BMC Genomics

RESEARCH

Open Access

Improving genetic variant identification for quantitative traits using ensemble learning-based approaches

Jyoti Sharma[1†], Vaishnavi Jangale[1†], Rajveer Singh Shekhawat[1] and Pankaj Yadav[1,2*]

Abstract

Background Genome-wide association studies (GWAS) are rapidly advancing due to the improved resolution and completeness provided by Telomere-to-Telomere (T2T) and pangenome assemblies. While recent advancements in GWAS methods have primarily focused on identifying genetic variants associated with discrete phenotypes, approaches for quantitative traits (QTs) remain underdeveloped. This has often led to significant variants being overlooked due to biases from genotype multicollinearity and strict *p*-value thresholds.

Results We propose an enhanced ensemble learning approach for QT analysis that integrates regularized variant selection with machine learning-based association methods, validated through comprehensive biological enrichment analysis. We benchmarked four widely recognized single nucleotide polymorphism (SNP) feature selection methods-least absolute shrinkage and selection operator, ridge regression, elastic-net, and mutual information-alongside four association methods: linear regression, random forest, support vector regression (SVR), and XGBoost. Our approach is evaluated on simulated datasets and validated using a subset of the PennCATH real dataset, including imputed versions, focusing on low-density lipoprotein (LDL)-cholesterol levels as a QT. The combination of elastic-net with SVR outperformed other methods across all datasets. Functional annotation of top 100 SNPs identified through this superior ensemble method revealed their expression in tissues involved in LDL cholesterol regulation. We also confirmed the involvement of six known genes (APOB, TRAPPC9, RAB2A, CCL24, FCHO2, and EEPD1) in cholesterol-related pathways and identified potential drug targets, including APOB, PTK2B, and PTPN12.

Conclusions In conclusion, our ensemble learning approach effectively identifies variants associated with QTs, and we expect its performance to improve further with the integration of T2T and pangenome references in future GWAS.

Keywords Genome-wide association studies, Machine learning, Feature selection, Elastic-net, Support vector regression, Functional enrichment

Background

Genome-wide association studies (GWAS) have transformed the field of genetics by enabling researchers to identify genetic variants associated with complex traits and diseases on a genome-wide scale. After the completion of the Human Genome Project (HGP) [1], GWAS has been recognised as the most effective approach for identifying variants associated with phenotypes of interest. The popularity of GWAS is expected to surge in the

† Jyoti Sharma and Vaishnavi Jangale contributed equally to this work.

*Correspondence:
Pankaj Yadav
pyadav@iitj.ac.in
[1] Department of Bioscience & Bioengineering, Indian Institute of Technology, Jodhpur 342030, Rajasthan, India
[2] School of Artificial Intelligence and Data Science, Indian Institute of Technology, Jodhpur 342030, Rajasthan, India

near future owing to the availability of newer high-quality gap-less reference genomes like Telomere-to-Telomere (T2T)-CHM13 and human pangenome [2, 3]. Over the past two decades, GWAS has identified genetic risk loci such as FTO for obesity [4] and PTPN22 for autoimmune diseases [5]. Additionally, GWAS also uncovered pathways such as the IL-12/IL-23 pathway linked to Crohn's disease [6], which encouraged clinical studies for medicines targeting the pathways.

Despite the widespread use and success of GWAS in identifying numerous disease-associated variants, conventional GWAS methods suffer from many challenges in analysing quantitative traits (QTs). The stringent p-value criteria used in GWAS may overlook variants with low or moderate effect sizes, potentially leading to false negatives [7] and hindering the detection of variants with modest yet biologically significant associations [8, 9]. Furthermore, conventional linear and logistic regression models used in GWAS [10] often fail to consider epistatic interactions between genetic variants [8]. Consequently, the estimated heritability derived from such GWAS analysis may not accurately reflect the true genetic component underlying complex traits/diseases, giving rise to "missing heritability" [11]. Missing heritability poses a significant challenge in GWAS analysis, underscoring the need for alternative approaches to enhance the detection of genetic associations. Moreover, the reliability of conventional GWAS results is often questioned due to the lack of functional annotations, making it challenging to interpret the biological significance of identified variants. Without this contextual information, it becomes challenging to discern whether a detected association is causative or merely a marker in linkage disequilibrium (LD) with the true functional variant.

To address these issues of GWAS, researchers have employed machine learning (ML) methods, such as decision tree-based and penalized regression-based approaches, for association analyses. Extreme Gradient Boosting (XGBoost), based on gradient-boosted decision trees, effectively incorporates pairwise epistatic interactions of single nucleotide polymorphisms (SNPs) within one tree. A notable feature of XGBoost is its capacity to restrict interactions between SNPs within single trees, facilitating the study of SNP interactions and enabling prediction models to include complex non-linear interactions in a non-additive form [12]. In contrast to decision tree approaches, penalized regression-based methods, such as ridge regression and least absolute shrinkage and selection operator (LASSO), are comparatively less complex ML methods. These methods simultaneously select variants and estimate their effects on phenotype by imposing constraints on model coefficients [13]. Ridge regularization stabilizes parameter estimation by shrinking predictors, while LASSO regularization facilitates variant selection by driving many regression coefficients to zero [14]. Elastic-net regularization, another ML approach, combines ridge and LASSO penalties to provide shrinkage and automated variant selection. These ML methods are particularly beneficial for GWAS analysis, where epistatic interactions and multicollinearity among nearby SNPs are prevalent due to LD.

However, ML methods encounter challenges due to the high dimensionality of GWAS data, characterized by a large number of SNPs relative to sample size, also known as the "curse of dimensionality" [15]. To address this issue, pre-selection of SNPs using various feature selection methods has been explored [16], albeit with limited success, particularly for complex continuous traits like low-density lipoprotein (LDL) cholesterol levels. To overcome these challenges and enhance our understanding of complex traits and diseases, innovative approaches integrating advanced statistical methodologies, association methods, functional annotations, and biological insights are essential.

Here, we employ elastic-net regularization as a feature selection approach, leveraging its ability to address computational challenges associated with high-dimensional data and mitigate spurious associations [17]. Other widely used feature selection methods such as LASSO, ridge, and mutual information are also implemented to compare the performance of elastic-net method. Additionally, we utilize linear regression (LR), random forest (RF), XGBoost, and support vector regression (SVR) ML methods for association testing on selected features/SNPs. Kernel-based ML methods like SVR for association testing can identify interactions by exploring all possible combinations of SNPs within a GWAS dataset. The effectiveness of our ensemble approach is validated on simulated, real and imputed datasets. We perform functional enrichment analysis of identified associated SNPs across diverse biological processes to confirm their biological relevance, ensuring the validity of our findings. The step-wise illustration of the proposed framework is displayed in Fig. 1.

Methods

Our proposed framework includes quality control (QC) of the data, four feature selection methods such as LASSO [14], ridge [18], elastic-net [19], and mutual information [20], followed by three association methods (i.e. LR [21], RF [22], and SVR [23]).

Let $X, Y = (X, Y) \in \mathbb{R}^{n \times snps}$ is a GWAS dataset of a complex trait for n individual and $snps$ number of SNPs. Let $X = (x_{ij})_{i=1,j=1}^{n,snps}$, where x_{ij} is the allele of i^{th} individual at j^{th} SNP and $Y = (y_i)_{i=1}^{n}$ indicates the $n \times 1$ quantitative phenotype, where y_i is phenotype value of i^{th} individual.

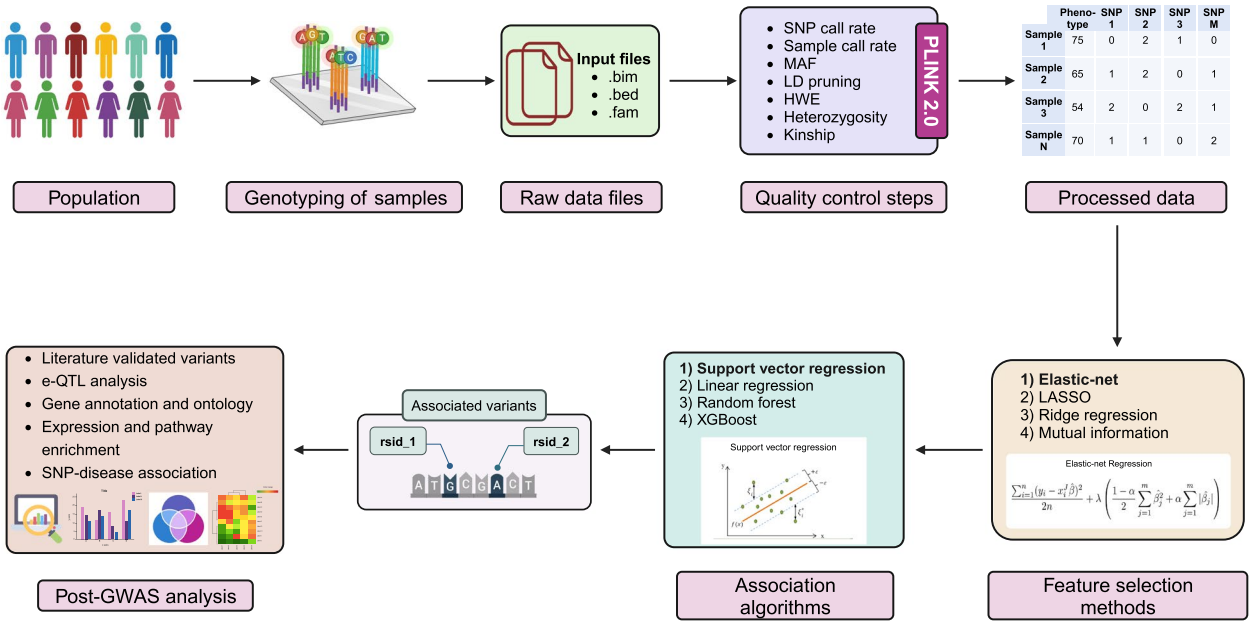

Fig. 1 The stepwise workflow of the proposed framework, highlighting the optimal combination of ML methods (shown in **bold**) for the identification of trait-associated variants, followed by validation through post-GWAS analysis. *Abbreviations*- GWAS: genome-wide association studies; SNP: single nucleotide polymorphism; MAF: minor allele frequency; LD: linkage disequilibrium; HWE: Hardy-Weinberg equilibrium; ML: machine learning; LASSO: least absolute shrinkage and selection operator; SVR: support vector regression; XGBoost: Extreme Gradient Boosting; e-QTL: Expression Quantitative Trait Locus

QC of dataset

The efficiency of GWAS to identify true genetic connections is dependent on the overall quality of the dataset. Even simple statistical tests of association are hampered by unprocessed genome-wide SNP data, potentially leading to false-negative and false-positive associations. Furthermore, concerns with genotype data quality will most likely affect subsequent analyses and studies beyond the initial GWAS [24]. The QC steps of all datasets were performed by Plink2.0 [25]. SNP call rate serves as the initial step in QC. This process entails filtering out SNPs that exhibit high rates of missing data across individuals, with a threshold set at 100%. Subsequently, the sample call rate filter involves exclusion of samples with substantial amounts of missing data across SNPs, employing a threshold of 95%. The minor allele frequency (MAF) is then assessed, representing the frequency of the least common allele within a particular population, with a threshold of 0.01. LD analysis examines the random association of different genetic loci within the same population, employing a threshold of 0.3. Hardy-Weinberg equilibrium (HWE) evaluates the relation between allele and genotype frequencies, with SNPs having a p-value $< 1 \times 10^{-5}$ considered as outliers and subsequently excluded from the dataset. Given the population-based nature of our study, individuals such as twins, first cousins, and other family members are excluded based on a kinship threshold of 0.0884 [26]. We implemented a relatively relaxed threshold for LD and HWE to facilitate the feature selection approaches.

Feature selection methods

In GWAS, a genotyping dataset may contain thousands of samples and up to four million SNPs, leading to the curse of dimensionality. Additionally, irrelevant and insignificant variants can hinder ML techniques from accurately identifying true SNP-SNP relationships within the dataset [27]. Directly using the entire dataset to train an ML model may result in the model learning noise and random fluctuations, leading to overfitting/underfitting. Feature selection methods address these challenges by focusing on significant variants, thereby reducing data dimensionality, optimizing the model, and enhancing prediction performance [16, 28, 29]. For this purpose, we employed four methods, including LASSO, ridge, elastic-net, and mutual information, to select a subset of 5000 SNPs. This selection was based on the Bonferroni correction method, which adjusts the significance p-value threshold (0.05) to account for the number of tests performed [26]. To comply with this correction, we used a suggestive p-value threshold of 1×10^{-5} for association test, as referenced from the NHGRI GWAS Catalog [30]. Consequently, feature selection

techniques identified 5000 SNPs, aligning with the corrected p-value threshold calculated using the formula 0.05/number of tests performed (i.e. independent variants).

LASSO

Equation (1) is LR in which $\hat{y}_i$ is a dependent variable/predicted phenotype and has a linear relationship with allele x_{ij}. β_j is regression coefficient, i.e. effect sizes of alleles/features, ϵ_i is normal errors with mean 0 and known variance σ^2.

$$\hat{y}_i = \sum_{i=1,j=1}^{n,snps} \beta_j x_{ij} + \epsilon_i \tag{1}$$

LASSO uses L1 penalty ($\sum_{j=1}^{snps} \|\beta_j\|_1$) on residual sum of squares (RSS) [14]. The main objective of LASSO is to regularize and select features by minimizing the RSS between actual and predicted values of phenotype $(y_i - \hat{y}_i)^2$ with L1 added penalty.

$$L(\hat{\beta}_{lasso}) = \min_{\beta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{snps} \beta_j x_{ij} \right)^2 + \alpha \sum_{j=1}^{snps} \|\beta_j\|_1 \tag{2}$$

Depending on the penalization parameter or the strength of penalty α in Eq. (2), some coefficients may be precisely zero, which causes only relevant features in the model. It provides a sparse solution and minimizes the model's variance and bias.

Ridge regression

Unlike LASSO, ridge regression uses L2 penalty ($\sum_{j}^{snps} \|\beta_j\|_2^2$) on RSS [18]. The ridge regression objective is to update effect sizes by minimizing the RSS.

$$L(\hat{\beta}_{ridge}) = \min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{snps} \beta_j x_{ij} \right)^2 + \alpha \sum_{j=1}^{snps} \|\beta_j\|_2^2 \tag{3}$$

The larger the value of penalization parameter α in the above Eq. (3), the smaller the coefficients become, but never reduce to absolute zero. So, ridge regression does not perform feature selection and cannot reduce model complexity. However, GWAS data have multicollinearity among their features; ridge regression is instrumental in avoiding it [31].

Elastic-net

Elastic-net combines both LASSO (L1 penalty) and ridge regression (L2 penalty) [19]. Elastic-net may create a greater number of accurately connected features than LASSO. It also has a substantially lower false positive rate than ridge regression [32].

$$L(\hat{\beta}_{elastic-net}) = \min_{\beta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{snps} \beta_j x_{ij} \right)^2 + \alpha(l1_{ratio} \|\beta_j\|_1 + (1 - l1_{ratio}) \|\beta_j\|_2^2) \tag{4}$$

Here, in the above Eq. (4) α value is a strength of penalty, and $l1_{ratio}$ is elastic-net mixing parameter with $0 <= l1_{ratio} <= 1$. If $l1_{ratio}$ is set to zero, resulting in ridge regression, and if its value is set to one, it is equivalent to a LASSO penalty. Even if there is collinearity among features, elastic-net handles it effectively and keeps mean square error at a minimal [33].

Mutual information

Mutual information was first proposed by Shannon [20]. This method is widely used for feature selection. Mutual information shows how much dependence between two random variables.

$$\mathrm{I(X, Y)} = \mathrm{H(X)} - \mathrm{H(X|Y)} \tag{5}$$

In the above Eq. (5), I(X, Y) is the mutual information of X and Y, where H(X) is entropy of X, and H(X | Y) is conditional entropy of X given Y. Mutual information values can be equal to zero (independent variables) or larger than zero (dependent variables). The larger the mutual information value is, the stronger the correlation between two random variables. In GWAS, phenotypes and genotypes are considered random variables [34].

Association analysis methods

Among selected SNPs through feature selection, association of SNPs was performed to determine significant SNPs. For that, conventional GWAS has some challenges, such as not taking SNP-SNP interactions into account. To address these challenges, we utilized advanced machine learning models, including RF [22] and SVR [23], alongside the traditional statistical method, LR [21]. For applying these association methods, the dataset with selected features was split into training and testing subsets, with 70% of the data allocated for training and 30% for testing.

Linear regression

LR analysis predicts the value of one variable depending on another. The variable that we predict is known as the dependent variable, i.e., phenotype. The variable used to predict the value of another variable is known as the independent variable, i.e. genotype. This type of analysis determines the coefficients of a linear equation using one or

more independent variables that best predict the value of dependent variable. LR finds a straight line that minimizes the difference between expected and actual output values. As explained in Eq. (1), x_{ij} is allele that will predict the value of $\hat{y}_i$. The regression coefficient is defined as the slope of regression line β_j and measures effect sizes of allele x_{ij} [35].

Random forest

The RF method is an extension of the bagging method, as it uses both bagging and feature randomness to generate an uncorrelated forest of decision trees. The RF algorithm is composed of a collection of decision trees, with each tree in the ensemble consisting of data points selected from a training set with replacement, known as the bootstrap sample. Variables that are not used for training each tree due to random sampling, also known as the out-of-bag (OOB) set, used for internal validation. Another instance of randomization is then introduced by feature bagging, which increases dataset variety while decreasing correlation among decision trees. Feature bagging typically occurs at each split within a tree. This ensures greater diversity among the decision trees in the ensemble and contributes to reducing the correlation between them, thereby enhancing the overall robustness of the model. In a regression job, the individual decision trees will be averaged. Finally, the OOB sample is used for cross-validation to finalize the prediction [22]. The parameters in RF are *n_estimators*, which is the number of trees, and *max_depth*, which is the maximum depth of the tree [36].

Support vector regression

SVR is a potent ML technique for regression applications. It works by maximizing the margin between hyperplane and nearest data points, or support vectors, and determining which hyperplane best fits the training set by minimizing cost function as described in Eq. (6). SVR uses different kinds of kernels such as sigmoid, polynomial, radial basis function, and linear [23].

$$min\frac{1}{2}\boldsymbol{\beta_j}^2 + C\sum\nolimits_j(\xi_j + \xi_j^*) \tag{6}$$

In the above Eq. (6), C is inverse regularization parameter that controls the strength of penalty, and ξ parameter defines the epsilon tube of width $|\xi - \xi^*|$ that training loss function does not penalize within, with points expected to be close to the actual value of phenotype [36].

XGBoost

XGBoost is a proficient ML method with great efficiency and predicted accuracy. It is a tree-boosting methodology, which sequentially creates an ensemble of decision trees to correct errors from previous trees, making it highly successful for a wide range of predictive modelling problems. XGBoost's regularization prevents overfitting, which is critical for increasing model generalization (see Eq. (7)).

$$\text{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\beta_j^2 \tag{7}$$

where l is loss function, T is number of leaves, γ is regularization parameter for number of leaves, and λ is regularization parameter for leaf weights (β_j). XGBoost also enables parallel processing and handles missing information, which improves its adaptability and performance. This method is well known for its capacity to handle large-scale data and generate robust models with excellent predictive power [37].

Evaluation metrics

The performance of each combination of feature selection and association methods was evaluated on the test dataset after the models were trained on the training dataset. The evaluation was performed using the coefficient of determination (R^2) [38]. This approach ensured that the assessment of performance was based on unseen data, providing a robust evaluation of the methods.

$$R^2 = 1 - \frac{\text{Sum squared residual (SSR)}}{\text{Sum of squares total (SST)}}, \tag{8}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{9}$$

In the above Eq. (9), y_i is actual phenotype, $\hat{y}_i$ is predicted phenotype, and $\bar{y}$ is mean of the actual phenotype. R^2 elucidates the extent to which variance in phenotype is explained by genotype data of samples. A $R^2 = 1$ represents that the given genotype dataset accounts for all variations in the phenotype observed in sample data. Conversely, a $R^2 = 0$ indicates that genotype dataset does not explain any of variations observed in the phenotype. R^2 serves as a measure of how well the genotype predicts the phenotype, with higher values indicating stronger predictive power.

Furthermore, another evaluation metric, mean absolute error (MAE), is used, which serves as a measure of model accuracy. MAE calculates an average of the absolute difference between values predicted by a model ($\widehat{y_i}$) and the actual value (y_i), as shown in the Eq. (10).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \widehat{y}_i)| \quad (10)$$

MAE is less influenced by outliers due to its linear weighting, offering a more balanced assessment of model performance. A lower MAE indicates higher accuracy in predicting the phenotype based on genotype data.

Dataset

The effectiveness of the proposed ensemble approach was evaluated using simulated datasets and validated on real and imputed PennCATH datasets.

Simulated dataset

We describe our simulation approach within the ADEMP (Aim, Data-generating mechanisms, Estimands, Methods, Performance) framework [39].

Aim: The simulation aimed to efficiently identify phenotype-associated SNPs in GWAS, as outlined above.

Data-generating mechanisms: The simulated dataset utilized genotype data from HapMap3 [40], with QT data generated using the G2P (Genotype-to-Phenotype) simulation tool [41]. The HapMap3 dataset provided comprehensive genotypic information comprising 1,397 samples and 1,457,897 SNPs. In the G2P simulation, a single continuous trait was simulated based on the sum of genetic effects derived from HapMap3 genotypes. Heritability was set to 0.45, reflecting the average heritability of total cholesterol [42, 43]. This approach ensured that the simulated phenotype data mirrored realistic genetic architectures of quantitative traits, such as total cholesterol. The simulation process was repeated 100 times to generate 100 independent simulated datasets.

Estimands: The number of quantitative trait nucleotides (QTNs) was set to 5,000, aligning with the objective of selecting 5,000 SNPs following feature selection. The effects of these QTNs were modelled using a normal distribution to reflect realistic genetic contributions.

Methods: The proposed feature selection methods, including LASSO, ridge regression, elastic-net, and mutual information, were applied to identify SNPs associated with the phenotype. These selected SNPs were subsequently tested using machine learning-based association methods, including LR, RF, SVR, and XGBoost. The ensemble method's effectiveness was compared against individual methods in identifying significant SNPs associated with the simulated phenotype.

Performance: The R^2 metric was used to evaluate the predictive performance of genotype-phenotype associations. In practical terms, R^2 indicates the extent to which genetic variants account for observed phenotypic variance, with higher values reflecting stronger predictive capability. This metric served as a critical measure for comparing the performance of different feature selection and association methods. Additionally, MAE was used as a complementary evaluation metric. Unlike R^2, MAE is less influenced by outliers, providing a balanced measure of model accuracy.

Real PennCATH dataset

In the PennCATH cohort study, 3,850 individuals participated in an angiographic CAD case-control GWAS. For this study, we utilized a subset of 1,401 samples, which are publicly available [44]. For the subset, 861,473 SNPs were genotyped using the Affymetrix 6.0 GeneChip platform. The Institutional Review Board of the University of Pennsylvania approved the study, and all participants provided informed consent along with details regarding their ethnicity [45]. The details of PennCATH data are explained in the supplementary.

Imputed PennCATH dataset

Before the imputation process, the PennCATH dataset was phased using SHAPEIT4 method [46]. Afterwards, the phased dataset was imputed using two widely used tools, IMPUTE5 [47] and Beagle5.4 [48]. We implemented these two methods based on their strengths in addressing different variant frequencies. IMPUTE5 is more effective for imputing low-frequency and rare variants, while Beagle5.4 demonstrates superior accuracy for common variant imputation[49]. To generate reliable and high-quality imputed datasets, stringent thresholds are applied, including imputation scores $>= 90\%$ in IMPUTE5-generated datasets and probability scores $>= 70\%$ in Beagle5.4-generated datasets. The total number of SNPs generated after IMPUTE5 and Beagle5.4 on the PennCATH dataset was 14,848,075 and 8,803,043, respectively, while maintaining a consistent sample size of 1401 as in the original PennCATH study.

Results

We performed a comprehensive assessment of four different feature selection methods and four association analysis methods using real PennCATH, imputed and simulated datasets. The results of our analyses are described below in detail.

Performance evaluation on simulated datasets

Initially, the hapmap genotype dataset consisted of 1397 samples and 1,457,897 SNPs. The above-mentioned QC filters were then applied to this dataset. Following the application of SNP call rate filter < 100%, 376,084 SNPs remained. None of these SNPs had MAF < 0.01 or showed significant deviation from HWE ($P < 1 \times 10^{-5}$). All 1397 samples maintained sample call rate above 95%, and no samples exhibited familial relationships in the

kinship analysis. After LD pruning step, 191,575 SNPs were removed, resulting in a final dataset containing 1397 samples with 184,509 SNPs.

Further, feature selection methods were applied to the processed simulated datasets to select 5,000 SNPs. For the LASSO method, we set $\alpha = 9 \times 10^{-4}$, for ridge regression $\alpha = 5 \times 10^{-3}$, and for the elastic-net method, we set $\alpha = 0.09$ with an optimal $l1_{ratio}$ of 0.5. From each method, the top 5000 SNPs were selected based on their rankings. Additionally, the mutual information method selected the top 5000 SNPs based on their mutual information scores. This entire process was repeated across all 100 simulated datasets.

After feature selection, the selected SNPs were tested for association with LDL-cholesterol levels using four methods: LR, RF, SVR, and XGBoost. The performance of each method was evaluated using R^2 and MAE. Among all combinations, elastic-net combined with SVR consistently outperformed others, achieving an average $R^2 = 0.91$ and average MAE = 10.15 (see Table 1). Across the 100 simulated datasets, R^2 values ranged from 0.83 to 0.95, while MAE ranged from 4.82 to 19.38, reflecting the robustness and reliability of the proposed approach.

Performance validation on PennCATH-real dataset

The PennCATH-real dataset comprised 1401 samples and 861,473 SNPs. Several stringent standards were used during data preprocessing to ensure the quality. Some 688,840 SNPs with call rate < 100% were removed for further analysis. Next, SNPs were filtered using two criteria: 1) divergence from HWE ($P < 1 \times 10^{-5}$) removed 31 SNPs; 2) MAF > 0.01 removed 41,118 SNPs. Some 61,582 SNPs were eliminated by LD pruning ($r^2 < 0.3$). All 1401 samples had call rate > 95%. Kinship analysis was performed using a threshold of 0.09, which did not remove any samples. After implementing all QC measures, the final dataset consisted of 1282 samples and 69,902 SNPs.

We applied four different feature selection approaches to the above preprocessed PennCATH dataset. For LASSO method, we used an $\alpha = 4.5 \times 10^{-4}$, which resulted in 5003 SNPs. The ridge regression identified top 5000 SNPs at an α value of 5×10^{-3}. Likewise, elastic-net approach resulted in 5037 SNPs with an α value of 3.3×10^{-3} and an $l1_{ratio}$ of 0.5. From the mutual information method, we selected the top 5000 SNPs based on their mutual information scores. The Venn diagram shows the number of overlapping SNPs selected by four different feature selection methods (Fig. 2). Some 80 SNPs were shared by all four methods. The maximum

Table 1 Performance evaluation of various combinations of feature selection and association methods for all five datasets

Association methods →	Linear Regression (LR)		Random Forest (RF)		Support Vector Regression (SVR)		XGBoost	
Feature selection ↓	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
Simulated dataset								
LASSO	0.78	12.46	0.39	20.64	0.79	12.16	0.34	21.47
Ridge	0.84	10.57	0.10	24.13	0.77	12.86	0.04	25.53
Elastic net	0.90	7.36	0.43	20.18	**0.91**	**10.15**	0.35	21.23
Mutual Information	−0.53	32.10	0.02	25.11	−0.51	31.80	−0.05	26.31
PennCATH dataset								
LASSO	0.63	16.056	0.02	26.329	0.65	15.416	−0.04	26.94
Ridge	0.79	11.576	0.04	25.972	0.79	11.597	−0.37	28.73
Elastic net	0.86	9.961	0.04	26.018	**0.89**	**8.666**	−0.14	27.82
Mutual Information	−0.36	31.257	0.01	26.451	−0.45	32.073	−0.20	28.32
Imputed PennCATH-dataset with IMPUTE5								
LASSO	0.68	14.69	0.04	25.91	0.70	14.31	−0.03	26.65
Ridge	0.85	10.27	0.06	25.31	0.84	10.64	−0.04	27.09
Elastic net	0.92	7.29	0.10	26.02	**0.94**	**6.00**	0.04	26.06
Mutual Information	−0.35	30.88	−0.01	26.74	−0.45	31.41	−0.08	27.32
Imputed PennCATH-dataset with Beagle5.4								
LASSO	0.71	14.07	0.05	25.98	0.73	13.68	−0.06	26.44
Ridge	0.77	9.82	0.07	25.81	0.84	10.33	−5.58	27.26
Elastic net	0.93	6.86	0.07	25.80	**0.94**	**5.88**	0.09	24.98
Mutual Information	−0.43	32.26	−0.01	27.03	−0.50	32.92	−0.10	28.17

Bold values denote the best performance value

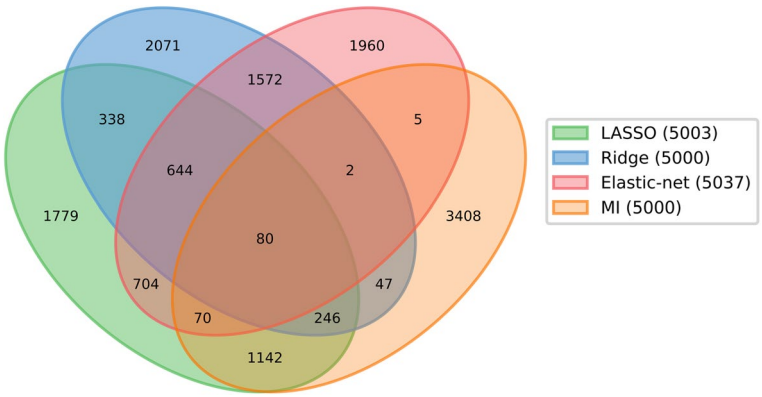

Fig. 2 Venn diagram illustrating the distribution of shared and distinct SNPs of real-PennCATH dataset among four feature selection methods. Some 80 SNPs were common across all four methods. The legend colors show the four feature selection methods, and the number of SNPs selected by each method is shown in parentheses. *Abbreviations*- LASSO: least absolute shrinkage and selection operator; MI: mutual information

overlap of 1572 SNPs was observed between elastic-net and ridge regression methods. LASSO and mutual information methods shared 1142 SNPs.

After the feature selection step, the selected SNPs were tested for association with LDL level using four different methods viz. LR, RF, SVR, and XGBoost. The performance of each method was evaluated using the R^2 and MAE. When different feature selection methods were paired with various association methods, majority of the combinations displayed lower R^2 values and higher MAE values. For instance, combinations of elastic-net with RF under-performed with $R^2 = 0.04$ and MAE = 26.02 (Table 1). Notably, elastic-net combined with SVR outperformed other combinations, achieving an R^2 of 0.89 and an MAE of 8.67.

The hyperparameter tuning for RF and XGBoost models was done, as described in the supplementary material, to optimize their performance. For the RF model, we obtained the following optimal parameters: *max_depth* = None, *min_samples_leaf* = 4, *min_samples_split* = 2, and *n_estimators* = 400. Similarly, optimal parameters for XGBoost model were as follows: *learning_rate* = 0.1, *max_depth* = 3, and *n_estimators* = 300. Despite using above-optimized parameters, both RF and XGBoost could not perform well, as shown in Table 1. The negative R^2 indicates that the regression model performs worse than a simple horizontal line (mean predictor). This is probably due to the fact that tree-based methods like RF and XGBoost face challenges in predicting unseen values, often due to high variance or poorly detected patterns in the data [50]. Their predictions tend to be constrained between the maximum and minimum values observed in the training data. This limitation is particularly evident when predicting patterns for QTs [51].

Further, we proceeded with 5000 SNPs that were obtained from the best-performed combination of feature selection and association method, namely, elastic-net with SVR method. We calculated the permutation importance scores for these 5000 SNPs, as per details provided in [52]. These scores were used to evaluate the relative importance of each SNP in explaining the variance in LDL-cholesterol levels. Based on these scores, we identified the top 100 SNPs having the most impact on LDL-cholesterol levels. A comparative graph of the permutation importance scores for all 5000 SNPs and the top 100 SNPs selected through the elastic-net and SVR combination is presented in the supplementary material (see Figure S1).

Post-GWAS analyses

We performed further post-GWAS analyses of top 100 SNPs identified from the permutation importance score of PennCATH-real dataset to elucidate their functional and biological role in LDL-cholesterol regulation. These analyses provided deeper insights into the genetic mechanisms underlying LDL-cholesterol levels and identified potential targets for therapeutic intervention.

Comparison with previous findings

At first, we conducted a comprehensive literature search to ascertain the experimentally validated functions of SNPs identified in our analysis. The SNP rs4591370 obtained from PennCATH-real dataset has been previously reported to be significantly ($P = 8.2 \times 10^{-9}$) associated with circulating LDL-cholesterol concentrations [53]. Further, rs7232775 showed association with blood urea nitrogen ($P = 8.0 \times 10^{-14}$) [54]; and rs12438724 has been associated with fibrotic idiopathic interstitial pneumonias ($P = 4.10 \times 10^{-8}$) [55]. Both these SNPs

have been related to cholesterol levels in blood serum. Next, we also searched the genome-wide repository of associations between SNPs and phenotypes (GRASP) [56] database to get relevant literature associated with our resultant SNPs. Notably, among 100 identified SNPs, 23 were experimentally validated to be associated with LDL-cholesterol and/or LDL-cholesterol-related diseases such as coronary artery disease (CAD), high blood pressure, and Alzheimer's disease. Table S2 provides detailed information on these 23 SNPs with their corresponding *p*-values.

eQTL analysis

The expression quantitative trait locus (eQTL) analysis serves as a fundamental tool for elucidating the regulation of gene expression. The genotype-tissue expression (GTEx) [57] and eQTLGen consortium [58] were utilized to investigate SNP expression across specific tissues and whole blood. Additionally, we assessed expressions of genes using functional mapping and annotation of genome-wide association studies (FUMAGWAS) [59]. The GENE2FUNC tool of FUMAGWAS was used to assess tissue specificity based on genes annotated from identified SNPs, that are more (or less) expressed in a specific tissue compared to all other tissues. Among the top 100 SNPs identified through our framework, eQTL analysis resulted in 64 SNPs and 60 genes with significantly higher expression levels ($P < 0.05$) in various tissues, including whole blood, adipose (both visceral and subcutaneous), fibroblast, oesophagus (gastroesophageal junction and mucosa), heart, artery, pancreas, brain, and thyroid. These organs are also identified in the expression analysis of genes annotated from identified SNPs from FUMAGWAS. The differentially expressed genes (DEG) were significantly enriched for heart, liver, pancreas, blood, etc. These tissues play a key role in regulating cholesterol levels. Figure 3A highlights the affected tissues/organs resultant from both eQTL and expression analyses. All eQTLs and non-eQTLs identified from our framework are shown in Fig. 3B. Furthermore, differential expression analysis of annotated genes highlights the involvement of previously identified genes known to regulate LDL-cholesterol. For instance, APOB gene, a well-known carrier of LDL-cholesterol, exhibited differential expression in the liver, small intestine, and heart [60, 61]. Also, RAB2A gene is involved in cholesterol efflux mechanisms [62], and TRAPPC9 is associated with serum LDL-cholesterol levels [63]. Both these genes showed differential expression patterns, further emphasizing their importance in cholesterol homeostasis. A heatmap of the corresponding genes and their expression is visualized in Fig. 3D. The *p*-values of these 64 SNPs across various tissues can be found in supplementary Table S1. These findings identify SNPs that influence specific tissues/organs, potentially contributing to traits and diseases associated with fluctuating cholesterol levels. The tissue-specific effects of these SNPs may drive the distinct biological processes involved in modifying cholesterol levels.

Variants involved in transcriptional regulation

In order to interpret the effects of SNPs, it is crucial to study their impact on gene regulation mechanisms. SNPs often alter gene regulation through changes to transcription factor binding sites (TFBS) [64]. Since intronic and intergenic regions have the highest concentration in identified SNPs, searching for TFBS could be beneficial. This can help to understand the molecular pathways through which SNPs impact the phenotype of interest. SNP2TFBS database [65, 66] was used to identify and visualize the SNPs affecting the TFBSs. Transcription factor (TF) enrichment plot generated from SNPs identified in our proposed pipeline highlights significant associations ($P < 0.05$) with cholesterol-related TF (Fig. 3C). For instance, the absence of GATA4 TF leads to increased plasma cholesterol levels [67] and has an association with LDL-cholesterol [68]. Overexpression of GATA2 leads to increased cholesterol efflux from macrophages [69]. GATA2 also play a role in regulating cholesterol storage [70]. Inhibition of FOS TF impacts cholesterol biosynthesis, and JUN family TFs, acting as heterodimeric partners of FOS, influence cell membrane composition in the presence of cholesterol [71]. The absence of MECOM TF reduces LDL uptake by human umbilical vein endothelial cells by four-fold [72]. Additionally, NR4A2 TF inhibits oxidized LDL uptake by macrophages, diminishes

(See figure on next page.)

Fig. 3 Post-GWAS analysis of top 100 SNPs identified using PennCATH-real dataset through our proposed framework. **A** Visualization of tissues/organs exhibiting enrichment of identified SNPs and annotated genes, affected by fluctuations in LDL cholesterol levels, **B** Chromosomal mapping of eQTLs and non-eQTLs in identified SNPs, **C** Enrichment-plot of TFs detected from identified SNPs, **D** heatmap of DEG annotated from identified SNPs in corresponding tissues, **E** GO analysis of annotated genes, **F** Top pathways, phenotypes, and proteins enriched from annotated genes. Abbreviations- GWAS: genome-wide association studies; SNP: single nucleotide polymorphism; LDL: low-density lipoprotein; eQTL: expression quantitative trait locus; TF: transcription factor; DEG: differentially expressed genes; GO: gene ontology; BP: biological process; CC: cellular component; MF: molecular function

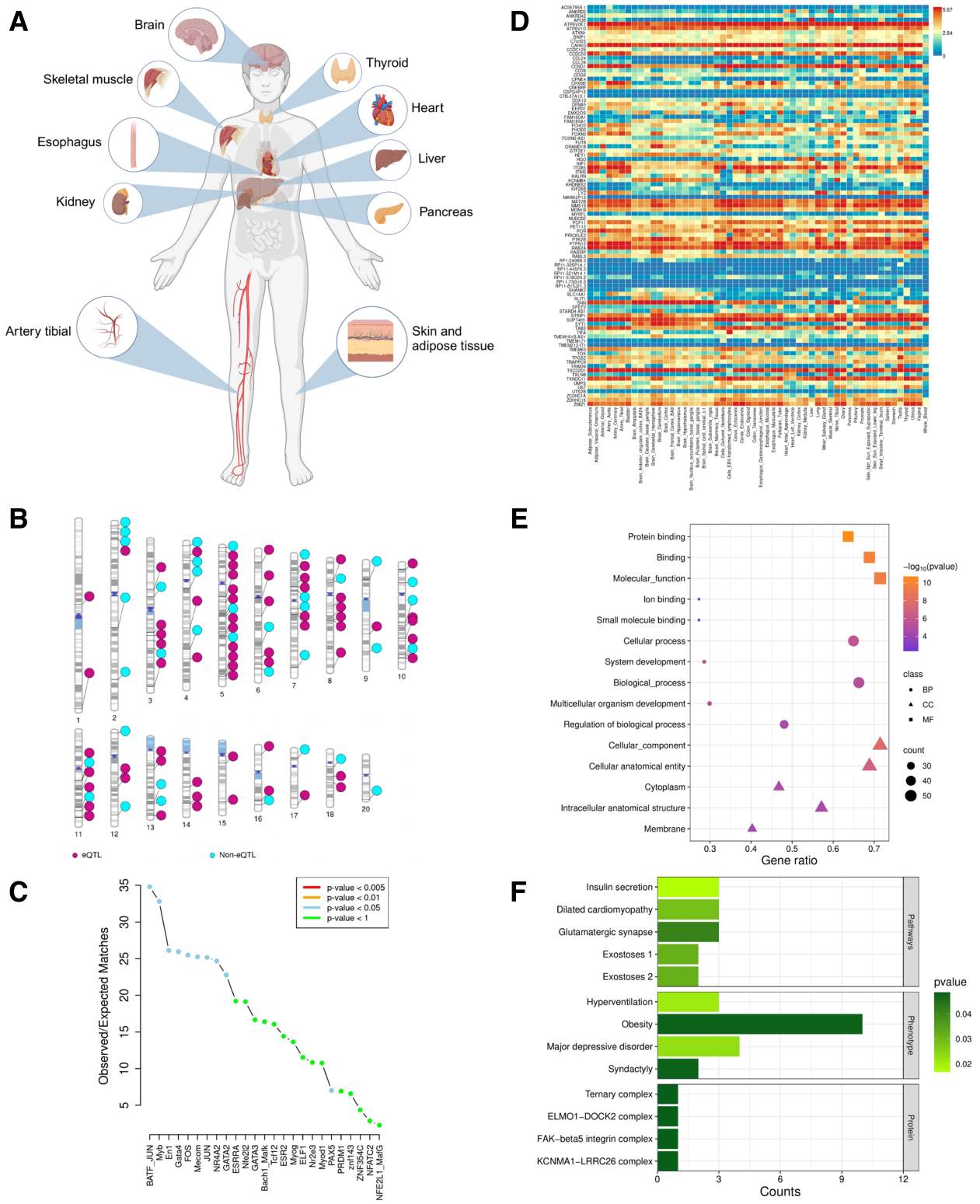

Fig. 3 (See legend on previous page.)

pro-inflammatory cytokine and chemokine expression, and is associated with blood pressure regulation [73, 74]. Moreover, we identified important TF binding sites such as BATF_JUN, Myb, En1, and PAX5, enriching our understanding of cholesterol-related transcriptional regulation.

Functional enrichment analysis

Since most of the resultant SNPs are located in noncoding regions of the genome, we conducted regulatory region enrichment analysis using HaploReg v4.2 [75], based on data from the Roadmap Epigenomics Consortium. Majority of the identified SNPs, along with their proxy SNPs within the LD range ($r^2 > 0.8$), show significant enrichment for enhancer regions (including promoter and enhancer histone modifications) DNase sensitivity, and motif alterations, particularly in whole blood, heart (left ventricle), and pancreas tissues. Detailed results are provided in the supplementary Table S7.

The enrichment analysis was performed to reveal significantly enriched biological processes, pathways, regulatory motifs and protein complexes. We used g:Profiler, a web server for functional enrichment analysis, to perform the gene ontology (GO) analysis [76]. The significantly enriched molecular functions, biological processes, and cellular components involved majority of binding functions like protein, ion and small molecule, cellular and developmental processes, and cytoplasm and membrane, respectively. The top GO terms are shown in Fig. 3E. The pathway enrichment analysis revealed pathways related to insulin secretion, synapse and heparan sulphate.

Functional gene annotation

The annotated genes from identified SNPs were also involved in various cholesterol-modulated phenotypes. For instance, TRAPPC9 gene [63] is responsible for obesity-related traits (a condition characterized by elevated LDL-cholesterol) [77]. Additionally, higher LDL-cholesterol has been linked to major depressive disorders [78], and genes associated with these conditions, such as CCL24 [79], have been identified as well. Further, LDL-cholesterol levels are suggestively associated with FCHO2 gene, which plays a crucial role in the clearance of LDL-cholesterol from the bloodstream [80]. Moreover, EEPD1 gene (endonuclease-exonuclease-phosphatase family domain containing 1) functions as part of the LXRs (Liver X Receptors)-regulated program, promoting ABCA1-dependent (ATP-binding cassette transporter A1) cholesterol efflux from macrophages [81]. Conversely, gene annotation identified genes associated with syndactyly, a condition characterized by fused digits, which showed a correlation with decreased serum cholesterol levels [82, 83]. Notably, proteins such as fructose-bisphosphatase-1 (FBPase) and synaptotagmin-1 were identified as regulators of abnormal cardiac atrium morphology, such as atrial fibrillation (AF), where both low cholesterol-levels and high cholesterol variability were linked to increased AF risk [84]. Furthermore, these two proteins are also implicated in hyperventilation conditions. Liver FBPase is known to be upregulated by obesity and dietary fat intake, suggesting a potential link between LDL and FBPase activity [85]. Cholesterol depletion disrupts synaptotagmin-1-induced membrane bending, which impairs synaptic transmission and neuronal function [86]. Additionally, glypican-6 regulates cardiomyocyte growth via the ERK1/2 signalling pathway [87], which is also involved in cholesterol trafficking [88]. Gustafsen, Camilla, et al. suggested in their study to investigate the specificity of interaction between glypicans and PCSK9 [89], which may give a biological understanding of increasing circulating LDL-cholesterol levels [90].

The identified SNPs were annotated to proteins using the CORUM database within g:Profiler. Among the annotated protein complexes, we identified several noteworthy associations with cholesterol. Firstly, the ternary complex containing GATA4, SRF, and MYOCD was observed, wherein cholesterol loading suppressed the expression of MYOCD [91]. Additionally, the ELMO1-DOCK2 complex, annotated from identified SNPs, revealed cholesterol sulfate as a potent inhibitor of DOCK2 [92]. The FAK-beta5 integrin complex (VEGF-induced complex) demonstrated cholesterol's regulatory role in VEGF:VEGFR-1 signalling, influencing cell migration and viability, potentially impacting disease progression [93] and also regulating the innate immune response [94]. Furthermore, the KCNMA1-LRRC26 complex highlighted the importance of KCNMA1 in cholesterol transport, as evidenced by significant cholesterol accumulation in the absence of KCNMA1 in mouse embryonic fibroblasts. This finding suggests a conserved role for KCNMA1 in the efficient cholesterol transport, implicating its importance in human physiology as well [95]. The top enriched pathways, phenotypes, and proteins are displayed in Fig. 3F.

Drug targets analysis

In this study, we analyzed potential drug targets among resultant genes associated with diseases influenced by cholesterol level fluctuations. For instance, Mipomersen, an antisense oligonucleotide drug targeting APOB gene, is used to treat homozygous familial hypercholesterolemia by inhibiting APOB synthesis [96]. Similarly, Baricitinib and Leflunomide, anti-rheumatic drugs targeting PTK2B gene, are used to treat Alzheimer's disease by inhibiting Janus kinases [97]. Auranofin, acting through PTPN12-ErbB-2 signalling axis, reduces damage

from myocardial ischemia/reperfusion injuries by targeting the PTPN12 gene [98].

Gene targets for personalized medicine treatments

Our pipeline also identifies genes that can be utilized in personalized medicine treatments. For instance, deficiency in the conserved oligomeric Golgi complex subunit 6 (COG6) gene causes congenital disorders of glycosylation (CDG), a rare autosomal recessive disease. Li G. et al. conducted a targeted NGS study on COG6, uncovering compound heterozygous variants that broaden the mutation spectrum and extend the genotype-phenotype relationship in CDG [99]. Moreover, ATXN1, a dosage-sensitive gene, is involved in neurodegenerative disorders like spinocerebellar ataxia type 1 and Alzheimer's disease. Nitschke L. et al. reported that mutations in miR760's binding site within the 5' UTR or in the 3' UTR binding sites of miRNAs and RNA-binding proteins could increase ATXN1 expression, causing ataxia symptoms. These findings underscore the importance of identifying ATXN1 regulatory regions and performing whole-genome sequencing in ataxia patients to identify potential disease-causing mutations in non-coding regions [100].

Performance evaluation on other datasets

The performance of feature selection and association analysis methods was further assessed on two imputed datasets of the same PennCATH study. The basic QC steps were applied to these two datasets, as described in the Methods section. Next, we separately passed the preprocessed datasets to four feature selection methods to select approximately 5000 SNPs from each method. Afterwards, we performed association analysis on these selected SNPs using four ML-based methods, adhering to the same hyperparameters as outlined for the PennCATH-real dataset. The results obtained from different combinations of feature selection and association analysis methods using imputed datasets are summarized in Table 1. Again, elastic-net combined with SVR outperformed other combinations of feature selection and association analysis methods, consistent with the results obtained from PennCATH-real dataset.

Performance comparison with conventional GWAS approach

Conventional GWAS methods often face challenges in accurately identifying phenotype-associated variants due to their reliance on stringent p-value thresholds and linear modelling approaches. In our analysis of the processed PennCATH-real dataset, which consists of 1282 samples and 69,902 SNPs, we performed a conventional GWAS pipeline using Plink2.0 with LDL cholesterol as the phenotype and 10 principal components as covariates. The results revealed no SNPs meeting the standard genome-wide significance p-value threshold (5×10^{-8}), while only two SNPs achieved a p-value threshold of 5×10^{-6}, and nine SNPs fell within the suggestive threshold of 5×10^{-5}. Among these 11 SNPs, two exhibited negative effect sizes, and only two SNPs were associated with LDL cholesterol or cholesterol-related diseases. Notably, rs3017499, expressed in adipose tissue [57], was identified as common between the conventional and proposed GWAS pipelines. Additionally, two genes were also common between conventional and proposed approaches, including TRAPPC9, which has been linked to serum LDL cholesterol levels [63].

In addition to the PennCATH-real dataset, we compared conventional GWAS methods with the proposed ensemble approach across 100 simulated datasets. The results demonstrated that the conventional GWAS pipeline identified between 0 to 5 QTNs, whereas the proposed pipeline consistently identified 8 to 29 QTNs. These QTNs were simulated as normally distributed during the generation of 5000 QTNs for the continuous trait. These findings underscore the challenges of conventional GWAS in detecting meaningful associations, particularly for traits with complex genetic architectures.

Polygenic risk score analysis

Polygenic risk score (PRS) analysis is a widely used approach to quantify the cumulative genetic risk of a trait by aggregating the effects of multiple SNPs. PRS models are especially valuable in predicting disease susceptibility and validating feature selection methods in genomic studies. Therefore, we performed an additional comparative analysis to assess the predictive power of PRS models using our proposed feature selection method in comparison to conventional GWAS approaches. For this purpose, the PennCATH-real dataset was split into a training set (897 samples, 69,902 SNPs) and a validation set (385 samples, 69,902 SNPs). PRS analysis was performed using *PRSice-2* package available in R software [101]. We used various p-value thresholds to determine the optimal set of SNPs contributing to polygenic risk prediction. The conventional GWAS-based PRS model included all SNPs that met a predefined significance threshold, whereas the elastic-net-based PRS model was built using the 5000 SNPs selected through our proposed feature selection approach.

The performance of PRS models was evaluated based on the R^2. Under the conventional GWAS framework, the PRS model achieved an R^2 of 0.086 at the best p-value threshold (10^{-5}), demonstrating modest predictive power. In contrast, PRS model incorporating elastic-net-selected SNPs showed a substantial improvement,

achieving an R^2 of 0.56. These findings demonstrate that our proposed feature selection approach enhances PRS predictive performance by prioritizing the most informative SNPs. By selecting the most relevant SNPs, this method effectively reduces noise and enhances signal detection, ultimately leading to a better estimation of genetic risk.

While our proposed feature selection approach enhances PRS predictive performance, we recognize a small sample size of our analysis, and the base and target datasets are from the same GWAS cohort. Since PRS models rely on effect sizes estimates obtained from GWAS summary statistics dataset(i.e. base dataset), using samples from the same cohort can inflate predictive performance due to overfitting or shared population structure. In contrast, PRS models generally provide more reliable and generalizable results when applied to larger, independent datasets, where the genetic effect sizes are estimated from a broader and more diverse population [102].

Despite these constraints, our approach demonstrates the potential of integrating machine learning-based feature selection with PRS analysis to improve disease susceptibility prediction. Future studies should validate this approach using larger datasets to establish its robustness and clinical relevance in genomics research.

Discussion

The emergence of GWAS has significantly enriched our understanding of human disease genetics in the past two decades, transitioning from analyzing common variants to exploring rare variants. GWAS findings provide valuable insights into disease biology, aiding in clinical application and revealing population-level risk stratification.

In this work, we integrated ML-based feature selection with association analysis to address the challenges of false negatives, epistasis interactions, and missing heritability. Our proposed approach offers alternative strategies to the conventional dependency on stringent p-value thresholds used in GWAS. We used various ML techniques for feature selection, including LASSO, ridge, elastic-net, and mutual information. Each of these techniques selected approximately 5000 SNPs on the PennCATH-real dataset. Further, we evaluated the associations between these selected SNPs and LDL-cholesterol levels using LR, RF, SVR, and XGBoost methods. Notably, employing the elastic-net for feature selection in combination with SVR yielded promising results, as evidenced by an R^2 of 0.89 and an MAE of 8.66. To evaluate the robustness of our proposed approach, we conducted additional assessments using two imputed datasets, 100 simulated datasets and one low-frequency variant dataset. Similar to the PennCATH-real dataset, the combination of elastic-net with SVR outperformed other feature selection and association method combinations across all these datasets. The low-frequency variant datasets were generated from the imputed PennCATH-real dataset, and details for the same are provided in the supplementary material, specifically in Tables S3, S4, S5, and S6. These promising results are likely due to the selection of 5000 SNPs using the Bonferroni correction method rather than the conventional stringent p-value threshold. Furthermore, the improved performance can be attributed to the ability of elastic-net method to account for SNP-SNP interactions during feature selection.

However, other feature selection and association algorithm combinations did not perform well. We also considered a popular non-linear ML method, XGBoost, for association tests, which is known to have a comparative performance in case-control GWAS [12]. However, in our analysis, this method did not perform well for QT studies. Next, in the PennCATH-real dataset, based on permutation feature importance scores, we ranked top 100 most effective SNPs out of 5000 selected SNPs from the combination of elastic-net and SVR methods.

In post-GWAS analyses, we conducted functional enrichment to gain insights into the biological significance of 100 SNPs identified from our proposed framework. Initially, we studied the expression patterns of these SNPs across different tissues and observed significant association to expression levels in tissues/organs known to play crucial roles in regulating cholesterol levels, including whole blood, adipose tissue, heart, and pancreas. Additionally, our study demonstrates the significant enrichment of down-regulated and both-sided DEG sets in the same tissues/organs. Notably, genes such as APOB, RAB2A, and TRAPPC9 exhibit differential expression patterns across various tissues, highlighting their potential role in cholesterol homeostasis. In addition to these three genes, our study also identified CCL24, EEPD1, and FCHO2, which are already known to be associated with LDL-cholesterol.

Although most of the identified SNPs are located in non-coding regions, they are enriched for regions that activate transcription through TFBS. Key findings of such SNPs include the role of GATA4 TF in plasma cholesterol regulation, the impact of GATA2 on cholesterol efflux and storage, and the influence of FOS and JUN TFs on cholesterol biosynthesis and cell membrane composition. Additionally, MECOM deficiency affects LDL uptake, while NR4A2 inhibits oxidized LDL uptake. Moreover, the identified SNPs play significant roles in enhancer regions, exhibit DNase activities, and show changes in motif regions in whole blood, heart, and pancreas tissues.

Furthermore, GO analysis highlighted the involvement of annotated genes in binding functions, cellular

and developmental processes, as well as cytoplasmic and membrane-related activities. Pathway enrichment analysis unveiled significant enrichment of pathways associated with insulin secretion, synapse function, and heparan sulfate metabolism, indicating potential biological mechanisms underlying the observed associations with LDL-cholesterol. Additionally, the annotated genes exhibit intricate role of LDL-cholesterol in regulating obesity, major depressive disorders, syndactyly, abnormal cardiac morphology, and neuronal function. Furthermore, the involvement of key proteins such as FBPase, synaptotagmin-1, and glypican-6 suggests potential pathways through which LDL-cholesterol modulates various physiological processes. The potential of synaptotagmin-1 as a therapeutic target for neurodegenerative and neurodevelopmental diseases is noteworthy due to its involvement in synaptic transmission and neuronal function. Addressing the specificity of interactions between glypicans and PCSK9 may offer insights into mechanisms influencing circulating LDL-cholesterol levels. Moreover, the annotation of identified SNPs to protein complexes further reveals the importance of cholesterol in modulating cellular functions and signalling pathways, highlighting its potential implications in innate immunity, disease progression and cellular homeostasis.

In addition, our proposed framework identifies potential drug targets among genes associated with cholesterol-related diseases. For example, Mipomersen targets APOB for treating familial hypercholesterolemia, Auranofin targets PTPN12 to reduce myocardial ischemia, while Baricitinib and Leflunomide target PTK2B for Alzheimer's. Further, genes like COG6 and ATXN1 hold promise for personalized medicine in conditions like CDG and neurodegenerative disorders, respectively. In summary, the identified SNPs, genes, and proteins provide valuable insights into potential therapeutic targets for managing cholesterol-related diseases.

The current challenge in our study is the inability to use other large GWAS datasets in our framework due to their unavailability. Future studies should focus on incorporating these large datasets to improve the robustness and reliability of our proposed framework for detecting genetic associations related to complex traits. This would enable a more thorough evaluation of our methodology and its effectiveness in identifying meaningful genetic associations across diverse populations, thereby advancing our understanding of complex diseases.

Conclusion

In conclusion, our proposed ensemble learning approach, which integrates elastic-net with SVR, effectively identifies variants associated with QTs. Our comprehensive analyses demonstrate that elastic-net effectively mitigates the issue of multicollinearity, while SVR alleviates limitations imposed by stringent p-value thresholds. Moreover, our approach follows an in-depth biological enrichment analysis to further reduce the false positive rate. We anticipate that the integration of T2T and pangenome references will further enhance the utility of our approach in future GWAS.

Abbreviations

AF	Atrial Fibrillation
CAD	Coronary Artery Disease
CDG	Congenital Disorders of Glycosylation
COG6	Conserved Oligomeric Golgi Complex Subunit 6
DEG	Differentially Expressed Genes
eQTL	Expression Quantitative Trait Locus
FBPase	Fructose-Bisphosphatase-1
FUMAGWAS	Functional Mapping And Annotation of Genome-wide Association Studies
GO	Gene Ontology
GRASP	Genome-wide Repository of Associations Between SNPs and Phenotypes
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide Association Studies
GxE interaction	Gene-Environment Interactions
HGP	Human Genome Project
HWE	Hardy-Weinberg Equilibrium
LASSO	Least Absolute Shrinkage and Selector Operator
LD	Linkage Disequilibrium
LDL	Low-Density Lipoprotein
LR	Linear Regression
LXR	Liver X Receptors
MAE	Mean Absolute Error
MAF	Minor Allele Frequency
ML	Machine Learning
OOB	Out-Of-Bag
QC	Quality Control
QT	Quantitative Trait
QTNs	Quantitative Trait Nucleotides
RF	Random Forest
RSS	Residual Sum of Squares
SNP	Single Nucleotide Polymorphisms
SST	Sum of Squares Total
SVR	Support Vector Regression
T2T	Telomere-to-Telomere
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
XGBoost	Extreme Gradient Boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11443-x.

Supplementary Material 1.

Authors' contributions

J.S.: Conceptualization, Methodology, Validation, Visualization, Writing -Original Draft. V.J.: Methodology, Software, Formal analysis, Writing- Original Draft. R.S.S.: Visualization, Writing - Original Draft. P.Y.: Conceptualization, Supervision, Resources, Writing- Reviewing and Editing. J.S. and V.J. contributed equally to this work.

Funding

This work was supported by the GenomeIndia grant (BT/GenomeIndia/2018) from the Department of Biotechnology and partly by the Ministry of Education, Government of India.

Data availability
This study utilizes publicly available data, and the access identifiers for these data are provided inside the manuscript. The PennCATH cohort study is available at https://pbreheny.github.io/adv-gwas-tutorial/quality_control.html. Code availability: https://github.com/VaishnaviJangale/GWAS-with-ML. Additionally, we have cited the original study (PMID: 21239051), which contains all relevant details regarding this dataset. We are unable to provide a GWAS catalog number for this dataset as the authors of the PennCATH study did not make this dataset available in the GWAS catalog database.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 16 October 2024 Accepted: 4 March 2025
Published online: 12 March 2025

References

1. Loos RJ. 15 years of genome-wide association studies and no signs of slowing down. Nat Commun. 2020;11(1):5900.
2. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.
3. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022;604(7906):437–46.
4. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007;316(5826):889–94.
5. Siminovitch KA. PTPN22 and autoimmune disease. Nat Genet. 2004;36(12):1248–9.
6. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, et al. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet. 2009;84(3):399–405.
7. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet. 2010;18(1):111–7.
8. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet. 2010;86(1):6–22.
9. Zhang Q, Long Q, Ott J. AprioriGWAS, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. PLoS Comput Biol. 2014;10(6):e1003627.
10. Han B, Chen XW, Talebizadeh Z, Xu H. Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks. BMC Syst Biol. 2012;6:1–12.
11. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446–50.
12. Medvedev A, Mishra Sharma S, Tsatsorin E, Nabieva E, Yarotsky D. Human genotype-to-phenotype predictions: Boosting accuracy with nonlinear models. PLoS ONE. 2022;17(8):e0273293.
13. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25(6):714–21.
14. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol. 1996;58(1):267–88.
15. Chen Z, Boehnke M, Wen X, Mukherjee B. Revisiting the genome-wide significance threshold for common variant GWAS. G3. 2021;11(2):jkaa056.
16. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinforma. 2022;2:927312.
17. Cho S, Kim K, Kim YJ, Lee JK, Cho YS, Lee JY, et al. Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis. Ann Hum Genet. 2010;74(5):416–28.
18. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for non-orthogonal problems. Technometrics. 1970;12(1):55–67.
19. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.
20. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423.
21. Hocking RR. Developments in linear regression methodology: 1959–l982. Technometrics. 1983;25(3):219–30.
22. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
23. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
24. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Curr Protocol Hum Genet. 2011;68(1):1–19.
25. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):s13742-015.
26. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27(2):e1608.
27. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics. 2010;26(4):445–55.
28. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(Mar):1157–82.
29. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res. 2004;5:1205–24.
30. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2013;42(D1):D1001–6.
31. Malo N, Libiger O, Schork NJ. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. Am J Hum Genet. 2008;82(2):375–85.
32. Tutz G, Ulbricht J. Penalized regression with correlation-based penalty. Stat Comput. 2009;19:239–53.
33. Buhlmann P, van de Geer S. Statistics for high-dimensional data. 2011th ed. Springer series in statistics. Berlin: Springer; 2011.
34. Guo H, Yu Z, An J, Han G, Ma Y, Tang R. A two-stage mutual information based Bayesian Lasso algorithm for multi-locus genome-wide association studies. Entropy. 2020;22(3):329.
35. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Ärzteblatt Int. 2010;107(44):776.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
37. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 785–794. New York: Association for Computing Machinery; 2016.
38. Nagelkerke NJ, et al. A note on a general definition of the coefficient of determination. Biometrika. 1991;78(3):691–2.
39. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102.
40. NHGRI — Sample Repository for Human Genetic Research. http://ccr.coriell.org/Sections/Collections/NHGRI/?SsId=11. Accessed 2 Jan 2025.
41. Tang Y, Liu X. G2P: a genome-wide-association-study simulation tool for genotype simulation, phenotype simulation and power evaluation. Bioinformatics. 2019;35(19):3852–4.

42. Williams PT. Quantile-specific heritability of total cholesterol and its pharmacogenetic and nutrigenetic implications. Int J Cardiol. 2021;327:185–92.
43. van Dongen J, Willemsen G, Chen WM, de Geus EJ, Boomsma DI. Heritability of metabolic syndrome traits in a large population-based sample [S]. J Lipid Res. 2013;54(10):2914–23.
44. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. Stat Med. 2015;34(28):3769–92.
45. Aea Helgadottir. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. Nat Genet. 2006;38:68–74.
46. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019;10(1):5436.
47. Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the positional burrows wheeler transform. PLoS Genet. 2020;16(11):e1009049.
48. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103(3):338–48.
49. Naito T, Okada Y. Genotype imputation methods for whole and complex genomic regions utilizing deep learning technology. J Hum Genet. 2024;69(10):481–6.
50. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj Comput Sci. 2021;7:e623.
51. Zhang H, Nettleton D, Zhu Z. Regression-enhanced random forests. ArXiv. 2019;abs/1904.10416.
52. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340–7.
53. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, Zhao JH, et al. LDL-cholesterol concentrations: a genome-wide association study. Lancet. 2008;371(9611):483–91.
54. Kim YJ, Go MJ, Hu C, Hong CB, Kim YK, Lee JY, et al. Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. Nat Genet. 2011;43(10):990–5.
55. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet. 2013;45(6):613–20.
56. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014;30(12):i185–94.
57. GTEx Consortium, Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60.
58. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53(9):1300–10.
59. Watanabe K, Taskesen E, Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1):1826.
60. Devaraj S, Semaan JR, Jialal I. Biochemistry, Apolipoprotein B. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2024.
61. Behbodikhah J, Ahmed S, Elyasi A, Kasselman LJ, De Leon J, Glass AD, et al. Apolipoprotein B and cardiovascular disease: biomarker and potential therapeutic target. Metabolites. 2021;11(10):690.
62. Robichaud S, Fairman G, Vijithakumar V, Mak E, Cook DP, Pelletier AR, et al. Identification of novel lipid droplet factors that regulate lipophagy and cholesterol efflux in macrophage foam cells. Autophagy. 2021;17(11):3671–89.
63. Liang ZS, Cimino I, Yalcin B, Raghupathy N, Vancollie VE, Ibarra-Soria X, et al. Trappc9 deficiency causes parent-of-origin dependent microcephaly and obesity. PLoS Genet. 2020;16(9):1–26.
64. Bryzgalov LO, Antontseva EV, Matveeva MY, Shilov AG, Kashina EV, Mordvinov VA, et al. Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data. PLoS ONE. 2013;8(10):e78833.
65. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2013;42(D1):D142–7.
66. Kumar S, Ambrosini G, Bucher P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res. 2016;45(D1):D139–44.
67. Patankar JV, Chandak PG, Obrowsky S, Pfeifer T, Diwoky C, Uellen A, et al. Loss of intestinal GATA4 prevents diet-induced obesity and promotes insulin sensitivity in mice. Am J Physiol Endocrinol Metab. 2011;300(3):E478–88.
68. Bideyan L, Rodríguez ML, Priest C, Kennelly JP, Gao Y, Ferrari A, et al. Hepatic GATA4 regulates cholesterol and triglyceride homeostasis in collaboration with LXRs. Genes Dev. 2022;36(21–24):1129–44.
69. Yin C, Vrieze AM, Rosoga M, Akingbasote J, Pawlak EN, Jacob RA, et al. Efferocytic defects in early atherosclerosis are driven by GATA2 overexpression in macrophages. Front Immunol. 2020;11:594136.
70. Kohlmeier A, Sison CAM, Yilmaz BD, Coon VJS, Dyson MT, Bulun SE. GATA2 and progesterone receptor interaction in endometrial stromal cells undergoing decidualization. Endocrinology. 2020;161(6):bqaa070.
71. Choi Y, Jeon H, Akin JW, Curry TE Jr, Jo M. The FOS/AP-1 regulates metabolic changes and cholesterol synthesis in human periovulatory granulosa cells. Endocrinology. 2021;162(9):bqab127.
72. Lv J, Meng S, Gu Q, Zheng R, Gao X, Kim JD, et al. Epigenetic landscape reveals MECOM as an endothelial lineage regulator. Nat Commun. 2023;14(1):2390.
73. Safe S, Jin UH, Morpurgo B, Abudayyeh A, Singh M, Tjalkens RB. Nuclear receptor 4A (NR4A) family-orphans no more. J Steroid Biochem Mol Biol. 2016;157:48–60.
74. Kardys I, van Tiel CM, de Vries CJ, Pannekoek H, Uitterlinden AG, Hofman A, et al. Haplotypes of the NR4A2/NURR1 gene and cardiovascular disease: the Rotterdam Study. Hum Mutat. 2009;30(3):417–23.
75. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2011;40(D1):D930–4.
76. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8.
77. Klop B, Elte JWF, Castro Cabezas M. Dyslipidemia in obesity: mechanisms and potential targets. Nutrients. 2013;5(4):1218–40.
78. Kim EJ, Hong J, Hwang JW. The association between depressive mood and cholesterol levels in Korean adolescents. Psychiatr Investig. 2019;16(10):737.
79. Trojan E, Chwastek J, Basta-Kaim A, et al. A potential contribution of chemokine network dysfunction to the depressive disorders. Current Neuropharmacol. 2016;14(7):705–20.
80. Zhang Q, Cai Z, Lhomme M, Sahana G, Lesnik P, Guerin M, et al. Inclusion of endophenotypes in a standard GWAS facilitate a detailed mechanistic understanding of genetic elements that control blood lipid levels. Sci Rep. 2020;10(1):18434.
81. Nelson JK, Koenis DS, Scheij S, Cook ECL, Moeton M, Santos A, et al. EEPD1 Is a Novel LXR Target Gene in Macrophages Which Regulates ABCA1 Abundance and Cholesterol Efflux. Arterioscler Thromb Vasc Biol. 2017;37(3):423–32.
82. Chinsky JM, Steiner RD. Chapter 30 - INBORN ERRORS OF METABOLISM. In: Carey WB, Crocker AC, Coleman WL, Elias ER, Feldman HM, editors. Developmental-Behavioral Pediatrics (Fourth Edition). 4th ed. Philadelphia: W.B. Saunders; 2009. p. 287–313.
83. Porter FD, et al. Malformation syndromes due to inborn errors of cholesterol synthesis. J Clin Investig. 2002;110(6):715–24.
84. Lee HJ, Lee SR, Choi EK, Han KD, Oh S. Low lipid levels and high variability are associated with the risk of new-onset atrial fibrillation. J Am Heart Assoc. 2019;8(23):e012771.
85. Visinoni S, Khalid NFI, Joannides CN, Shulkes A, Yim M, Whitehead J, et al. The role of liver fructose-1, 6-bisphosphatase in regulating appetite and adiposity. Diabetes. 2012;61(5):1122–32.
86. Ali Moussa HY, Shin KC, Ponraj J, Kim SJ, Ryu JK, Mansour S, et al. Requirement of Cholesterol for Calcium-Dependent Vesicle Fusion by Strengthening Synaptotagmin-1-Induced Membrane Bending. Adv Sci. 2023;10(15):2206823.
87. Thota LNR, Chignalia AZ. The role of the glypican and syndecan families of heparan sulfate proteoglycans in cardiovascular function and disease. Am J Physiol-Cell Physiol. 2022;323(4):C1052–60.

88. Zhou X, Yin Z, Guo X, Hajjar DP, Han J. Inhibition of ERK1/2 and activation of liver X receptor synergistically induce macrophage ABCA1 expression and cholesterol efflux. J Biol Chem. 2010;285(9):6316–26.
89. Gustafsen C, Olsen D, Vilstrup J, Lund S, Reinhardt A, Wellner N, et al. Heparan sulfate proteoglycans present PCSK9 to the LDL receptor. Nat Commun. 2017;8(1):503.
90. Canuel M, Sun X, Asselin MC, Paramithiotis E, Prat A, Seidah NG. Proprotein convertase subtilisin/kexin type 9 (PCSK9) can mediate degradation of the low density lipoprotein receptor-related protein 1 (LRP-1). PLoS ONE. 2013;8(5):e64145.
91. Xia XD, Yu XH, Chen LY, Xie SL, Feng YG, Yang RZ, et al. Myocardin suppression increases lipid retention and atherosclerosis via downregulation of ABCA1 in vascular smooth muscle cells. Biochim Biophys Acta (BBA) - Mol Cell Biol Lipids. 2021;1866(4):158824.
92. Kunimura K, Uruno T, Fukui Y. DOCK family proteins: key players in immune surveillance mechanisms. Int Immunol. 2019;32(1):5–15.
93. Casalou C, Costa A, Carvalho T, Gomes AL, Zhu Z, Wu Y, et al. Cholesterol regulates VEGFR-1 (FLT-1) expression and signaling in acute leukemia cells. Mol Cancer Res. 2011;9(2):215–24.
94. Pokharel SM, Shil NK, Gc JB, Colburn ZT, Tsai SY, Segovia JA, et al. Integrin activation by the lipid molecule 25-hydroxycholesterol induces a proinflammatory response. Nat Commun. 2019;10(1):1482.
95. Wang W, Zhang X, Gao Q, Lawas M, Yu L, Cheng X, et al. A voltage-dependent K+ channel in the lysosome is required for refilling lysosomal Ca2+ stores. J Cell Biol. 2017;216(6):1715–30.
96. Li Z, Zhang B, Liu Q, Tao Z, Ding L, Guo B, et al. Genetic association of lipids and lipid-lowering drug target genes with non-alcoholic fatty liver disease. EBioMedicine. 2023;90:104543.
97. Kwok MK, Lin SL, Schooling CM. Re-thinking Alzheimer's disease therapeutic targets using gene-based tests. EBioMedicine. 2018;37:461–70.
98. Yang CF, Chen YY, Singh JP, Hsu SF, Liu YW, Yang CY, et al. Targeting protein tyrosine phosphatase PTP-PEST (PTPN12) for therapeutic intervention in acute myocardial infarction. Cardiovasc Res. 2019;116(5):1032–46.
99. Li G, Xu Y, Hu X, Li N, Yao R, Yu T, et al. Compound heterozygous variants of the COG6 gene in a Chinese patient with deficiency of subunit 6 of the conserved oligomeric Golgi complex (COG6-CDG). Eur J Med Genet. 2019;62(1):44–6.
100. Nitschke L, Tewari A, Coffin SL, Xhako E, Pang K, Gennarino VA, et al. miR760 regulates ATXN1 levels via interaction with its 5′ untranslated region. Genes Dev. 2020;34(17–18):1147–60.
101. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience. 2019;8(7):giz082.
102. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protocol. 2020;15(9):2759–72.

Publisher's Note