects and making the estimation possible by assuming prior assumptions about the SNP effects' variance is a nonlinear solution like Bayes A, Bayes B (5,7), and Bayes C (8). While Bayes methods assume different emphasis on different loci, the genomic best linear unbiased prediction (gBLUP) gives equal importance to each SNP by assigning the same variance to entire SNPs (9,10). gBLUP relays on a genomic relationship matrix (GRM) for importing genomic similarity of individuals into the model (10–12).

While many studies have tried to increase genetic prediction accuracy by comparing and developing various methods (6,13–16) or even combining methods (17), shrinkage and selection of SNPs can enhance prediction accuracy in complex traits (18). Like the presence of related individuals in training and testing dataset increases the prediction accuracy (19), the relatedness between pairs of SNP markers, linkage disequilibrium might be influential in highly complex human traits (20,21). Lipid traits are risk factors for non-communicable diseases (NCDs) like cardiovascular disease and Type 2 diabetes. Genetic prediction of lipid traits, given their moderate heritability, could

substantially impact public health care and prevention as early as birth due to mostly unchanged genetic variants in an individual (22,23).

This study aimed to investigate the effect of different linkage disequilibrium (LD) patterns on genomic prediction accuracy and SNP-based heritability estimation. We tried to prove this hypothesis by using high-density lipoprotein cholesterol (HDL-C), Low-density lipoprotein cholesterol (LDL-C), triglycerides (TG), and total cholesterol (TC) from Tehran Lipid and Glucose Study (TLGS) project (24,25) along with their genomic information from Cardiometabolic Genetic Study (TCGS) projects (26).

2. Methods and Materials:

2.1 Study setting and participants

The study was carried out in the framework of two broader studies, Tehran Lipid and Glucose Study (TLGS) and Tehran Cardiometabolic Genetic Study (TCGS).

Tehran Lipid and Glucose Study (TLGS):

TLGS is a cohort study that started in 1997 in distinct 13 of Tehran municipality on about 19,905 family-based individuals, with age ranges from 3 to 80, in which subjects have been selected randomly and have been recorded every three years. This study is being carried out on non-communicable diseases (NCD) in the Iranian population for mortality and morbidity. TLGS aimed to determine the tendency of the lipid and glucose status' modifications of the individuals, in which the periodic examinations were done every three years(24).

Tehran Cardiometabolic Genetic Study (TCGS)

More recently, TLGS has started projects that include the genomics and biomarkers of the non-communicable disease, the first whole-genome sequencing study in Iran, and aims to develop personalized medicine

endocrinology. Tehran Cardiometabolic Genetic Study (TCGS), a prospective family-based cohort study, has created a genome-wide sequencing database of about 15,000 individuals of TLGS that was designed in collaboration with the Research Institute for Endocrine Sciences and the genetic company deCODE (26). This collaboration aimed to assess the underlying and complex genetic patterns of cardiovascular disease, as well as the epistasis and gene-environmental interactions by conducting Genome-wide Association Study. The individuals were genotyped with HumanOmniExpress-24-v1-0 bead chips, each with 649,932 SNP loci, and the genotyping rate for these participants was 0.9854.

2.2 Phenotype Measurement

Since the fourth phase of the TLGS (2009-2011) contained the most available information of recorded lipid profile traits measurements of the participants rather than the other phases, it was considered analyzed in the current study. Total Cholesterol (TC) was measured by the enzymatic colorimetric method using cholesterol esterase and cholesterol oxidase. High-Density Lipoprotein cholesterol (HDL-C) was measured after precipitation of the apolipoprotein B containing lipoproteins with phosphotungstic acid, Triglycerides (TG) were measured using glycerol phosphate oxidase. Finally, the LDL-C was measured as LDL-C = TC - HDL - (TG/5). The number of recorded observations for HDL-C was 10301 (45% males and 55% females), and for the LDL-C 10586 (45% males and 55% females), and TC and TG were 10303 (46% males and 54% females). According to the previous studies, we decided to consider age, sex, and body mass index (BMI)(kg/m^2) as covariates of the used models, as the participants' age was measured in years, and the BMI was calculated as the weight (kg) divided by the squared height in meters. Also, the 1% winsorization has been applied on phenotype variables to adjust the right-skewed distribution, in which the outlier values are modified to the specific quantiles, i.e., the 99^{th} percentile.

2.3 Genotyping and Quality control

Genotyping of the TCGS participants was done in the deCODE genetic company, where HumanOmniExpress-24-v1 bead chips used for genotyping the blood samples, which have provided us with 652919 single nucleotide polymorphism loci with an average mean distance of 4 kilobases (26). Pedigree Check and the parentage test were

conducted before the quality control stage in individual and marker levels. We used ped.info command for this regard in SAGE (Statistical Analysis for Genetic Epidemiology) software version 6.4 (27), to find any wrongly-recorded parental information. snp1101 software was applied for conducting the parentage test (28). In this stage, 887 subjects out of 11798 were excluded from the study, and 325 individuals turned to the founder. In our study, data were assessed quality-control by the Plink software (29) in terms of both individuals and the markers. In the first step of the quality control, SNPs and the samples with genotyping missingness rate of more than 0.2, which are considered low-quality SNPs and individuals, were excluded from the study, where 770 SNPs and 11 individuals were removed at this step. Similarly, SNPs and individuals with genotyping missingness rates of more than 0.02 were excluded from the study (17636 SNPs and zero individuals were removed). The second step of the quality control is checking the sex discrepancies from the difference between information related to the individuals' gender and the homozygosity/heterozygosity rate of their X chromosome. However, none of the individuals excluded from the study in this step. 72500 SNPs with a minor allele frequency (MAF) less than 0.05 were removed in the third step. Next, 1125 SNPs were excluded from the analysis due to deviation from Hardy-Weinberg Equilibrium, since their p-value in the Hardy-Weinberg test were below 10^{-6}. In the next step, 317 samples were excluded from our study due to their high/low heterozygosity rate and deviated from ±3×sample heterozygosity mean. Principle component analysis in the SNPRelate package (30) in R software was applied for checking population stratification. However, there were no inconsistencies in population stratification in the last step of our quality control process. After the quality control stage, imputing the missing genotypes was the crucial phase of the preprocessing steps. Beagle 5.1 (version: 18, May20.d20) software was used to impute missing genotypes (31) and finally provided us with 10583 individuals with 560888 genetic markers to commence the analysis. Readers may find a summary of the above procedure in Supplementary Figure 1.

2-4 Statistical method

Model Selection

We have considered a multiple linear regression model, including age, sex, and BMI, as fixed factors for lipid profile traits. The stepwise approach, which is a combination of the forward and backward selection, considered all three above covariates to be included in the predictor model for HDL-C, LDL-C, TC, and log transformation of TG (to control high skewness). Therefore, the phenotype prediction study has been done with SNP markers as random effects and age, sex, and BMI as fixed effects.

Genomic best linear unbiased prediction (gBLUP)

A mixed model was used as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \varepsilon \qquad \text{Equation 1}$$

Where $\boldsymbol{y}$ is defined as the vector of observed phenotypes, y_i, with $i = 1, \ldots, n$ (n = number of subjects), $\boldsymbol{\beta}$ indicates the vector of fixed effects (age, sex, and BMI), X is a design matrix relating the fixed effects to each individual, $u \sim N(0, \boldsymbol{G}\sigma_u^2)$ indicates a vector of SNP effects with a variance of σ_u^2, $\boldsymbol{G}$ is a square $n \times n$ genomic relationship matrix (GRM). $\varepsilon \sim N(0, \boldsymbol{I}\sigma_\varepsilon^2)$ is residual vector σ_ε^2 indicates the variance of residuals. **Z** is a matrix of genotypes which indicates the number of copies of reference allele (coded as 0,1and 2). If we transform the matrix **Z** to its standardized form, noted by **W**, we would have the following equation:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Wu} + \varepsilon \qquad \text{Equation 2}$$

With the variance of

$$\boldsymbol{var}(\boldsymbol{y}) = \boldsymbol{WW'}\boldsymbol{\sigma_u^2} + \boldsymbol{I}\boldsymbol{\sigma_\varepsilon^2}$$

In which **W** is a matrix that its ij^{th}(i^{th}individual and j^{th} SNP) element is $w_{ij} = (z_{ij} - 2p_j)/\sqrt{2p_j(1-p_j)}$, that p_j shows the frequency of j^{th} SNP (j=1, …, k). Regarding our objective, which is the aggregation of SNPs effects on the phenotype, if we define 1×n vector of **g** total genetic effects of the individuals, we have the equation two mathematically equal to:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \varepsilon \qquad \text{Equation 3}$$

With the variance of

$$\boldsymbol{var(y) = G\sigma_g^2 + I\sigma_\varepsilon^2}$$

Note that $\boldsymbol{G = WW'/N}$ can be defined as the GRM between individuals. Based on the estimated GRM from entire SNPs, we can estimate the phenotypic variance explained by all the SNPs (σ_g^2) as well as residual variance (σ_ε^2) by the restricted maximum likelihood (REML) method using GCTA software, which is applying the average information (AI) method to initiate its iterations.

Therefore, we can have the best linear unbiased prediction (BLUP) of the whole SNPs effects for all individuals ($\hat{\boldsymbol{g}}$ in equation 3). Straightforwardly, we can have the estimation of each SNP effect due to equations 2 and 3. In fact, having $\hat{\boldsymbol{g}}$, the BLUP of **u** ($\hat{\boldsymbol{u}}$) can be found with the following equation:

$$\hat{\boldsymbol{u}} = \boldsymbol{W'A^{-1}\hat{g}/N}$$

We know that $\hat{u}_j$ is the coefficient of w_{ij}. So to have an estimation of SNP effect corresponded to z_{ij} it is enough to transform it by $\hat{u}_j^* = \hat{u}_j/\sqrt{2p_j(1-p_j)}$. The BLUP effects that are achieved by GCTA in a dataset can be used to gain genetic value of the individuals for a given phenotype in a matched validation or test set, which means $\hat{g}_{test} = w_{test}\hat{u}$. This feature predicts genetic value or an individual's risk to disease (polygenic risk score) in complex traits by using the PLINK version 1.9 scoring approach in a test dataset (29).

GRM calculation

Among various approaches that calculate GRM, in this study, we applied the method presented by Yang(12). Genomic similarities between i^{th} and i'^{th} individuals with entire SNPs can be defined as below. In the following formula $A_{ii'}$indicates the similarity between i^{th} and i'^{th} individuals in the j^{th} SNP, so with summation on j we can capture the entire genomic resemblance between every two cases. Thus, when $i \neq i'$:

$$G_{ii'} = \frac{1}{k}\sum_{j=1}^{k} G_{jii'} = \frac{1}{k}\sum_{j=1}^{k} \frac{(x_{ij} - 2p_j)(x_{i'j} - 2p_j)}{2p_j(1-p_j)}$$

Similarly, when $i = i'$:

$$G_{jk} = \frac{1}{k}\sum_{j=1}^{k} G_{jii'} = 1 + \frac{1}{k}\sum_{j=1}^{k} \frac{x_{ij}^2 - (1-2p_j)x_{ij} + 2p_j^2}{2p_j(1-p_j)}$$

Where x_{ij} indicates the observed genotype of j^{th} SNP for i^{th} cases (coded as 0, 1, and 2 according to the number of copies of reference allele), and p_j is the frequency of j^{th} SNP.

2-5 Subset selection

The subset of SNPs was conducted based on their pairwise LD. We determined SNPs' allocation to subsets with different scenarios based on the correlation coefficient between loci pairs. The correlation of two biallelic loci with alleles A and in the first locus and alleles B and b in the second locos with the frequency calculated as follows

$$r^2 = \frac{D^2}{\pi_A \pi_B \pi_a \pi_b}$$

That is the deviation of the expected frequency of one haplotype from its observed value. Therefore, we created subsets of pruned SNPs with threshold (0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3, 0.5, 0.6, 0.7, 0.8, and 0.9) and the window size of 50 SNPs and 5-shifted forward in each step via plink. The number of SNPs in each subset is 17349, 27215, 35704, 44783, 54131, 58807, 106683, 153804, 247642, 294416, 341081, 387071, and 435798 SNPs, respectively. Then the effect of LD between markers on the predictive ability of gBLUP has been investigated within these subsets.

2-6 Estimation of accuracy

In this study, the 10-fold 10-repeat cross-validation (CV) was the evaluation mechanism for each approach's prediction performance. In this way, individuals were randomly divided into ten parts, the collection of 9 parts was utilized as a discovery set, and the one left part was used for validation. This process continued until each part was considered as the validation set. The SNPs' effect sizes, which were estimated based on the discovery set, were used to calculate individuals' whole-genome risk prediction in the validation set, which was not involved in estimating SNPs' effect sizes. The entire process was repeated ten times to reduce the variance of prediction accuracy. The evaluation was based on the correlation between genetic values and adjusted phenotypes (sex, age, and BMI). The average CV-correlation is the index to compare the performance of different subset selection strategies and the model with entire SNPs included. Besides, we have randomly selected an equal number of SNPs to form subsets to evaluate the corresponding selected subsets' performance. Finally, since the LD thresholds have a significant linear correlation with the number of SNPs in each subset, a proportion of the prediction accuracy variations are due to

the increase in the number of SNPs. So, we adjusted the effect of the number of SNPs to realize how much prediction accuracy variation do LD patterns account for.

3. Results:

Demographic information

Supplementary Table 1 describes the participants' lipid profile measurements and the mean age, BMI, and lipid profile traits by gender. HDL-C , TG, and TC were significantly different between the male and female participants (T-Test P-value < 0.0001), but there was not any significant difference in LDL-C (T-Test P-value = 0.26) (see Supplementary Figure 1). Also, all of the lipid profile traits were significantly different between the underweight (BMI <18.5), normal (18.5 < BMI < 24.5), overweight (25 < BMI < 29.5), and obese (BMI > 30) subjects (ANOVA P-Value < 0.0001) (see Supplementary Figure 2). Similarly, lipid profile traits were significantly different (ANOVA P-value < 0.0001) between the children (age < 10), teenagers (11< age < 18), and adults (age > 18) (see Supplementary Figure 3).

3.2 Prediction Accuracy

Prediction accuracy for each pattern and trait was visualized in Figure 1. And the trajectory of changes for the prediction accuracy adjusted for the number of SNPs was depicted in Figure 2. Here, the accuracy means the correlation between the genetic values and the adjusted phenotypes of each of the lipid profile traits for both random and each of the selection patterns (LD thresholds = 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9).

As is shown in Figure 1, surprisingly, the LD-based selected SNP groups' prediction accuracy is always more than randomly selected SNPs in all traits but is slightly lower when the LD threshold is between 0.1 and 0.9. Generally, when the LD is increasing, the two curves are meeting each other more closely. The highest accuracy is obtained when the selection LD threshold is 0.9 (HDL-C =0.32, LDL-C= 0.26, TG =0.29, and TC=0.26 P-value=10^{-7}) and the lowest prediction accuracy is achieved when the selection LD threshold is 0.01 for all of the traits (HDL-C =0.26, LDL-C= 0.21, TG = 0.23, and TC=0.22 P-value=10^{-7}).

Since we considered the LD thresholds cumulatively, a proportion of the variations in prediction accuracy are due to the increase in the number of SNPs. Therefore, we adjusted the effect of the number of SNPs to realize how much prediction accuracy variation do LD patterns account for. Figure 2 illustrates that prediction accuracy adjusted for the number of SNPs in each selected subset based on LD thresholds is upper than randomly selected ones at all LD thresholds. As the curves show, there has been an overall rise in the adjusted accuracy as the LD threshold grows. For all traits, between LD threshold 0.01 and 0.2, the adjusted accuracy has a dramatic rise. After substantial growth, there was a steady decline and depend on the traits, the adjusted accuracy reaches a peak at an LD threshold between 0.2 and 0.3. These results show that having selected subsets of SNPs based on the LD threshold is always better than randomly selected SNPs for prediction objectives. The LD threshold for achieving the most adjusted accuracy of prediction is not just the highest threshold picked.

3.3 Heritability

SNP-based heritability and SNP-based heritability adjusted for the number of SNPs are shown in Figure 3 and Figure 4, respectively. Also, genetic variances and heritability of each trait according to the selection LD threshold and their p-values are shown in the supplementary Table (2-5).

As is shown, in general, as the selection LD threshold increases, genetic variance tends to increase. In participants with HDL-C, the lowest level of heritability belongs to the LD threshold of 0.01 ($h^2_{HDL-C} = 0.30, SE = 0.01, \mathrm{P-value} < 0.0001$) and the highest measure of heritability belongs to the 0.7 ($h^2_{HDL-C} = 0.51, SE = 0.02, \mathrm{P-value} < 0.0001$). Also, the selection LD thresholds of 0.6, 0.8, and 0.9 have a relatively high heritability measure. For adjusted SNP-based heritability, as Figure 4 shows, something similar to adjusted prediction accuracy happened here. For all traits, between LD threshold 0.01 and 0.2, the adjusted SNP-based heritability rose dramatically. Then, it declined steadily, and the peak of adjusted SNP-based is at an LD threshold between 0.2 and 0.3.

4. Discussion

This study was conducted to investigate the effect of the LD pattern between SNP markers on the genomic prediction ability of the gBLUP model, and our results revealed that increasing LD between markers, including more SNPs remaining for prediction, increased the predictive accuracy of gBLUP only up to 6% per trait.

Whole-genome predictors perfectly learn the information behind entire genetic variants with no demand for pre-selection SNPs based on their association ability with phenotype, which is regularly done with univariate models (12). This ability becomes brighter in highly human complex traits where the trait architecture is unknown, and many SNPs with small effect sizes are influential. The predictive power of whole-genome predictors, gBLUP in this study, is influenced by many factors; the relatedness of individuals and alleles' nonindependence at different sites (LD). Previous studies showed that LD plays a critical role in genomic prediction (5). Strong LD led to higher accuracy of genomic prediction when gBLUP was performed using chip data (20,32).

The effect of LD between markers on the predictive ability of gBLUP has been investigated with different R^2cut-points of LD (0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3, 0.5, 0.6, 0.7, 0.8, and 0.9). Our results revealed that increasing LD between markers, including more SNPs for prediction, increased the predictive accuracy of gBLUP up to 6% per trait. A possible reason is that higher LD resulted in including causal mutation and a larger number of variants incorporated in the prediction process, which would increase the prediction power in human highly complex traits(33). Including stronger LD between SNPs can capture the LD between markers and Quantitative Trait Loci (QTL), and more importantly, they can capture the familial relationship between samples as well (9). Since attending high LD between SNPs remained familial in training and testing samples, prediction in subsets with stronger LD improved, which agrees with previous studies (19,34). Long-ranged LD can enhance the possibility of capturing markers adjacent to causative mutations, which increases the accuracy of prediction.

Lipid traits are known as moderately heritable traits, as we estimated were 0.495, 0.388, 0.390, and 0.431 for HDL-C, LDL-C, TC, and TG, respectively (with adjustment on sex, age, and BMI). As previously through simulation analysis have been shown, approximately the total genotypic variation of phenotype can be capture by genetic variants

when the perfect linkage between QTL and markers exist (18). However, it is virtually impossible in real data of highly human complex traits due to new traits architecture and low LD in the human population, even in related studies (18,35). Therefore, as we concluded, including SNPs with high LD can increase genomic data information to capture the phenotype's genotypic variation.

The increasing genotypic variance of lipid traits can be seen in our result by including stronger LD between SNPs, consistent in HDL-C, LDL-C, TC, and TG. This increase resulted in increasing the gene-based heritability. Long-rage LD within genetic variants in highly complex traits like lipid traits that many markers influence phenotype with moderately small effects give rise to including more markers adjacent to causative loci. As a result, SNP markers can capture the phenotype's most possible variation explained by genetic factors.

Conclusion:

The genetic architecture of traits is based on genome prediction, which determines the accuracy of prediction and genetic variation that explains the phenotype's variance. The genomic prediction accuracy would deteriorate as the trait's complexity increases, so shrinkage and variable selection approaches perform poorly in this regard. LD-based marker pruning is not always a good idea in human complex traits to estimate individuals' genetic values. This research indicated that having selected subsets of SNPs based on the LD threshold is always better than randomly selected SNPs for prediction objectives. However, the LD threshold we may use for achieving the most adjusted genetic prediction accuracy is not just the best threshold. We have proved this theory on the Iranian population's genomic prediction using the gBLUP model on HDL-C, LDL-C, TC, and TG, among highly human complex traits. Finally, we concluded and recommended that choosing the LD threshold as a tool for prediction purposes should be used with intense care to boost genetic prediction accuracy.

Acknowledgments

The authors would like to express their gratitude to the staff and participants in the TCGS project. Also, special thanks for the scientific and financial support of the deCODE genetic company (Reykjavik, Iceland). We would also like to express our special thanks of gratitude to Asieh Zahedi, Sajedeh Masjoodi, and Atefeh Seyed Hamzehzadeh for doing quality controls for TCGS phenotypes. The present study was funded by the RIES, Shahid Beheshti University of Medical Sciences (Tehran, Iran), and recognizes the scientific support of deCODE (Reykjavik, Iceland).

Author information

1. **Cellular, and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran**

   Mahdi Akbarzadeh & Saeid Rasekhi Dehkordi & Parisa Riahi & Kamran Guity & Bahareh Sedaghati-Khayat & Hoeein Lanjanian & Maryam S Daneshpour

2. **Endocrine research center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran**
   Fereidoun Azizi

3. **Department of Animal Science, Safiabad-Dezful Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education & Extension Organization (AREEO), Dezful, Iran**
   Mahmoud Amiri Roudbar

4. **Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, Canada, and Select Sires Inc., Plain City, USA**
   Mehdi Sargolzaei

Contributions

Mahdi Akbarzadeh: Conceptualization, Writing - Original Draft, Writing- Reviewing and Editing, Software. **Saeid Rasekhi Dehkordi:** Software, Formal analysis, Writing - Original Draft. **Mahmoud Amiri Roudbar:** Validation, Reviewing. **Parisa Riahi:** Visualization. **Mehdi Sargolzaei:** Validation. **Kamran Guity:** Resources. **Bahareh Sedaghati-khayat:** Resources. **Hossein Lanjanian:** Investigation. **Fereidoun Azizi:** Supervision. **Maryam S Daneshpour:** Supervision, Writing- Reviewing, and Editing. All authors read and approved the final manuscript.

Corresponding author

Correspondence to **Maryam S Daneshpour.**

Ethics approval and consent to participate

The local ethics committee approved this study at Research Institute for Endocrine Sciences; Shahid Beheshti University of Medical Sciences (Research Approval Code:*98104* & Research Ethical Code: *IR.SBMU.Endocrine.REC.1398.121*). In this study, all participants provided written informed consent for participating in the study. The research has been performed in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

All parts of this research work, design of the study, data collection, analysis, interpretation of data, and writing the manuscript, were funded by the Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. The funding body played no role in publication costs.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Reference:

1. Guttmacher AE, Collins FS. Genomic Medicine — A Primer. Guttmacher AE, Collins FS, editors. N Engl J Med [Internet]. 2002 Nov;347(19):1512–20. Available from: http://www.nejm.org/doi/10.1056/NEJMra012240

2. Guttmacher AE, McGuire AL, Ponder B, Stefánsson K. Personalized genomic information: Preparing for the future of genetic medicine [Internet]. Vol. 11, Nature Reviews Genetics. Nature Publishing Group; 2010. p. 161–5. Available from: https://www.nature.com/articles/nrg2735

3. Maher B. Personal genomes: The case of the missing heritability [Internet]. Vol. 456, Nature. Nature Publishing Group; 2008. p. 18–21. Available from: https://pubmed.ncbi.nlm.nih.gov/18987709/

4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases [Internet]. Vol. 461, Nature. Nature Publishing Group; 2009. p. 747–53. Available from: www.1000genomes.org/page.php

5. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics [Internet]. 2001 Apr;157(4):1819–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11290733

6. Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, et al. Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. Cai X, editor. PLoS One [Internet]. 2014 Mar;9(3):e93017. Available from: https://dx.plos.org/10.1371/journal.pone.0093017

7. Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G, Fernando RL, et al. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Dec; Available from: http://www.ncbi.nlm.nih.gov/pubmed/18073436

8. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics [Internet]. 2011 Dec;12(1):186. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21605355

9. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics [Internet]. 2007 Dec;177(4):2389–97. Available from: https://www.genetics.org/content/177/4/2389

10. VanRaden PM. Efficient Methods to Compute Genomic Predictions. J Dairy Sci [Internet]. 2008 Nov;91(11):4414–23. Available from: https://www.sciencedirect.com/science/article/pii/S0022030208709901

11. Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet [Internet]. 2011 Dec;128(6):409–21. Available from: http://doi.wiley.com/10.1111/j.1439-0388.2011.00964.x

12. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet [Internet]. 2010 Jul;42(7):565–9. Available from: http://www.nature.com/articles/ng.608

13. Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, Schön CC. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics [Internet]. 2013 Oct;195(2):573–87. Available from: https://www.genetics.org/content/195/2/573

14. Verbyla KL, Bowman PJ, Hayes BJ, Goddard ME. Sensitivity of genomic selection to using different prior distributions. BMC Proc [Internet]. 2010 Dec;4(S1):S5. Available from: /pmc/articles/PMC2857847/?report=abstract

15. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci [Internet]. 2009;92(2):433–43. Available from: http://www.sciencedirect.com/science/article/pii/S0022030209703479

16. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited Review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci [Internet]. 2009;92(1):16–24. Available from: http://www.sciencedirect.com/science/article/pii/S0022030209703054

17. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet [Internet]. 2013 Feb;9(2):1003264. Available from: www.plosgenetics.org

18. Berger S, Pérez-Rodríguez P, Veturi Y, Simianer H, de los Campos G. Effectiveness of Shrinkage and Variable Selection Methods for the Prediction of Complex Human Traits using Data from Distantly Related Individuals. Ann Hum Genet [Internet]. 2015 Mar;79(2):122–35. Available from: http://doi.wiley.com/10.1111/ahg.12099

19. Spiliopoulou A, Nagy R, Bermingham ML, Huffman JE, Hayward C, Vitart V, et al. Genomic prediction of complex human traits: Relatedness, trait architecture and predictive meta-models. Hum Mol Genet [Internet]. 2015 Jul;24(14):4167–82. Available from: https://europepmc.org/articles/PMC4476450

20. Calus MPL, Bouwman AC, Schrooten C, Veerkamp RF. Efficient genomic prediction based on

whole-genome sequence data using split-and-merge Bayesian variable selection. Genet Sel Evol [Internet]. 2016 Jun;48(1):49. Available from: http://gsejournal.biomedcentral.com/articles/10.1186/s12711-016-0225-x

21. Amiri Roudbar M, Mohammadabadi MR, Ayatollahi Mehrgardi A, Abdollahi-Arpanahi R, Momen M, Morota G, et al. Integration of single nucleotide variants and whole-genome DNA methylation profiles for classification of rheumatoid arthritis cases from controls. Heredity (Edinb) [Internet]. 2020 May;124(5):658–74. Available from: https://www.nature.com/articles/s41437-020-0301-4

22. Musunuru K, Kathiresan S. Surprises from Genetic Analyses of Lipid Risk Factors for Atherosclerosis. Circ Res [Internet]. 2016 Feb;118(4):579–85. Available from: /pmc/articles/PMC4762058/?report=abstract

23. Stitziel NO. Human genetic insights into lipoproteins and risk of cardiometabolic disease [Internet]. Vol. 28, Current Opinion in Lipidology. Lippincott Williams and Wilkins; 2017. p. 113–9. Available from: /pmc/articles/PMC5584563/?report=abstract

24. Azizi F, Ghanbarian A, Momenan AA, Hadaegh F, Mirmiran P, Hedayati M, et al. Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. Trials. 2009 Jan;10:5.

25. Azizi F, Rahmani M, Emami H, Mirmiran P, Hajipour R, Madjid M, et al. Cardiovascular risk factors in an Iranian urban population: Tehran lipid and glucose study (phase 1). Sozial-und präventivmedizin. 2002;47(6):408–26.

26. Daneshpour MS, Fallah M-S, Sedaghati-Khayat B, Guity K, Khalili D, Hedayati M, et al. Rationale and Design of a Genetic Study on Cardiometabolic Risk Factors: Protocol for the Tehran

Cardiometabolic Genetic Study (TCGS). JMIR Res Protoc [Internet]. 2017 Feb;6(2):e28. Available from: https://pubmed.ncbi.nlm.nih.gov/28232301/

27. Elston RC, Gray-McGuire C. A review of the "Statistical Analysis for Genetic Epidemiology" (S.A.G.E.) software package. Hum Genomics [Internet]. 2004;1(6):456–9. Available from: /pmc/articles/PMC3500199/?report=abstract

28. M Sargolzaei. SNP1101 User's guide. Version 1.0. HiggsGene Solut Inc. 2014;

29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

30. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics [Internet]. 2012 Dec;28(24):3326–8. Available from: /pmc/articles/PMC3519454/?report=abstract

31. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet [Internet]. 2007 Nov;81(5):1084–97. Available from: www.ajhg.org

32. Ye S, Gao N, Zheng R, Chen Z, Teng J, Yuan X, et al. Strategies for Obtaining and Pruning Imputed Whole-Genome Sequence Data for Genomic Prediction. Front Genet [Internet]. 2019 Jul;10(JUN):673. Available from: https://www.frontiersin.org/article/10.3389/fgene.2019.00673/full

33. de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of Complex

Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genet [Internet]. 2013 Jul;9(7):e1003608. Available from: www.plosgenetics.org

34. Wientjes YCJ, Veerkamp RF, Calus MPL. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics [Internet]. 2013;193(2):621–31. Available from: /pmc/articles/PMC3567749/?report=abstract

35. Goddard M. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica [Internet]. 2009 Jun;136(2):245–57. Available from: https://pubmed.ncbi.nlm.nih.gov/18704696/

Figures

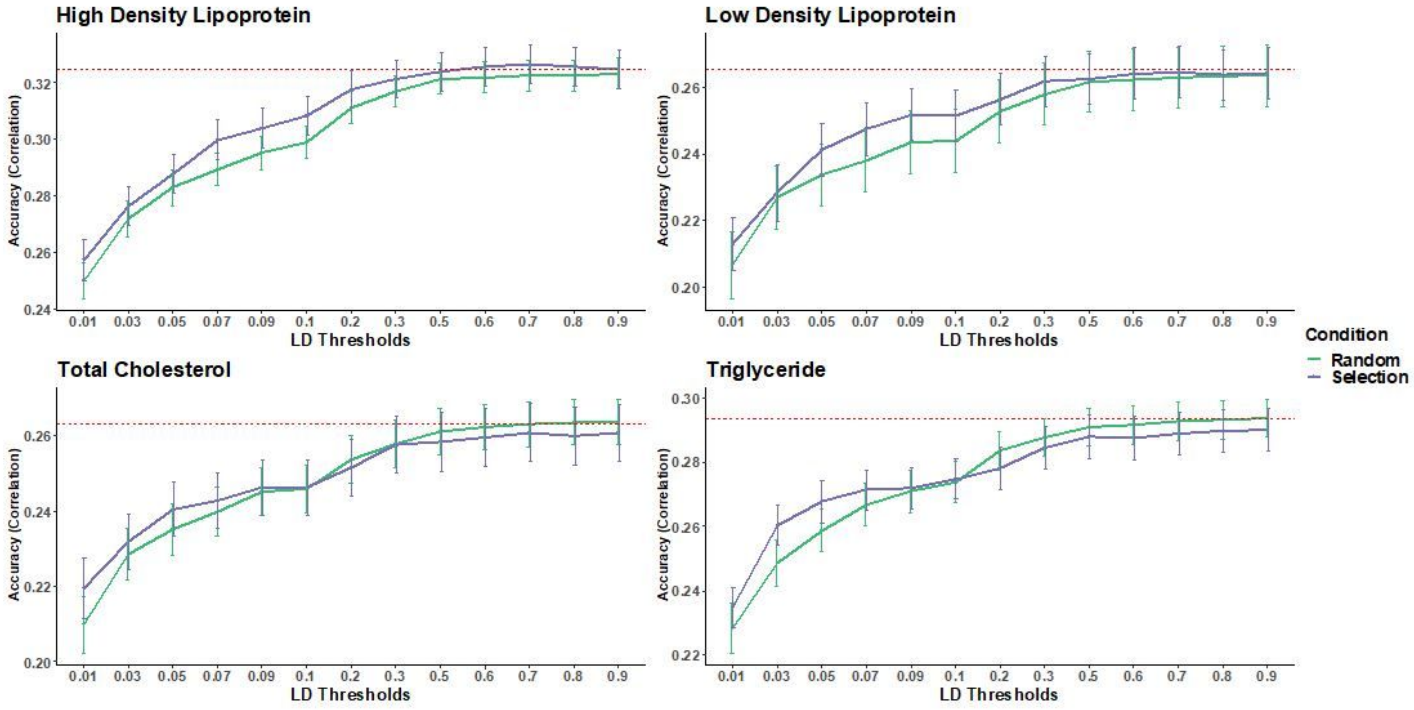

Figure 1

Prediction Accuracy Correlation between genetic values and lipid profile traits is considered as prediction accuracy. The dashed red line indicates the prediction accuracy when all of the SNPs are included in the model. The blue curve indicates the selected SNPs group's accuracy based on an LD threshold, and the green curve indicates the accuracy of the randomly selected SNPs. Up to the point 0.1, the accuracy of selected SNPs based LD thresholds is higher than random SNPs selected for all traits. After point 0.1, with increasing the LD threshold, the difference decreases, and from somewhere onwards, this difference becomes even negative for each trait. In such a way that the accuracy of random SNPs selected will be higher than selected SNPs based LD thresholds.

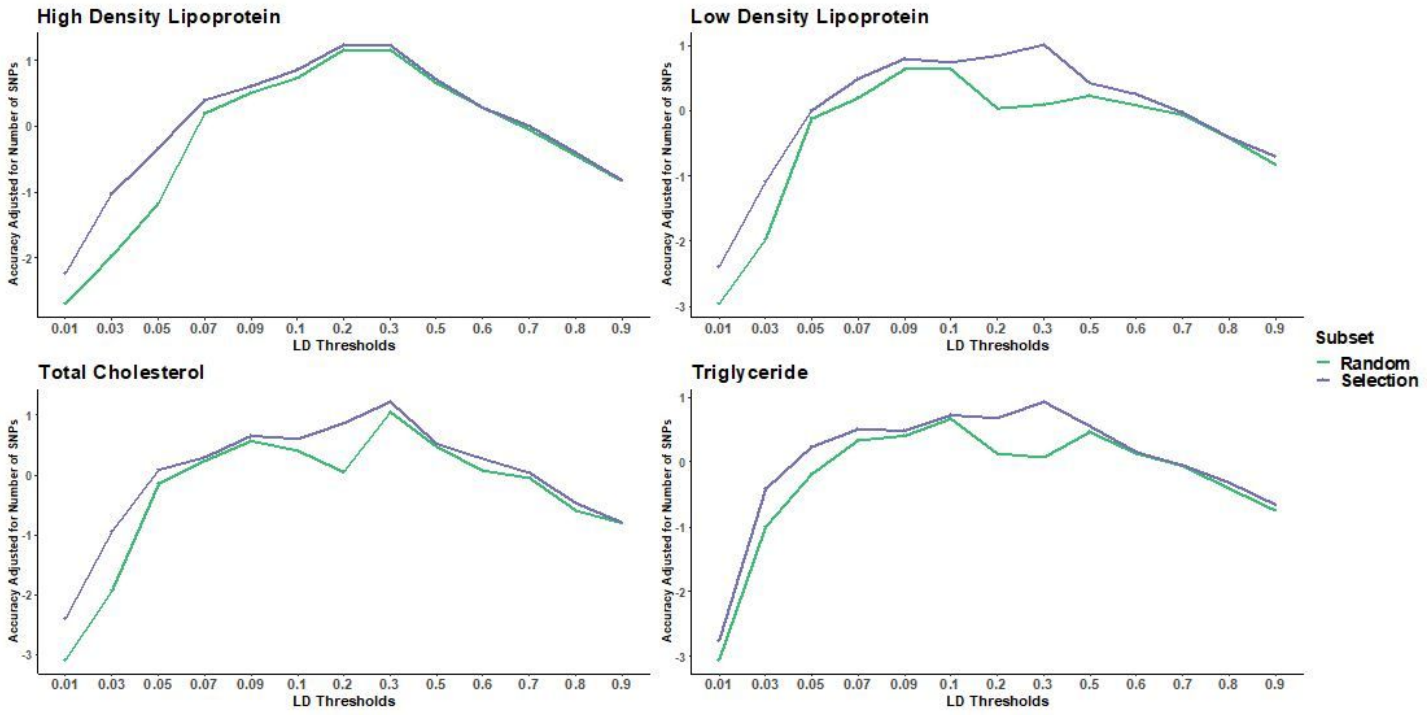

Figure 2

Prediction Accuracy adjusted for the number of SNPs for each pattern There has been an overall rise in the adjusted accuracy as the LD threshold increases. For all traits, between LD threshold 0.01 and 0.2, the adjusted accuracy has a dramatic rise. After substantial growth, there was a steady decrease, and depended on the traits; the adjusted accuracy prevailed consistently at an LD threshold between 0.2 and 0.3.

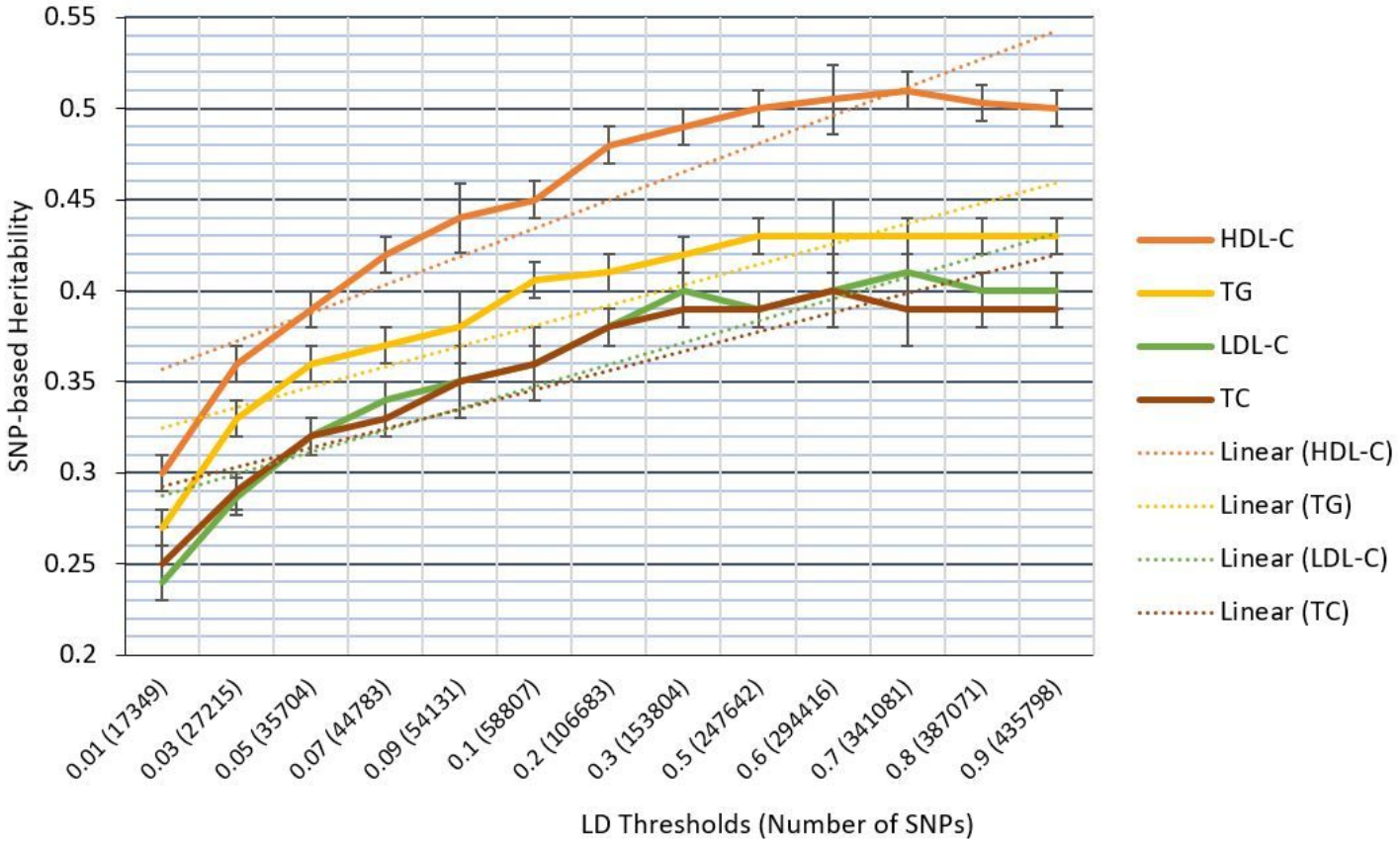

Figure 3

SNP-based heritability estimation by linkage disequilibrium patterns with a linear trendline. The solid-colored lines show the SNP-heritability for each LD threshold. The dashed-colored line indicates a linear trendline for each trait. Total SNP-based heritability for HDL-C is 0.495 (0.02), for LDL-C is 0.38 (0.019), for TG is 0.43 (0.02), and for TC is 0.39 (0.02).

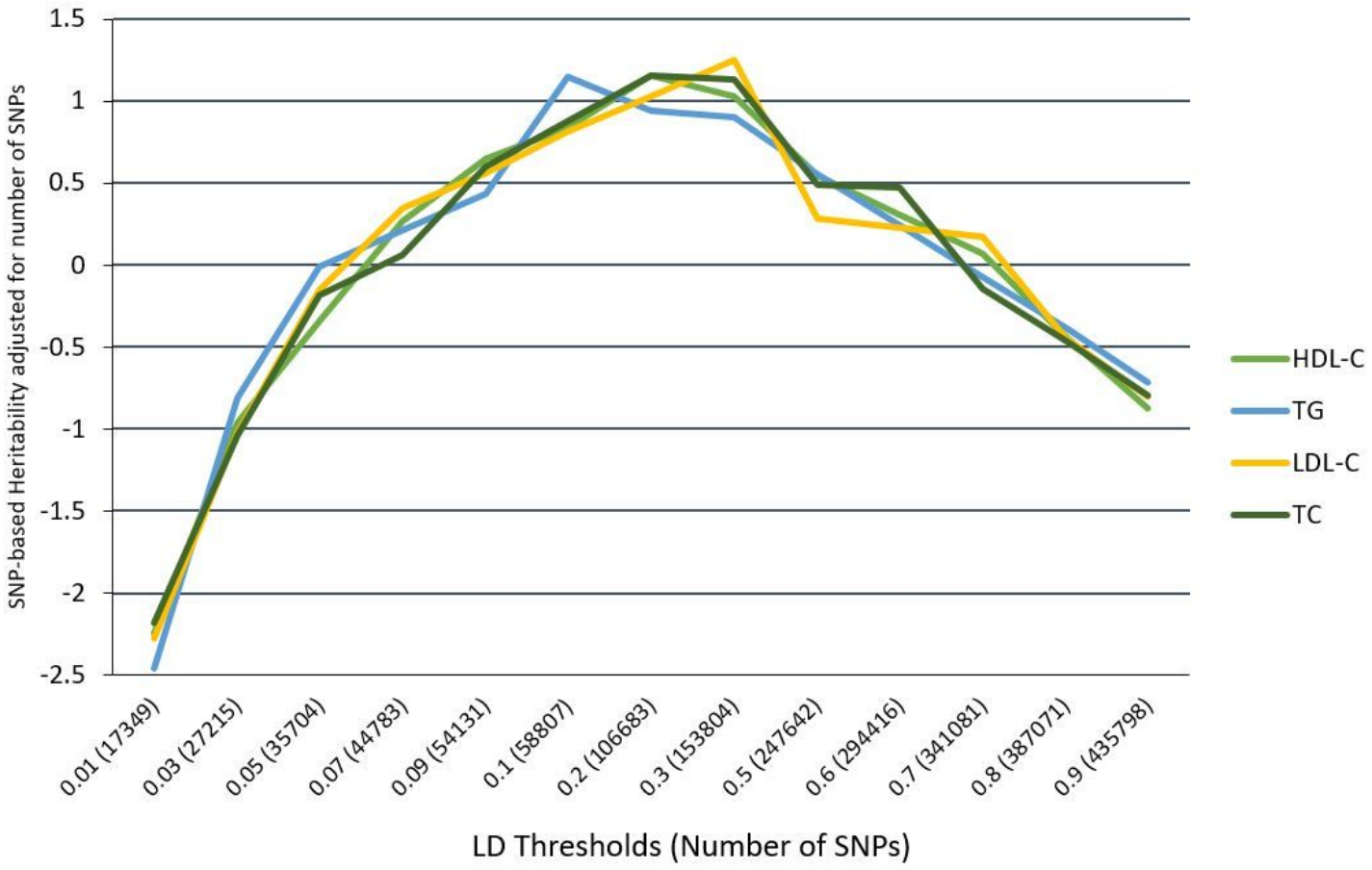

Figure 4

SNP-based heritability estimation by linkage disequilibrium patterns adjusted for the number of SNPs for each pattern The adjusted SNP-based heritability rose dramatically between LD threshold 0.01 and 0.2. The peak of adjusted SNP-based is at an LD threshold between 0.2 and 0.3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfiguresandtables.docx