scientific reports

OPEN

Development and validation of an RNA-seq-based transcriptomic risk score for asthma

Xuan Cao[1], Lili Ding[2] & Tesfaye B. Mersha[3]✉

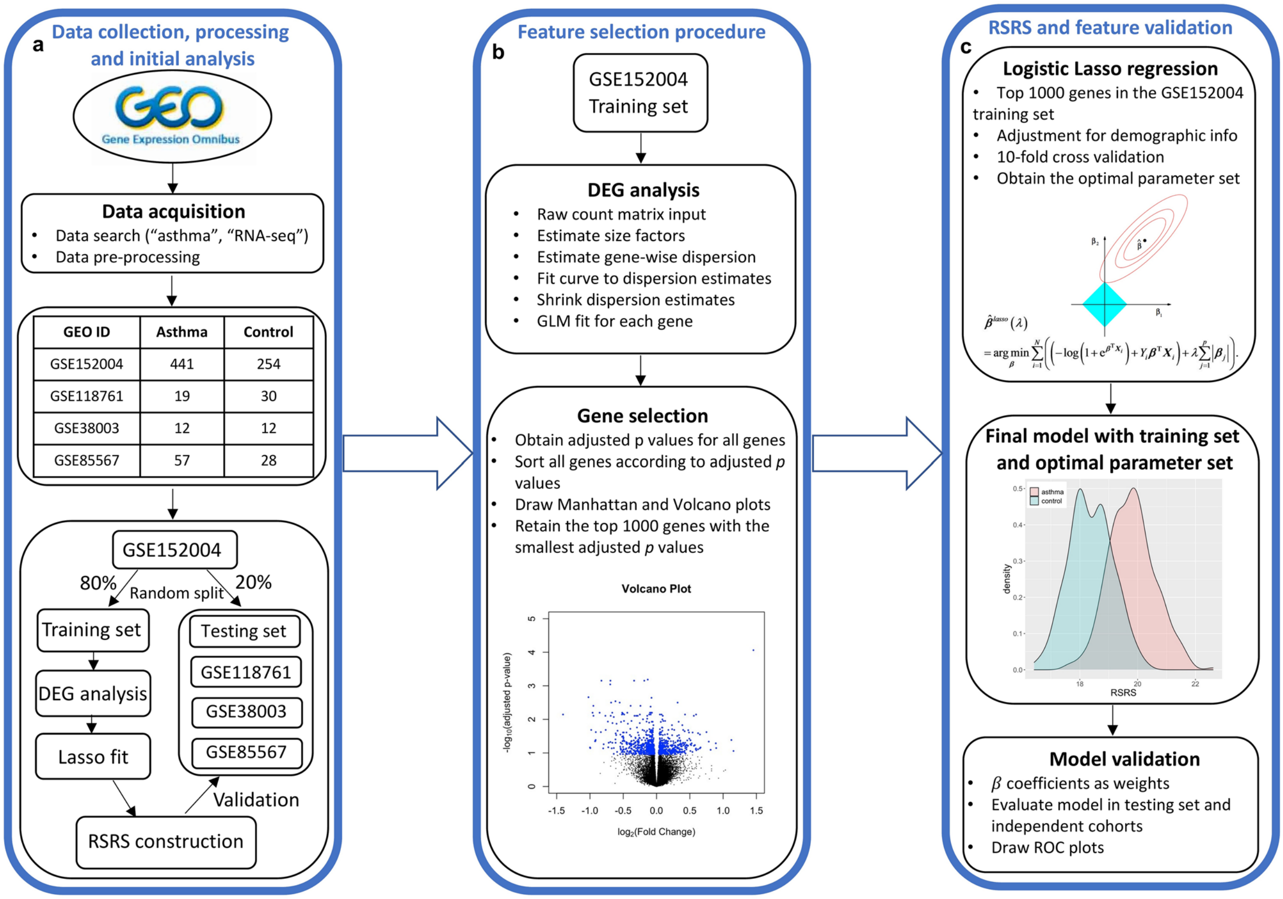
Recent progress in RNA sequencing (RNA-seq) allows us to explore whole-genome gene expression profiles and to develop predictive model for disease risk. The objective of this study was to develop and validate an RNA-seq-based transcriptomic risk score (RSRS) for disease risk prediction that can simultaneously accommodate demographic information. We analyzed RNA-seq gene expression data from 441 asthmatic and 254 non-asthmatic samples. Logistic least absolute shrinkage and selection operator (Lasso) regression analysis in the training set identified 73 differentially expressed genes (DEG) to form a weighted RSRS that discriminated asthmatics from healthy subjects with area under the curve (AUC) of 0.80 in the testing set after adjustment for age and gender. The 73-gene RSRS was validated in three independent RNA-seq datasets and achieved AUCs of 0.70, 0.77 and 0.60, respectively. To explore their biological and molecular functions in asthma phenotype, we examined the 73 genes by enrichment pathway analysis and found that these genes were significantly ($p < 0.0001$) enriched for DNA replication, recombination, and repair, cell-to-cell signaling and interaction, and eumelanin biosynthesis and developmental disorder. Further in-silico analyses of the 73 genes using Connectivity map shows that drugs (mepacrine, dactolisib) and genetic perturbagens (PAK1, GSR, RBM15 and TNFRSF12A) were identified and could potentially be repurposed for treating asthma. These findings show the promise for RNA-seq risk scores to stratify and predict disease risk.

Genome-wide association studies (GWAS) have been used to identify individual variants influencing disease risk[1]. However, most complex diseases are influenced by several loci, each with a small effect on its own, and polygenic approaches that group individual variants collectively influence a phenotypic trait offer a more predictive value than is possible by single variant approaches[2]. The most popular polygenic approach is polygenic risk score (PRS). PRS is defined as weighted sums of risk alleles of a pre-specified set of single nucleotide polymorphisms (SNPs)[2,3]. Weights in PRS are typically defined by estimated effect sizes of the SNPs and determined externally from independent studies, but there are also approaches that accommodate cases where appropriate external weights are not available, and internal weights from within the study are adopted instead[4–6]. PRS has proven to be statistically powerful for testing marginal genetic effects[7] and gene-environment interaction effects[8], and for predicting risks of complex disease[9].

Studies are now adapting PRS approaches to transcriptomics data[10–14]. A 20-gene microarray gene expression-based risk score was built to predict the overall survival and risk classification for patients with chronic lymphocytic leukemia[10]. Zhu et al.[13] developed a microarray expression-based risk score with 16 survival-associated autophagy-related genes for prognostic assessment of multiple myeloma. With technological advancement, RNA-seq becomes a more unbiased profiling method for the entire transcriptome than microarray platform. Compared with microarray analysis, RNA-seq can detect novel transcripts, quantify expression over a wider dynamic range, and detect rare and low-abundance transcripts[15–18].

Asthma has been recognized as a systemic disease consisting of networks of genes showing inflammatory changes that involve a broad spectrum of adaptive and innate immune systems. Utilizing measurable characteristics including gene expression can help to stratify asthma patients and develop strategies to predict asthma

[1]Division of Statistics and Data Science, Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, USA. [2]Division of Biostatistics and Epidemiology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH, USA. [3]Division of Asthma Research, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH, USA. ✉email: tesfaye.mersha@cchmc.org


a
Data collection, processing and initial analysis
GEO
Gene Expression Omnibus
Data acquisition
• Data search ("asthma", "RNA-seq")
• Data pre-processing
GSE152004
80%
Random split
20%
Training set
DEG analysis
Lasso fit
Testing set
GSE118761
GSE38003
GSE85567
Validation
RSRS construction
b
Feature selection procedure
GSE152004 Training set
DEG analysis
• Raw count matrix input
• Estimate size factors
• Estimate gene-wise dispersion
• Fit curve to dispersion estimates
• Shrink dispersion estimates
• GLM fit for each gene
Gene selection
• Obtain adjusted p values for all genes
• Sort all genes according to adjusted p values
• Draw Manhattan and Volcano plots
• Retain the top 1000 genes with the smallest adjusted p values
Volcano Plot
-log10(adjusted p-value)
0
1
2
3
4
5
log2(Fold Change)
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
c
RSRS and feature validation
Logistic Lasso regression
• Top 1000 genes in the GSE152004 training set
• Adjustment for demographic info
• 10-fold cross validation
• Obtain the optimal parameter set
β_2
$\hat{\beta}$
β_1
$\hat{\beta}^{lasso}(\lambda)$
$= \arg\min_{\beta} \sum_{i=1}^{N}\left(\left(-\log\left(1+e^{\beta^T X_i}\right)+Y_i\beta^T X_i\right)+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right).$
Final model with training set and optimal parameter set
asthma
control
density
0.0
0.1
0.2
0.3
0.4
0.5
RSRS
18
20
22
Model validation
• β coefficients as weights
• Evaluate model in testing set and independent cohorts
• Draw ROC plots

GEO ID	Asthma	Control
GSE152004	441	254
GSE118761	19	30
GSE38003	12	12
GSE85567	57	28

Figure 1. Study workflow for constructing the RSRS containing the steps of data acquisition and analysis. (**a**) Public data collection, processing and initial data analysis; (**b**) feature selection pipeline including DEG analysis and gene selection; (**c**) RSRS formulation and model validation in the testing set and independent cohorts.

severity and risk[19]. Castro-Rodriguez et al.[20] developed clinical data based asthma predictive index (API). Belsky et al.[21] derived a PRS based on multi-locus profiling from published GWAS. Recently, our group developed the pediatric asthma risk score (PARS) algorithm that integrates clinical and demographic factors[22]. PARS showed improvement compared with previous tools such as API[23]. However, both are based on clinical data and did not incorporate biological information including transcriptomic data.

The objectives of this study were to uncover differentially expressed signature genes between asthmatic and healthy controls using RNA-seq data and to identify an optimal subset of signature genes to construct a RNA-seq-based risk score (RSRS) that allows the accommodation of demographic factors including age and gender. The performance of RSRS was evaluated in an independent RNA-seq dataset. Additionally, we explored the DEGs in various publicly available resources such as asthma GWAS catalog and further determine their biological roles with enrichment analysis. Finally, we applied Connectivity Map (CMap) analysis approach to explore potential drug targets by systematically mining functional connections between asthma, RSRS, and perturbagens[24].

Materials and methods

To construct an RSRS for asthma, RNA-seq datasets were obtained from two independent studies[25,26] downloaded from the publicly available NCBI GEO (Gene Expression Omnibus, NCBI) database. Individuals in discovery dataset were randomly split into training and testing sets. DEGs were determined in the training set according to genome-wide adjusted p values. The selected DEGs were used to construct the RSRS in training dataset using logistic Lasso regression and develop a prediction model. The final model was tested in the testing set and validated in the other independent datasets. Datasets and analysis steps were summarized in Fig. 1.

RNA sequencing (RNA-seq) datasets.

Eligible GEO RNA-seq asthma datasets were selected based on the following inclusion criteria: (a) the dataset must compare asthma patients to non-asthma controls, and (b) the dataset must be generated from same tissue type. Four asthma RNA-seq studies (accession: GSE152004, GSE118761, GSE38003 and GSE85567)[25–28] fulfilled the inclusion criteria were used in the subsequent analysis. The following information was extracted from each study: (1) GEO accession numbers, (2) asthma and healthy status of the samples, (3) demographic information, and (4) count data from sequencing reads. In addition, demographic information including age and gender was obtained for GSE152004 and GSE118761. The raw count datasets were downloaded from GREIN (GEO RNA-seq Experiments Interactive Navigator)[29]. There were 695 individuals in the GSE152004 dataset including 441 asthmatic subjects and 254 non-asthmatic subjects[26]. RNA-seq data in GSE152004 was used for discovery of DEGs and model training and testing, while GSE118761 with 19 asthma subjects and 30 controls[25], GSE38003 with 12 asthmatics and 12 controls[28], and GSE85567 with 57 asthmatics and 28 controls[27] were used for model validation. The statistical software R was used to extract the expression values of individual genes and demographic information of different samples in asthma and control groups.

Data pre-processing. Normalized datasets were download from GREIN (GEO RNA-seq Experiments Interactive Navigator)[29], which uses the normalization method of Trimmed Mean of M-values (TMM) as implemented in the Bioconductor package edgeR[30]. DESeq2 R package was used to parse RNA-seq data and store the read counts in the form of a matrix of integer values[31]. Pre-filtering was performed to keep only genes that have at least 10 reads total.

RNA-seq-based risk score (RSRS) development and validation. The RNA-seq data GSE152004 was randomly split into 80% as training and 20% as testing while maintaining the same asthma: control ratio. DEG selection and risk score development were solely based on the training set. Three independent RNA-seq asthma datasets (GSE118761, GSE38003 and GSE85567) were used for model validation.

Constructing RNA-seq-based risk score (RSRS). Analogous to the polygenetic risk score (PRS)[32,33], RSRS of the ith individual was constructed as weighted sum of the individual's transformed and normalized RNA-seq expression values of K identified genes ($g_{i1}^{\log}, g_{i2}^{\log}, \ldots, g_{iK}^{\log}$),

$$RSRS_i = \beta_1 g_{i1}^{log} + \beta_2 g_{i2}^{log} + \cdots + \beta_K g_{iK}^{log}.$$

Here, $\beta_k, k = 1, 2, \ldots, K$ denotes weights and $g_{ik}^{\log} = \log(g_{ik} + 1)$ for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$ denotes normalized and log-transformed RNA-seq expression values. As the normalized counts could contain null values, we shifted them by one before log-transforming.

The log-transformed RNA-seq gene expression values are typically distributed more symmetrically than before transformation and have fewer extreme values compared to the untransformed data[34].

Variable selection for RSRS. To identify the signature genes to be included in the RSRS and estimate the corresponding weights of those genes, DEG analysis will be conducted on the training data to select a candidate set of DEGs, followed by Lasso logistic regression for further variable selection and weight estimation.

DEG analysis. Differential expression analysis of the training set was conducted using the DESeq2 R package. We extracted DEG results including log2 fold changes, p values and adjusted p values for all genes, where the p values were attained by the Wald test and corrected for multiple testing using the Benjamini and Hochberg method[35]. Manhattan plot of the genome-wide DEG analysis results was composed via the ggbio R package[36], where genes were annotated using the biomaRt R package from the Bioconductor project and mapped to corresponding chromosome locations[37,38].

Lasso logistic regression. The 1000 genes with the smallest adjusted p values in training set, as well as demographic factors such as age and sex, were included in Lasso logistic regression to select the optimal subset of DEGs to construct RSRS. To compare with the 1000 gene list, we also considered the DEG list with the top genes ranked by fold change and adjusted p value less than 0.05[39,40]. The same Lasso logistic regression was applied to the resulting gene list. Lasso is a penalized regression approach that performs variable selection and regularization by maximizing the log-likelihood function with the constraint that the sum of the absolute values of the coefficients is less than or equal to some positive constant[41]. Lasso logistic regression was carried out using the glmnet R package[42] with tenfold cross validation on the training set to select the optimal parameters. Age and sex were included in Lasso logistic regression to control the potential impact of demographic information on the disease risk. Genes with non-zero estimated beta weights were the optimal subset of features used to construct the RSRS.

RSRS and prediction model for disease risk. Given the optimal tuning parameters and the optimal subset of DEGs and demographic factors with non-zero beta weights identified in Lasso logistic regression with tenfold cross validation, an RSRS and prediction model for disease risk was developed in the whole training set using a logistic regression model. For the i th individual with a binary disease outcome $y_i = 1, 0$, $i = 1, 2, \ldots, n$, we consider the following logistic regression model,

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + RSRS_i + \boldsymbol{X_i}\boldsymbol{\beta_{K+1}} = \beta_0 + \beta_1 g_{i1}^{log} + \beta_2 g_{i2}^{log} + \cdots + \beta_K g_{iK}^{log} + \boldsymbol{X_i}\boldsymbol{\beta_{K+1}}.$$

Here μ represents the probability of having the disease, i.e., $\mu = P\left(y_i = 1\right) = \frac{\exp\left(\beta_0 + RSRS_i + \boldsymbol{X_i}\boldsymbol{\beta_{K+1}}\right)}{1+\exp\left(\beta_0 + RSRS_i + \boldsymbol{X_i}\boldsymbol{\beta_{K+1}}\right)}$ and β_0 is the intercept. Any demographic factors that had non-zero estimated coefficient in Lasso logistic regression will be included in the model by the term $\boldsymbol{X_i}\boldsymbol{\beta_{K+1}}$, where $\boldsymbol{X_i}$ is the matrix of demographic data, and $\boldsymbol{\beta_{K+1}}$ is the vector of regression coefficients to be estimated from the whole training set. The exponential function e^{β_k} of the kth regression coefficient β_k is the odds ratio associated with a one-unit increase in the log of the normalized gene expression count for the kth gene[43,44]. The estimated regression coefficient $\widehat{\beta_k}(k = 1, 2, \ldots, K)$ is the estimated weight for the kth gene in the RSRS.

Pairwise correlation. To investigate if the K signature DEGs retained in the RSRS provide independent information to asthma risk, we visualized the pair-wise Pearson correlation among the normalized and log-transformed gene expression levels of the selected genes by plotting the heat map using the corrplot R package.

Model validation. Finally, the prediction model with RSRS and demographic information was both tested in the testing dataset and validated in the independent sample GSE118761. RSRS without age and gender was implemented to predict the disease risk in the independent cohorts GSE38003 and GSE85567. Different prediction models based on the DEG list ranked by fold change with adjusted p value < 0.05, and the top 10, 50, 100 genes ranked by p value were also formulated and compared with the prediction models with RSRS utilizing the genes and demographic factors selected by Lasso. The confidence interval (CI) for the area under the receiver operating characteristics curve (ROC) was deduced based on the covariance matrix derived from generalized U-statistics[45] using an accelerated algorithm[46]. Both the ROC curves and AUC values were implemented in the R package pROC[47]. AUC was used to compare the models. We used the R package cutpointr[48] to estimate the optimal cut points that maximizes the Youden-Index[49] for determining the binary disease outcome and validate performance using bootstrapping.

Pathway and network analyses of RSRS. Ingenuity pathway analysis (IPA) software (Qiagen, USA) was used to generate putative networks and pathways based on the manually curated knowledge database of pathway interactions. The networks were generated using the genes retained in the RSRS after Lasso logistic regression in both direct and indirect relationships/connectivity. These networks were ranked by scores that measured the probability that the genes were included in the network beyond chance[50]. Canonical pathways associated with input genes were elucidated with a ratio to examine pathway enrichment and statistical significance adjusted for multiple testing. The p value is calculated using a right-tailed Fisher Exact test and indicates the likelihood of the pathway association under the random model.

Linking asthma RSRS with asthma GWAS catalog datasets. There were 168 asthma studies resulting 2811 SNPs from GWAS Catalog (Hindorff et al. 2009) (accessed January 2022). Inclusion of asthma GWAS catalog-based associations was limited to those studies with p values of less than 5 × 10E−8 (http://www.ebi.ac.uk/gwas/). Overlap between genes from RSRS after Lasso logistic regression and asthma genes from GWAS catalog were examined.

Connectivity Map (CMap) analysis. Next, we used Connectivity Map (CMap) analysis approach to explore potential drugs targeting asthma by systematically mining functional connections between asthma disease, RSRS, and perturbagens[24]. CMap currently covers > 1309 compounds connected with 7000 expression profiles[51]. This approach can identify drugs that affect the RNA-seq-based risk score and perturbagens that might potentially reverse asthma risk[52]. DEGs retained in the RSRS were divided into two groups, one for upregulated and the other for downregulated group. The CMap analysis was performed through the web interface CLUE (https://clue.io/), a cloud-based platform used to analyze perturbation-driven gene signatures, following a standard protocol described by Wang et al.[53,54]. CMap instance was measured by an enrichment score, which ranged from – 1 to 1, and a permutation p value. Connectivity score of below -0.85 or above 0.85 was considered for this analysis[54].

Results

Gene selection based on DEG analysis. Genome-wide DEG analysis results of the training set were visualized using volcano plot and MA plot (Fig. 2). The Manhattan plot of DEGs was given in Fig. S1. We retained 1000 genes with smallest adjusted p values in the training set.

Predictive value of RSRS. Of the 1000 genes, 73 genes were selected by logistic Lasso regression with tenfold cross validation (Fig. 2a). Weights, log2 fold changes and adjusted p values from logistic regression models of the 73 genes on the whole training set were provided in Table S1. As shown in Table S1, the odds ratios of 33 genes including SIK1, RAB3A, KRT76, UBE2V1 among others are larger than 1. Therefore, these genes can increase the odds and show an up-regulated effect on the outcome of asthma, while 40 genes including CNBP, POLL, ZNF696, HUS1 among others impose a down-regulated effect on the disease response. In Fig. 3, we generated the heat map of the pair-wise Pearson correlation among these 73 genes in the training data and the result of principal component analysis (PCA) was shown in Fig. S2. The results showed that the 73 genes are not correlated and provide independent information to asthma risk. The gender factor also contributed to the disease risk. In particular, girls would be at a higher risk for asthma compared with boys.

The density distributions of RSRS for asthma vs. healthy controls in both the training and testing sets are depicted in Fig. 4. Note that a higher value of RSRS indicates a higher probability for asthma. For both the asthmatic and control groups, the density distribution resembles mostly a unimodal distribution with thin tails, which indicates only a few individuals have higher probabilities of developing asthma compared with others.

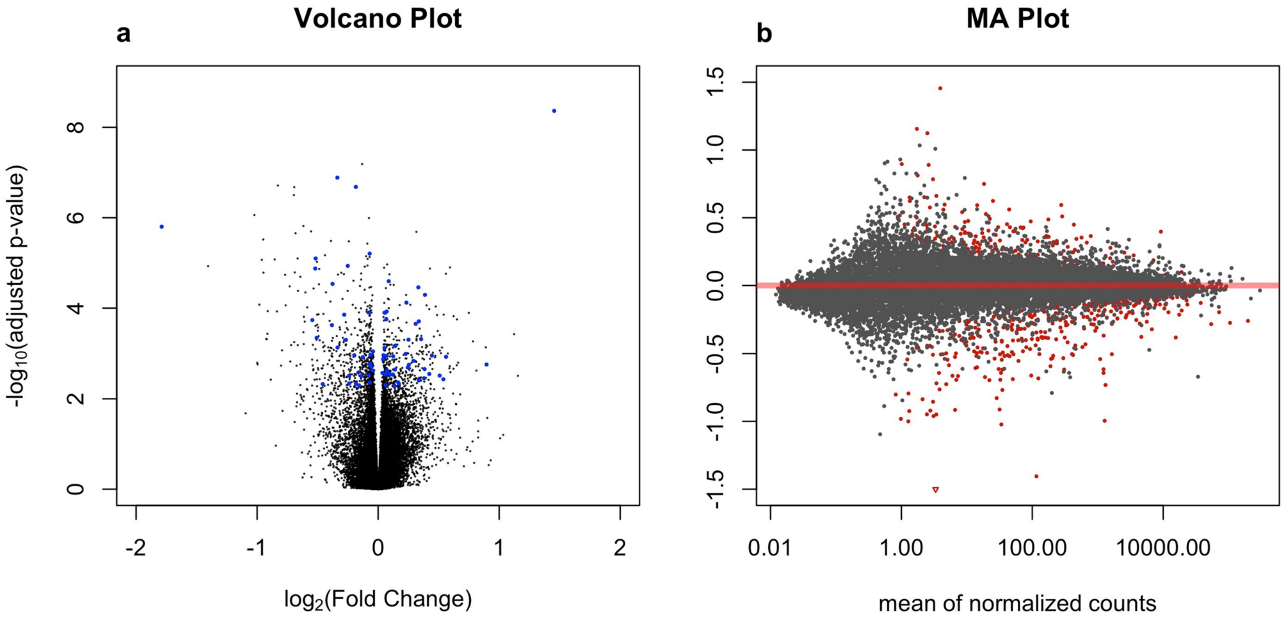


Figure 2. Initial quality checking of the RNA-seq data based on the training set. (**a**) Volcano plot of − $\log_{10}$ adjusted p values (on the y-axis) versus $\log_2$ fold changes (on the x-axis) using the training set. The blue points correspond to the 73 RSRS genes. (**b**) MA plot, which is a scatter plot of log2 fold changes (on the y-axis) versus the mean of normalized counts (on the x-axis), where points were colored red if the genes were selected. Points which fall out of the window were plotted as open triangles pointing either up or down.

Testing set. The ROC curves and AUC values in the testing set corresponding to different risk scores utilizing different numbers of top ranked genes were given in Fig. 5a. The RSRS based on 73 genes with adjustment for demographic information achieved the highest AUC of 0.80 (CI 0.73–0.88) compared with other risk scores based on the DEG list ranked by fold change with adjusted p value < 0.05, and the top 10, 50, 100 genes ranked by p value. The optimal cut point for thresholding the probabilities obtained from the prediction model is 0.54, which yields an accuracy of 0.76. Figure S3 lists various distribution plots for cutoff points and out-of-bag performance based on bootstrapping.

Validation dataset. For the independent validation dataset GSE38003, the 73-gene-based RSRS and sex achieved an AUC of 0.70 (95% CI 0.55–0.85) (Fig. 5b). The optimal cut point for thresholding the probabilities yielded an accuracy of 0.61. Figure S4 lists various distribution plots for cutoff points and out-of-bag performance based on bootstrapping for GSE118761. For GSE38003 and GSE85567, the 73-gene-based RSRS achieved AUCs of 0.77 (95% CI 0.58–0.97) (Fig. 5c) and 0.60 (95% CI 0.47–0.74) (Fig. 5d), respectively. The optimal cut point for thresholding the probabilities yielded an accuracy of 0.67 for GSE38003 and an accuracy of 0.60 for GSE85567. Figures S5 and S6 list various distribution plots for cutoff points and out-of-bag performance based on bootstrapping for GSE38003 and GSE85567, respectively.

Enrichment analysis of RSRS for network and pathways.

The biological and molecular functions of the 73 RSRS genes were examined for enriched pathways using the Ingenuity Pathway Analysis system.

Using the Ingenuity Pathway Analysis of the 73 RSRS genes, we found six and five enriched networks and pathways, respectively (score ≥ 2). The functions of the top networks and pathways are shown in Table S2 and Fig. S7. Enriched diseases and functions involve organismal injury and abnormalities, developmental disorder, and connective tissue disorders.

Overlap of genes between RSRS and asthma GWAS catalog.

Overlap between the 73 RSRS genes and asthma GWAS catalog is given in Fig. 6. There were seven genes identified by both RSRS and asthma GWAS catalog. Among these, each of gene CPS1 and gene NOL4 was mapped to 2 SNPs that belong to functional classes including the intronic variant and intergenic variant. Gene ZFP36L1 was mapped to 2 SNPs belonging to functional classes including the intronic variant and regulatory region variant, while each of SIK1, UQCC2, GBA3 and UBE2D3 was involved with one intronic SNP. Detailed information including p values are provided in Table S3.

Connectivity map identifies potential asthma target signature.

Using publicly available perturbagens, we identified drug targets related RSRS. The identified perturbagens (genetic or chemical) were primarily associated with immune function, cellular transport, regulation of transcription and inflammation. The perturbagens associated with asthma are shown in Table S4.

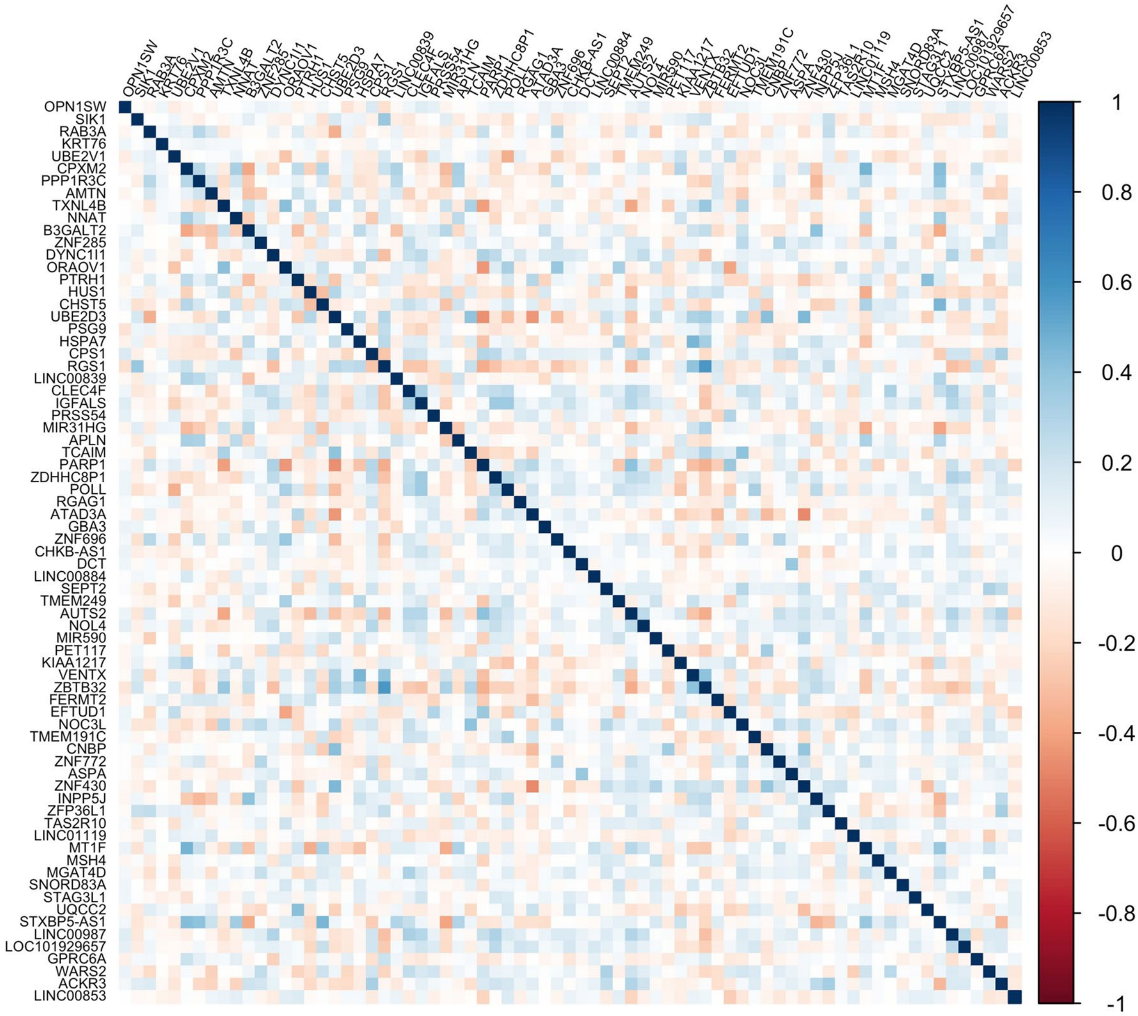


Figure 3. The heat map of the pair-wise Pearson correlation among the normalized and log-transformed gene expression values of 73 genes.

Discussion

To the best of our knowledge, current risk prediction methods that are implemented using the combination of variants into polygenetic risk scores (PRS) did not take the opportunity to incorporate RNA-seq and demographic information. The current work filled this gap by constructing the RAN-seq analog of PRS, RNA-seq-based transcriptomic risk score (RSRS), based differential gene expression risk scores with adjustment for age and gender. In particular, we developed a novel 73-gene RNA-seq based score for predicting asthma risk by determining RSRS in a sample size of 695. We further validated the risk score in an independent validation set and obtained high predictive accuracy based on the AUC value. These findings demonstrate the RNA-seq based gene profiling score has the potential to support clinical diagnosis and to achieve satisfactory prediction accuracy for asthma.

Compared to PRSs, RSRSs have many advantages. Transcriptomics bridges the gap between genotype and phenotype, is physiologically closer to phenotype, and links genetic associations to biological mechanism. Therefore, RSRS should provide biologically tractable prediction that could potentially outperforms PRSs. Marigorta et al. showed that transcriptomic risk scores (TRSs) outperform PRSs in distinguishing Crohn's disease from healthy samples[55]. Using data on 17 quantitative traits in UK Biobank, Liang et al. found that prediction accuracy of TRS was significantly higher than PRS in the African samples relative to the European reference set[56]. Other advantages of RSRS as a gene-level risk score include more biologically interpretable than SNP-level PRS, smaller and more manageable number of features, which in turn requires smaller samples to train prediction models[56].

Several of the 73 RSRS genes were already linked with asthma. Leukocyte transcriptomes analysis from preschool children with acute wheeze identified genes including UBE2D3 involved that were significantly enriched in the innate immune responses[57]. Yucesoy et al.[58] identified a novel Locus (18q12.1, NOL4) linked with diisocyanate-Induced occupational asthma via a genome-wide association study. Dysregulation of ZFP36L1/L2 in severe asthma epithelium was found to contribute to glucocorticoid non-responsiveness as well as epithelial barrier disruption[59]. The salt-inducible kinases (SIKs) are required for producing cytokines that regulate airway hyper-responsiveness, immune cell infiltration and inflammation[60]. Genome-wide significant loci (6p21.31, UQCC2) was identified by cross-trait meta-analysis associated with asthma and ADHD. Several top-ranking

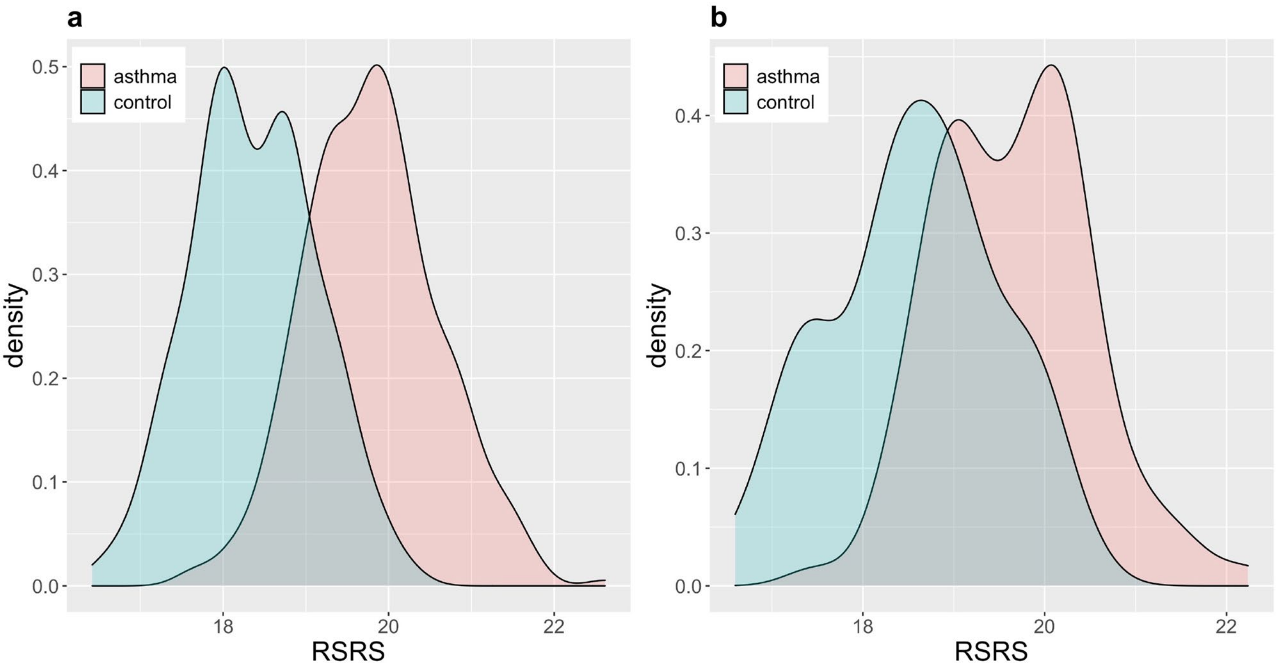


Figure 4. Density plots showing the asthma risk based on RSRS for different groups. (**a**) In training set; (**b**) testing set.

genetic perturbagens including PAK1, GSR, RBM15 and TNFRSF12A have been indicated by previous studies to be involved in allergy/asthma[61–63]. In particular, TNFRSF12A belongs to the family of Tumour necrosis factor (TNF) receptor. Current studies suggest that tumor necrosis factor (TNF)-α found in asthmatic airways, may directly alter the contractile properties of the airway smooth muscle and lead to the development of bronchial hyper-responsiveness[64]. Preliminary studies have demonstrated when treated with anti-TNF-α therapy, asthma patients showed an improvement in lung function, and airway hyperresponsiveness and a decrease in exacerbation frequency[65]. Our results also suggest several chemical perturbagens including mepacrine (cytokine production inhibitor) and dactolisib, WYE-125132, AZD-8055 (MTOR inhibitor). Cytokines are responsible for initiating the early stages of asthma and play a critical role in the persistence of the chronic inflammatory process in asthma because of many pro-inflammatory effect characteristics produced by cytokines[66,67]. Steroid-dependent asthma patients treated with Th2 cytokine inhibitor showed improvement in their pulmonary function and symptom control, and became less dependent on the inhaled corticosteroid[68]. Studies have also shown that patients experiencing an asthma attack showed significantly elevated serum MTOR pathway activation compared with patients in asthma remission, which suggests potential targets of MTOR inhibitor for treating asthma[69]. The results clearly demonstrated that the connectivity scores of perturbagens have role in RSRS. Our result also suggested the importance of age in predicting childhood asthma risk. Multiple studies have shown gender heterogeneity in the prevalence of asthma[70,71]. As children, boys have been consistently reported to have more asthma incidence than girls[70], which was also confirmed in our study. As adults, compared with men, women exhibit an increased prevalence and severity of asthma[70]. Important factors in the gender heterogeneity related to asthma onset and severity include sex hormones, social and environmental factors, and responses to asthma therapeutics[70,72]. To examine the effect of gender-specific differences in changes of asthma prevalence, larger sample size and multi-omics should be investigated[70,71].

Our study has some limitations. First, when the model includes several highly correlated variables, Lasso tends to pick only one or a few of them and shrinks the effect of the rest to 0[73]. For future studies, the elastic-net[74] or adaptive Lasso[75] may be adopted to alleviate the possible limitation of Lasso. Second, information provided in the public-domain gene expression data are limited and we only have access to the demographic information (age and gender) but no clinical data. However, even without clinical data, we were able to develop an RSRS with significant prediction accuracy for RNA-seq based data. One could anticipate when more demographic information, clinical parameters and symptoms (such as race, wheezing and polysensitization[22]) become available, the prediction power will be further improved. Third, the current datasets and sample sizes are somewhat small. Recently, efforts are made to generate big multi-omics data which could be incorporated in risk prediction. Further analysis of RNA-seq data with a large sample size taking social and environmental exposures into account will provide the opportunity to improve the accuracy of RSRS in predicting asthma risk.

In summary, using RNA-seq data, Logistic Lasso regression identified 73 gene-based RSRS that is predictive of asthma risk with AUC of 0.80. Our findings reveal the potential of RSRS for asthma risk and generated a new set of pathways and networks that may assist in defining genetic signatures linked with asthma. The pathways affected in our data can provide deeper insight in diseases mechanisms and to identify the most critical genes and drug or chemical that can be used to perturb this mechanism. Our RSRS method offers new statistical methodology to develop risk scores based on transcriptomic data in complex diseases.

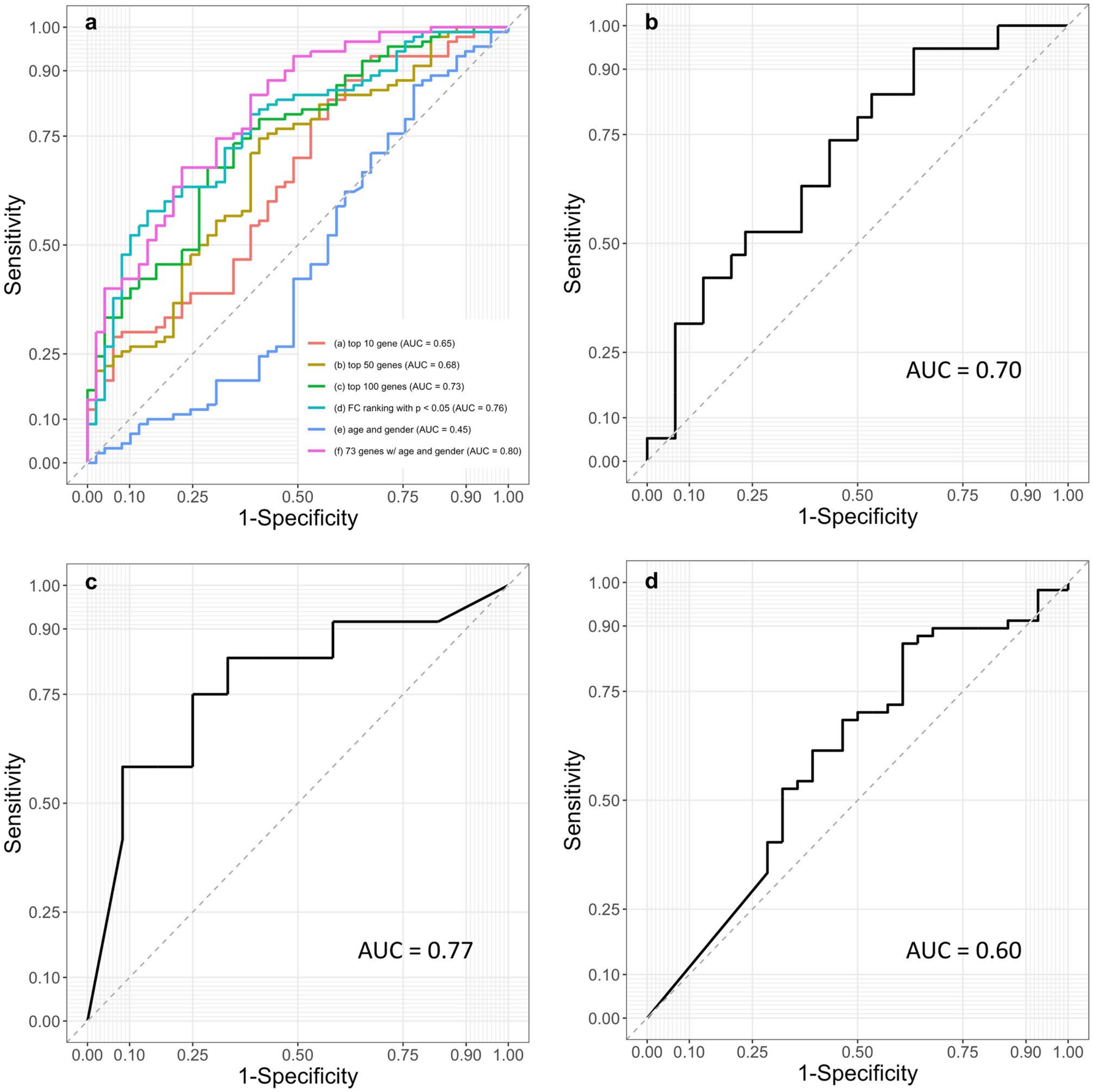


Figure 5. ROC curves for the asthma prediction performance of RSRS. (**a**) testing set and comparison with risk scores based on the DEG list ranked by fold change (FC) with $p < 0.05$, and the top 10, 50,100 genes ranked by p value; (**b**) in the independent cohort GSE118761 (AUC = 0.70); (**c**) in the independent cohort GSE38003 (AUC = 0.77); (**d**) in the independent cohort GSE85567 (AUC = 0.60).

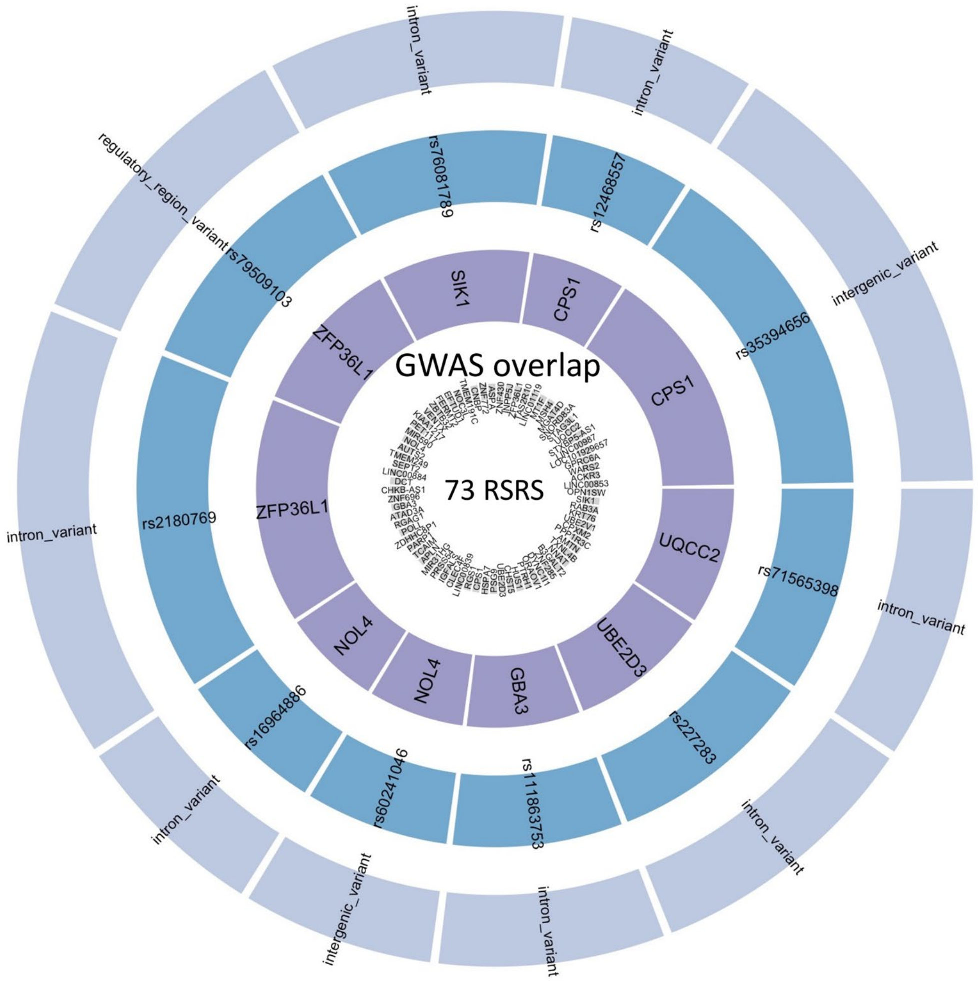


Figure 6. The overlap between the 73 RSRS genes and asthma genome wide association study catalog. The sector width for the SNP is proportional to the – log10 (adjusted p value) corresponding to each SNP.

Data availability

The data supporting this work is publicly available from NCBI GEO (Gene Expression Omnibus): https://www.ncbi.nlm.nih.gov/gds/?term=asthma.

Received: 15 February 2022; Accepted: 4 May 2022
Published online: 23 May 2022

References

1. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121. https://doi.org/10.1038/s41588-018-0147-3 (2018).
2. Huls, A. & Czamara, D. Methodological challenges in constructing DNA methylation risk scores. *Epigenetics* **15**, 1–11. https://doi.org/10.1080/15592294.2019.1644879 (2020).
3. Wray, N. R. *et al.* Research review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55**, 1068–1087. https://doi.org/10.1111/jcpp.12295 (2014).
4. Hüls, A., Ickstadt, K., Schikowski, T. & Krämer, U. Detection of gene-environment interactions in the presence of linkage disequilibrium and noise by using genetic risk scores with internal weights from elastic net regression. *BMC Genet.* **18**, 55. https://doi.org/10.1186/s12863-017-0519-1 (2017).
5. Huls, A. *et al.* Comparison of weighting approaches for genetic risk scores in gene-environment interaction studies. *BMC Genet.* **18**, 115. https://doi.org/10.1186/s12863-017-0586-3 (2017).
6. Martin, A. R., Daly, M. J., Robinson, E. B., Hyman, S. E. & Neale, B. M. Predicting polygenic risk of psychiatric disorders. *Biol. Psychiatry* **86**, 97–109. https://doi.org/10.1016/j.biopsych.2018.12.015 (2019).
7. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776. https://doi.org/10.1038/s41467-019-09718-5 (2019).
8. Meisner, A., Kundu, P. & Chatterjee, N. Case-only analysis of gene-environment interactions using polygenic risk scores. *Am. J. Epidemiol.* **188**, 2013–2020. https://doi.org/10.1093/aje/kwz175 (2019).
9. Sun, J. *et al.* Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nat. Commun.* **12**, 5276. https://doi.org/10.1038/s41467-021-25014-7 (2021).
10. BouSamra, E., Klein, B., Commes, T. & Moreaux, J. Identification of a 20-gene expression-based risk score as a predictor of clinical outcome in chronic lymphocytic leukemia patients. *Biomed. Res. Int.* **2014**, 423174. https://doi.org/10.1155/2014/423174 (2014).

11. Cho, J. Y. *et al.* Gene expression signature-based prognostic risk score in gastric cancer. *Clin. Cancer Res.* **17**, 1850–1857. https://doi.org/10.1158/1078-0432.CCR-10-2180 (2011).
12. Chu, J., Li, N. & Li, F. A risk score staging system based on the expression of seven genes predicts the outcome of bladder cancer. *Oncol. Lett.* **16**, 2091–2096. https://doi.org/10.3892/ol.2018.8904 (2018).
13. Zhu, F. X., Wang, X. T., Zeng, H. Q., Yin, Z. H. & Ye, Z. Z. A predicted risk score based on the expression of 16 autophagy-related genes for multiple myeloma survival. *Oncol. Lett.* **18**, 5310–5324. https://doi.org/10.3892/ol.2019.10881 (2019).
14. Kim, S. K. *et al.* A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol. Oncol.* **8**, 1653–1666. https://doi.org/10.1016/j.molonc.2014.06.016 (2014).
15. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.* **10**, 4706. https://doi.org/10.1038/s41467-019-12464-3 (2019).
16. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63. https://doi.org/10.1038/nrg2484 (2009).
17. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32**, 926–932. https://doi.org/10.1038/nbt.3001 (2014).
18. Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* **9**, e78644. https://doi.org/10.1371/journal.pone.0078644 (2014).
19. Carr, T. F. & Bleecker, E. Asthma heterogeneity and severity. *World Allergy Organ. J.* **9**, 41. https://doi.org/10.1186/s40413-016-0131-2 (2016).
20. Castro-Rodriguez, J. A. The asthma predictive index: A very useful tool for predicting asthma in young children. *J. Allergy Clin. Immunol.* **126**, 212–216. https://doi.org/10.1016/j.jaci.2010.06.032 (2010).
21. Belsky, D. W. *et al.* Polygenic risk and the development and course of asthma: An analysis of data from a four-decade longitudinal study. *Lancet Respir. Med.* **1**, 453–461. https://doi.org/10.1016/S2213-2600(13)70101-2 (2013).
22 Biagini Myers, J. M. *et al.* A pediatric asthma risk score to better predict asthma development in young children. *J. Allergy Clin. Immunol.* **143**, 1803–1810. https://doi.org/10.1016/j.jaci.2018.09.037 (2019).
23. Castro-Rodriguez, J. A., Holberg, C. J., Wright, A. L. & Martinez, F. D. A clinical index to define risk of asthma in young children with recurrent wheezing. *Am. J. Respir. Crit. Care Med.* **162**, 1403–1406. https://doi.org/10.1164/ajrccm.162.4.9912111 (2000).
24. Lamb, J. *et al.* The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935. https://doi.org/10.1126/science.1132939 (2006).
25. Kicic, A. *et al.* Assessing the unified airway hypothesis in children via transcriptional profiling of the airway epithelium. *J. Allergy Clin. Immunol.* **145**, 1562–1573. https://doi.org/10.1016/j.jaci.2020.02.018 (2020).
26. Jackson, N. D. *et al.* Single-cell and population transcriptomics reveal pan-epithelial remodeling in type 2-high asthma. *Cell Rep.* **32**, 107872. https://doi.org/10.1016/j.celrep.2020.107872 (2020).
27. Nicodemus-Johnson, J. *et al.* DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight*. https://doi.org/10.1172/jci.insight.90151 (2016).
28. Yick, C. Y. *et al.* Gene expression profiling of laser microdissected airway smooth muscle tissue in asthma and atopy. *Allergy* **69**, 1233–1240. https://doi.org/10.1111/all.12452 (2014).
29. Mahi, N. A., Najafabadi, M. F., Pilarczyk, M., Kouril, M. & Medvedovic, M. GREIN: An interactive web platform for re-analyzing GEO RNA-seq data. *Sci. Rep.* **9**, 7580. https://doi.org/10.1038/s41598-019-43935-8 (2019).
30. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140. https://doi.org/10.1093/bioinformatics/btp616 (2010).
31. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. https://doi.org/10.1186/s13059-014-0550-8 (2014).
32. Prive, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787. https://doi.org/10.1093/bioinformatics/bty185 (2018).
33. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348. https://doi.org/10.1371/journal.pgen.1003348 (2013).
34. Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **9**, e85150. https://doi.org/10.1371/journal.pone.0085150 (2014).
35. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).
36. Yin, T., Cook, D. & Lawrence, M. ggbio: An R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13**, R77. https://doi.org/10.1186/gb-2012-13-8-r77 (2012).
37. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191. https://doi.org/10.1038/nprot.2009.97 (2009).
38. Durinck, S. *et al.* BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440. https://doi.org/10.1093/bioinformatics/bti525 (2005).
39. SEQC Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.* **32**, 903–914. https://doi.org/10.1038/nbt.2957 (2014).
40. Guo, L. *et al.* Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.* **24**, 1162–1169. https://doi.org/10.1038/nbt1238 (2006).
41. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x (1996).
42. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
43. Hoppe, F. M., Hoppe, D. J. & Walter, S. D. Explaining odds ratios as conditional risk ratios. *J. Clin. Epidemiol.* **97**, 123–124. https://doi.org/10.1016/j.jclinepi.2017.10.009 (2018).
44. Szumilas, M. Explaining odds ratios. *J. Can. Acad. Child Adolesc. Psychiatry* **19**, 227–229 (2010).
45. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845 (1988).
46. Sun, X. & Xu, W. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **21**, 1389–1393. https://doi.org/10.1109/LSP.2014.2337313 (2014).
47. Robin, X. *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **12**, 77. https://doi.org/10.1186/1471-2105-12-77 (2011).
48. Thiele, C. & Hirschfeld, G. cutpointr: Improved estimation and validation of optimal cutpoints in R. *J. Stat. Softw.* **98**, 1–27. https://doi.org/10.18637/jss.v098.i11 (2021).
49. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35. https://doi.org/10.1002/1097-0142(1950)3:1%3c32::aid-cncr2820030106%3e3.0.co;2-3 (1950).
50. Raponi, M. *et al.* Microarray analysis reveals genetic pathways modulated by tipifarnib in acute myeloid leukemia. *BMC Cancer* **4**, 56. https://doi.org/10.1186/1471-2407-4-56 (2004).
51. Wang, L., Yu, Y., Yang, J., Zhao, X. & Li, Z. Dissecting Xuesaitong's mechanisms on preventing stroke based on the microarray and connectivity map. *Mol. Biosyst.* **11**, 3033–3039. https://doi.org/10.1039/c5mb00379b (2015).

52. Ravindranath, A. C. *et al.* Connecting gene expression data from connectivity map and in silico target predictions for small molecule mechanism-of-action analysis. *Mol. Biosyst.* **11**, 86–96. https://doi.org/10.1039/c4mb00328d (2015).
53. Wang, Y., Yella, J. & Jegga, A. G. Transcriptomic data mining and repurposing for computational drug discovery. *Methods Mol. Biol.* **1903**, 73. https://doi.org/10.1007/978-1-4939-8955-3_5 (1903).
54 Subramanian, A. *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452. https://doi.org/10.1016/j.cell.2017.10.049 (2017).
55. Marigorta, U. M. *et al.* Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat. Genet.* **49**, 1517–1521. https://doi.org/10.1038/ng.3936 (2017).
56. Liang, Y. *et al.* Polygenic transcriptome risk scores improve portability of polygenic risk scores across ancestries. *BioRxiv.* https://doi.org/10.1101/2020.11.12.373647 (2020).
57 Katayama, S. *et al.* Acute wheeze-specific gene module shows correlation with vitamin D and asthma medication. *Eur. Respir. J.* **55**, 1901330. https://doi.org/10.1183/13993003.01330-2019 (2020).
58. Yucesoy, B. *et al.* Genome-wide association study identifies novel loci associated with diisocyanate-induced occupational asthma. *Toxicol. Sci.* **146**, 192–201. https://doi.org/10.1093/toxsci/kfv084 (2015).
59. Rynne, J. *et al.* Dysregulation of ZFP36L1 and ZFP36L2 in asthma alters epithelial integrity and genome-wide glucocorticoid responses. *ERJ Open Res.* **7**, 111. https://doi.org/10.1183/23120541.Lsc-2021.111 (2021).
60. Darling, N. J., Arthur, J. S. C. & Cohen, P. Salt-inducible kinases are required for the IL-33-dependent secretion of cytokines and chemokines in mast cells. *J. Biol. Chem.* **296**, 100428. https://doi.org/10.1016/j.jbc.2021.100428 (2021).
61. Lu, M. *et al.* Inhibition of p21-activated kinase 1 attenuates the cardinal features of asthma through suppressing the lymph node homing of dendritic cells. *Biochem. Pharmacol.* **154**, 464–473. https://doi.org/10.1016/j.bcp.2018.06.012 (2018).
62. Polonikov, A. V. *et al.* Antioxidant defense enzyme genes and asthma susceptibility: Gender-specific effects and heterogeneity in gene-gene interactions between pathogenetic variants of the disease. *Biomed. Res. Int.* **2014**, 708903. https://doi.org/10.1155/2014/708903 (2014).
63. Dai, B. *et al.* Significance of RNA N6-methyladenosine regulators in the diagnosis and subtype classification of childhood asthma using the gene expression omnibus database. *Front. Genet.* https://doi.org/10.3389/fgene.2021.634162 (2021).
64. Amrani, Y., Chen, H. & Panettieri, R. A. Jr. Activation of tumor necrosis factor receptor 1 in airway smooth muscle: A potential pathway that modulates bronchial hyper-responsiveness in asthma? *Respir. Res.* **1**, 49–53. https://doi.org/10.1186/rr12 (2000).
65. Berry, M., Brightling, C., Pavord, I. & Wardlaw, A. TNF-alpha in asthma. *Curr. Opin. Pharmacol.* **7**, 279–282. https://doi.org/10.1016/j.coph.2007.03.001 (2007).
66. Chung, K. F. & Barnes, P. J. Cytokines in asthma. *Thorax* **54**, 825–857. https://doi.org/10.1136/thx.54.9.825 (1999).
67. Lambrecht, B. N., Hammad, H. & Fahy, J. V. The cytokines of asthma. *Immunity* **50**, 975–991. https://doi.org/10.1016/j.immuni.2019.03.018 (2019).
68. Tamaoki, J. *et al.* Effect of suplatast tosilate, a Th2 cytokine inhibitor, on steroid-dependent asthma: A double-blind randomised study. Tokyo Joshi-Idai Asthma Research Group. *Lancet* **356**, 273–278. https://doi.org/10.1016/s0140-6736(00)02501-0 (2000).
69. Zhang, Y. *et al.* Activation of the mTOR signaling pathway is required for asthma onset. *Sci. Rep.* **7**, 4532. https://doi.org/10.1038/s41598-017-04826-y (2017).
70. Almqvist, C. *et al.* Impact of gender on asthma in childhood and adolescence: A GA2LEN review. *Allergy* **63**, 47–57. https://doi.org/10.1111/j.1398-9995.2007.01524.x (2008).
71. Postma, D. S. Gender differences in asthma development and progression. *Gend. Med.* **4**(Suppl B), S133–S146. https://doi.org/10.1016/s1550-8579(07)80054-4 (2007).
72. Chowdhury, N. U., Guntur, V. P., Newcomb, D. C. & Wechsler, M. E. Sex and gender in asthma. *Eur. Respir. Rev.* https://doi.org/10.1183/16000617.0067-2021 (2021).
73. Wang, S., Nan, B., Rosset, S. & Zhu, J. Random lasso. *Ann. Appl. Stat.* **5**, 468–485. https://doi.org/10.1214/10-AOAS377 (2011).
74. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **67**, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x (2005).
75. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429. https://doi.org/10.1198/016214506000000735 (2006).

Acknowledgements

This work was supported by the National Institutes of Health (NIH) Grants (R01HL132344, R01HG011411, U01AI159087 and U19AI070235), the Simons Foundation's collaboration Grant (No. 635213) and the Charles Phelps Taft Research Center at the University of Cincinnati.

Author contributions

X.C., L.D., and T.B.M. conceived and designed the study. X.C., L.D., and T.B.M. wrote the main manuscript text. X.C. performed the statistical analyses. All authors reviewed and revised the manuscript prior to submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12199-0.

Correspondence and requests for materials should be addressed to T.B.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.