Integrated gene-based and pathway analyses using UK Biobank data identify novel genes for chronic respiratory diseases

Lijuan Wang
Nanjing Medical University
Meng Zhu
Nanjing Medical University
Na Qin
Nanjing Medical University
Yuzhuo Wang
Nanjing Medical University
Jingyi Fan
Nanjing Medical University
Qi Sun
Nanjing Medical University
Mengmeng Ji
Nanjing Medical University
Xikang Fan
Nanjing Medical University
Junxing Xie
Nanjing Medical University
Hongxia Ma
Nanjing Medical University
Juncheng Dai (djc@njmu.edu.cn)
Nanjing Medical University https://orcid.org/0000-0002-3909-5671

Research article

Keywords: Chronic respiratory diseases, gene-based analysis, functional annotation, pathway enrichment analysis

Posted Date: November 6th, 2019

DOI: https://doi.org/10.21203/rs.2.16894/v1

Version of Record: A version of this preprint was published at Gene on January 1st, 2021. See the published version at https://doi.org/10.1016/j.gene.2020.145287.

Abstract

Background Chronic respiratory diseases have become a nonnegligible cause of death globally. Although smoking and environmental exposures are primary risk factors for chronic respiratory diseases, genetic factors also play an important role in determining individual's susceptibility to diseases. Here we performed integrated gene-based and pathway analyses to systematically illuminate the heritable characteristics of chronic respiratory diseases. Methods UK (United Kingdom) Biobank is a very large, population-based prospective study with over 500,000 participants, established to allow detailed investigations of the genetic and nongenetic determinants of the diseases. Utilizing the GWAS-summarized data downloaded from UK Biobank, we conducted gene-based analysis to obtain associations of susceptibility genes with asthma, chronic obstructive pulmonary disease (COPD) and pneumonia using FUSION and MAGMA softwares. Across the identified susceptibility regions, functional annotation integrating multiple functional data sources was performed to explore potential regulatory mechanisms with INQUISIT algorithm. To further detect the biological process involved in the development of chronic respiratory diseases, we undertook pathway enrichment analysis with the R package (clusterProfiler). Results A total of 195 susceptibility genes were identified significantly associated with chronic respiratory diseases (P bonferroni <0.05), and 28/195 located out of known susceptibility regions (e.g. WDPCP in 2p15). Within the identified susceptibility regions, functional annotation revealed an aggregation of credible variants in promoter-like and enhancer-like histone modification regions and such regulatory mechanisms were specific to lung tissues. Furthermore, 110 genes with INQUISIT score ≥1 may influence diseases susceptibility through exerting effects on coding sequences, proximal promoter and distal enhancer regulations. Pathway enrichment results showed that these genes were enriched in immune-related processes and nicotinic acetylcholine receptors pathways. Conclusions This study conducted an integrated gene-based and pathway strategy to explore the underlying biological mechanisms and our findings may serve as promising targets for future clinical treatments of chronic respiratory diseases.

Background

With an aging global population, chronic respiratory diseases are growing up to be a more prominent cause of death and disability [1]. According to the data from the Global Burden of Diseases (GBD) 2017, chronic respiratory diseases have caused 3.91 million deaths, accounting for 15.8% of all-aged deaths globally [2]. Among chronic respiratory diseases, chronic obstructive pulmonary disease (COPD) and asthma are common obstructive diseases characterized by persistent airflow limitation and decline of lung function [3]. Infectious lung diseases are mainly accompanied by a massively activated inflammatory response [4]. Although smoking and environmental exposures are primary risk factors for chronic respiratory diseases, genetic factors also play an important role in determining individual's susceptibility to diseases [5-7], which could inform drug target identification, risk prediction, and stratified prevention or treatment.

Previous genome-wide association studies (GWASs) have identified dozens of variants associated with chronic respiratory diseases [8-10]. However, genome-wide significant loci only account for a small proportion of the genetic variants, which is insufficient to dissect the complex genetic structure. Besides, a single SNP typically has only mild effects while the common diseases are often influenced by the joint effects of multiple loci within a gene or the joint action of multiple genes within a pathway [11]. Thus, by integrating the effects of a group of genetic variants, the gene-based and pathway enrichment analyses can help us holistically unravel the mechanisms of complex diseases [12].

Recently, alternative approaches were developed to perform the gene-based analysis. One way was to integrate functional data with GWAS association results to explore the underlying biological mechanisms. For example, Gusev *et al.* introduced a method, referred as transcriptome-wide association study (TWAS), integrating gene expression data with large scale GWAS data to estimate the association of each gene to disease [13]. An additional approach was to aggregate variants to the level of the whole gene and detect the joint association of all variants in the gene with the phenotype. For instance, VEGAS performs permutation based simulation [14, 15], MAGMA employs multiple linear regression [16] and Pascal computes sum and maximum of chi-squared statistics [17] to obtain gene-based *P*-values. In this work, we carried out the gene-based analysis using TWAS strategy and MAGMA software to identify novel susceptibility genes for chronic respiratory diseases. Then, the identified genes were utilized to perform pathway enrichment analysis to explore the potential biological process. Our study provides important insight into a deeper understanding of the pathogenesis involved in chronic respiratory diseases.

Methods

Study design and data preparation

UK Biobank

The UK Biobank (UKB) cohort is a major data resource that contains genetic as well as a wide range of phenotypic data of ~500,000 participants of European ancestry aged 39-73 years at recruitment [18]. Genotyping was conducted using the Affymetrix UK BiLEVE Axiom or Affymetrix UK Biobank Axiom array. These arrays were augmented by imputation of ~90 million genetic variants from the 1000 Genomes and UK10K projects, Haplotype Reference Consortium. Detailed information of cohorts, genotyping, imputation, quality control approaches, and association analysis please refer to the published studies [19, 20].

Phenotype selection

We paid attention to the susceptibility mechanisms modified by genetic mutations for chronic lower respiratory diseases. Phenotypes with more than 5,000 cases were selected for subsequent analyses, including asthma with 28,628 cases, COPD with 9,266 cases and pneumonia with 9,774 cases. All

individuals clinically defined from hospital episode statistics were coded as cases, while all other individuals were considered as controls [20].

GWAS summary datasets

We downloaded the summarized data of genome-wide imputed variants from GeneALTAS website (http://geneatlas.roslin.ed.ac.uk). We included variants with the quality score of imputation >0.9 and 13,324,371 SNPs remained. We further performed the quality control based on European population from 1000 Genomes Project (Phase 3) with the following criteria: (i) variants having a minor allele frequency (MAF) >0.01; (ii) variants deviated from the Hardy-Weinberg equilibrium (P>1.0×10^{-6}). After the procedure of quality control, 7,330,104 variants were finally retained for the following analysis. The study design is shown in **Figure 1**.

Genetic correlations

We calculated the genetic correlations (r_g) between diseases using linkage disequilibrium score regression (LDSC), which requires only GWAS summarized statistics and is not biased by sample overlap [21]. LDSC quantifies the extent to which two phenotypes share genetic etiology based on the patterns of LD found across the genome (https://github.com/bulik/ldsc). A conservative Bonferroni-corrected method was used to determine significant correlations ($P_{bonferroni}$<0.05).

Gene-based analysis

TWAS analysis using FUSION

We downloaded FUSION [13] software (http://gusevlab.org/projects/fusion/) along with its prepackaged weights for gene expression data in lung tissue derived from the Genotype-Tissue Expression Project (GTEx). FUSION estimated the heritability of genes explained by cis-SNPs (SNPs within 1 Mb region surrounding the TSS) and restricted TWAS analysis to include cis-heritable genes (P<0.01). Then, the effect sizes of cis-SNPs on gene expression (i.e. expression weights) were estimated using predictive linear models (Elastic Net, LASSO, GBLUP, and BSLMM). For each gene, FUSION estimated the z-score of the expression and a complex trait (Z_{TWAS}) as a linear combination of the vector of GWAS summary Z scores at a given cis-locus with weight vector W derived from the reference panels. However, the imputed z-score of expression and trait (WZ) has variance WVW^t, where V is a covariance matrix across SNPs at the locus (i.e., LD), as:

$$Z_{TWAS} = \frac{WZ}{\sqrt{WVW^t}}$$

MAGMA's gene association test

The SNP-based *P* values were used for gene-based analysis using MAGMA [16] software (http://ctg.cncr.nl/software/magma), a novel tool for gene and gene-set analysis. Total 19,427 protein-coding genes from the database (NCBI 37.3) were used for SNP annotation. Then, MAGMA used a multiple regression approach to properly incorporate LD between markers and to detect multi-marker effects for a genome-wide gene association analysis. For both two gene-based methods, we applied a stringent Bonferroni correction to account for multiple testing and associations with $P_{bonferroni}$ <0.05 were considered as statistically significant.

Functional annotation

Functional enrichment of defined CRVs

We first defined credible risk variants (CRVs) as SNPs located within 500 kb upstream or downstream of the associated genes and with *P* values within two orders of magnitude of the most significant SNP in each locus. To investigate the enrichment of CRVs in chromatin regulatory regions, Fisher's exact test was used to estimate the distribution of the above CRVs in active promoter and enhancer regions defined in NHLF and A549 cell types by calculating the fold-enrichment against the background of 1,000 genomes (other SNPs in the defined locus). Chromatin state data in four human cell types of other tissues (HSMM, HESC, NHEK, and GM12878) were also included in our analysis for comparison. All the histone modifications of promoter and enhancer marks (H3K4me3, H3K9ac, H3K4me1, and K3K27ac) were downloaded from the UCSC Genome Browser.

Functional annotation using INQUISIT algorithm

To further detect the regulation mechanisms underlying identified genes, we performed functional annotation with the INQUISIT algorithm [22]. We calculated a score for each gene by assessing the potential impact of each CRV on regulatory (proximal or distal gene regulation) or coding features. The INQUISIT score was contributed with multiple lines of evidence including Hi-C chromatin interaction information, enhancer-target gene predictions, topologically associated domains, histone modification marks, transcription factor binding sites and expression quantitative trait loci (eQTLs) in lung-related tissues. Details of the algorithm and scoring strategy have been described previously elsewhere [22].

Functional exploration for the best GWAS SNP

To investigate the genetic effects of variants on gene regulation, we conducted functional annotation for the best hit in each novel susceptibility region. The best hit was referred to the most significant GWAS SNP within cis-locus (500 kb upstream or downstream of the susceptibility gene). We performed annotations for variants within promising genes using ANNOVAR software [23]. The functional effects of missense variations were predicted using the SIFT [24] and PolyPhen [25] databases. To investigate the potential function of the association at non-coding regions, we utilized data from the Genotype-Tissue Expression (GTEx, http://www.gtexportal.org/, version 7) to perform the expression quantitative trait loci (eQTL) analyses in 383 lung tissue samples. To further map the variants to potential regulatory elements,

we annotated SNPs according to the histone Chip-seq (H3K27AC, H3K4ME1, H3K4ME3) peaks and DNase I hypersensitivity sites (DHS) from ENCODE Project Consortium (http://genome.ucsc.edu/ENCODE). We downloaded these features measured in NHLF cell lines from the UCSC genome browser (http://genome.ucsc.edu/).

Pathway enrichment analysis

We performed pathway enrichment analysis on genes with integrated score ≥1 defined by functional annotation to further explore the biological process. The analysis was conducted with the combined genes of three traits as input considering the high genetic correlations between traits. We used the Reactome Pathway Database [26] as a reference, which was implemented in R package "clusterProfiler" [27]. Bonferroni method was used for multiple correction and pathways with adjusted *P*-value <0.05 were retained.

Results

Genetic correlation

To examine the genetic correlation among asthma, COPD and pneumonia, we used LDSC to calculate genetic correlations with GWAS summarized statistics obtained from UKB. We discovered significant overall genetic correlations of asthma with COPD (r_g=0.70, P=1.20×10^{-76}), asthma with pneumonia (r_g=0.74, P=3.47×10^{-5}) and COPD with pneumonia (r_g=1.00, P=1.25×10^{-5}, **Supplementary Table 1**). We also assessed genetic correlations of smoking behavior with the above three traits, as expected, we identified significant correlations of asthma with smoking (r_g=0.23, P=3.61×10^{-18}), COPD with smoking (r_g=0.59, P=9.36×10^{-99}), and pneumonia with smoking (r_g=0.74, P=3.38×10^{-06}), suggesting that the correlation between diseases may partially due to smoking.

Gene-based analysis identified significant susceptibility genes

In the gene-based analysis, we totally identified 221 significant associations for the three respiratory diseases at Bonferroni *P*-value <0.05, including 195 unique genes (**Supplementary Table 2**). Of these genes, 28/195 are located in novel susceptibility regions that independent from those identified by GWAS or candidate gene strategies (**Table 1**). We compared TWAS significant associations with those from MAGMA and found that 30/45 could be verified with nominal *P*-value <0.05, which further supported the validity of two gene-based methods. Manhattan plots of genes as well as SNPs associations are presented in **Figure 2**.

Approximately half of asthma associated genes reside in the MHC region, indicating the crucial role of immune response in asthma pathogenesis (**Table 1**). Nicotinic acetylcholine receptors including *CHRNA5*, *CHRNA3,* and *CHRNB4* at 15q25 locus, were found significantly linked to COPD susceptibility, suggesting that genetic variants integrating environmental exposures such as smoking contributed to the development of COPD. Besides, *IREB2* in this locus was also identified as a susceptibility gene for COPD

(P=2.94×10^{-13}). However, *IREB2* was the only protein-coding gene reached the gene association significance (P=1.17×10^{-6}) for pneumonia.

Functional annotation

We first defined a set of credible candidate variants (CRVs) in the identified loci and annotated these variants with publicly available genomic data. Then, we systematically evaluated these CRVs for evident of enrichment of genomic features such as histone modification marks. Interestingly, we observed significant over-representation of these CRVs in promoter-like (H3K4me3 and H3K9ac) and enhancer-like (H3K4me1 and H3K27ac) histone modification regions, and such enrichment was especially identified in lung (NHLF) or lung cancer (A549) related cell types (**Figure 3, Supplementary Table 3**).

We applied the INQUISIT functional annotation strategy to further detect the regulatory mechanisms underlying identified genes by evaluating the impact of CRVs on coding sequences, proximal promoter and distal enhancer regulations in each gene locus. Across the identified susceptibility regions, coding impact evaluation aligned CRVs to two genes, proximal regulatory gene mapping matched CRVs to 80 genes and distal regulatory gene mapping annotated CRVs to 103 genes. This resulted in 110 unique mapped genes, and only *TSPAN8* was implicated by all three mapping strategies (**Supplementary Table 4**).

To unravel the potential regulatory mechanisms in novel susceptibility regions, we performed functional annotation for 19 best GWAS SNPs (**Supplementary Table 5**). The best GWAS SNP rs11178649 at 12q21.1, was significantly associated with the expression of *TSPAN8* in 383 GTEx lung tissues (P=5.40×10^{-3}). Interestingly, we identified a missense variant rs3763978, which was in strong LD (r^2=0.99) with rs11178649, was predicted as 'probably damaging' by PolyPhen (PolyPhen-score 0.999) and had a 'deleterious' SIFT-score of 0.033 (**Supplementary Table 6**). The most significant SNP rs2084200 near *WDPCP* was confirmed to be an eQTL variant for *WDPCP* in lung tissue based on GTEx database (P=2.20×10^{-5}, **Supplementary Figure 1**). Additionally, variants in high LD (r^2>0.8) with rs2084200 reside in the promoter region of *CDH1*, where exhibits strong interaction with *WDPCP* identified by ChIA-PET. The tag SNP rs2416984 at 9q33.3 showed significant association with *PBX3* expression (P=1.80×10^{-7}, **Supplementary Figure 2**), with its related SNPs (r^2>0.8) mainly located in the histone modification marks targeting both promoters and enhancers in A549 and NHLF cell types. Although the best GWAS SNP rs17484235 was marginally correlated with the expression of *IREB2* (P=0.038, **Supplementary Figure 3**), rs17484235 as well as its related variants present strong regulatory signals by targeting the *IREB2* promoter.

Pathway enrichment analysis

To further explore the biological pathways involved in the process of chronic respiratory diseases, we performed pathway enrichment analysis with 110 functional annotated genes defined by INQUISIT (integrated score ≥1). The result revealed the enrichment of 15 pathways (P_{adj}<0.05) that were involved

in the immune system and nicotinic acetylcholine receptors signaling, such as *PD-1* signaling (P_{adj}=1.74×10^{-8}), interferon gamma signaling (P_{adj}=2.56×10^{-5}) and presynaptic nicotinic acetylcholine receptors (P_{adj}=1.72×10^{-2}, **Figure 4**, **Supplementary Table 7**).

Discussion

UK Biobank is a very large and detailed prospective study with over 500,000 participants aged 39-73 years when recruited in 2006-2010. Utilizing the GWAS-summarized data downloaded from UK Biobank website, we performed gene-based analysis and identified 28 susceptibility genes within novel regions associated with risk of chronic respiratory diseases. Functional annotation revealed that the regulation of gene expression tended to be a common way influencing diseases susceptibility. Pathway enrichment analysis demonstrated that the implicated genes were mainly aggregated in immune-related processes and nicotinic acetylcholine receptors pathways, proving the important role of immune system and smoking behavior in respiratory diseases development. By utilizing an integrated gene and pathway strategy, our study systematically evaluated susceptibility gene and potential biological process for chronic respiratory diseases, which provides important insights into the etiology of chronic respiratory diseases.

Based on the results of gene-based analyses, we newly identified 21 susceptibility genes for asthma, 10 genes for COPD and only one gene for pneumonia. For asthma, *WDPCP* at 2p15 associated with the susceptibility of asthma. The expression of *WDPCP* is significantly regulated by the best GWAS SNP rs2084200 and its related variants within the cis-locus. *WDPCP* plays a critical role in regulating planar cell polarity and ciliogenesis by mediating septin localization [28]. As the ciliated epithelium that covers the surface of the airways forms an immunologically active natural barrier to invasion and injury, ciliary dysfunction could increase susceptibility to infection and inflammation and has been found as a feature of moderate to severe asthma [29]. Another asthma susceptibility gene, *TSPAN8*, encodes a cell surface glycoprotein that mediates signal transduction events in the regulation of cell development, activation and motility [30]. The lead SNP at the *TSPAN8* locus, rs11178649, was in complete LD with rs3763978 (r^2=0.99), a missense variant in the exon region of *TSPAN8*, which causes a glycine to alanine substitution and was predicted to be deleterious by SIFT and PolyPhen databases. While little is known about the mechanism of this gene underlying asthma development and further studies are warranted.

The COPD susceptibility gene, *PBX3*, is one member of TALE class homeodomain family that are implicated in developmental gene expression through their abilities to form hetero-oligomeric DNA-binding complexes and function as transcriptional regulators in numerous cell types [31]. The expression of *PBX3* was greatly influenced by the lead SNP rs2416984 as well as its related variants at the *PBX3* locus. By developing a *PBX3*-deficient mice model, Rhee JW et al identified that *PBX3* was essential for the proper development of medullary respiratory control mechanisms and mutations of *PBX3* may promote the pathogenesis of central hypoventilation [32]. Besides, Heguy A's research observed that *PBX3* was up-regulated in the alveolar macrophages of smokers compared to nonsmokers, indicating the

potential association between *PBX3* and cigarette smoking or COPD [33]. However, the biological mechanism involved in COPD pathogenesis remains unclear, which needs further investigations.

Consistent with previous study [34], *IREB2* was identified as a susceptibility gene for COPD in our study. Interestingly, we also observed an association between *IREB2* and pneumonia susceptibility. *IREB2* encodes an RNA binding protein that acts to maintain human cellular iron metabolism by modulating the expression of those proteins relevant to iron uptake, export, and sequestration [35]. *IREB2* variants are in tight LD with SNPs associated with nicotine addiction. As we all know, conventional cigarette smoking is an important factor for developing both COPD and pneumonia, and prior studies have demonstrated that smoking was associated with lung iron imbalance in pulmonary inflammation [36, 37], supporting a role for *IREB2* in the pathogenesis of COPD and pneumonia.

Formal studies have identified that a subset of current and former smokers develops an asthma-COPD overlap condition that is associated with gene expression markers of Th2 inflammation in the airway [38, 39]. Torres A *et al.* found that adults with chronic conditions and other risk factors such as COPD, asthma, and smoking are at increased risk of pneumonia [40], indicating shared pathogenic mechanism among chronic respiratory diseases. Consistent with previous findings, our study revealed high genetic correlations between chronic respiratory diseases, as well as an aggregation of susceptibility genes in immune-related processes and nicotinic acetylcholine receptors pathways. Furthermore, we conducted functional annotation to explore potential regulatory mechanisms and found that the credible variants defined in the identified susceptibility regions were primarily mapped to non-coding regions, and showed a strong over-representation in eQTLs, as well as functional regions such as histone modification marks. These results suggested that the variants contributed to the development of chronic respiratory diseases mainly through transcriptional regulation. More importantly, we proved that such genetic mechanisms were specific to lung tissues.

Conclusions

In this study, we applied an integrated gene-based and pathway enrichment strategy to explore the potential susceptibility genes and biological processes involved in the development of chronic respiratory diseases. Our study identified 28 genes within novel susceptibility regions that are statistically significantly associated with diseases, providing important insight into the genetic causes of diseases and giving suggestions to future clinical treatment. However, other respiratory traits with limited cases such as interstitial lung disease were not included, for the sample size may have an influence on the reliability of the results. Another limitation was that further exploration of the relationship between smoking status and genetic variant could not be performed due to the lack of smoking information for UKB samples. Therefore, additional analyses such as stratified or interaction analyses are needed to validate our findings, and functional experiments are warranted to unravel the biological mechanisms behind the associations.

Declarations

Availability of data and materials

All summary results from the analyses performed are available at the GeneATLAS website, http://geneatlas.roslin.ed.ac.uk/.

Abbreviations

GWAS: genome-wide association studies; TWAS: transcriptome-wide association study; MAF: minor allele frequencies; HWE: Hardy-Weinberg equilibrium; LDSC: linkage disequilibrium score regression; eQTL: expression quantitative trait loci.

Acknowledgements

We thank the study participants and research staffs for their contributions and commitment to this study.

Funding

This work was funded by the National Key Research and Development Program of China (2017YFC0907900 and 2017YFC0907905), National Natural Science of China (81820108028, 81521004 and 81803306), Science Foundation for Distinguished Young Scholars of Jiangsu (BK20160046), Natural Science Foundation of Jiangsu Province (BK20180675), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (18KJB330002), the Priority Academic Program for the Development of Jiangsu Higher Education Institutions [Public Health and Preventive Medicine] and Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (PPZY2015A067).

Author's contributions

Juncheng Dai designed this study. Analysis was conducted by Lijuan Wang, Meng Zhu, Na Qin and Yuzhuo Wang. All authors contributed to the interpretation of the analysis and to the direction of the discussion. Hongxia Ma and Juncheng Dai reviewed, edited and commented on multiple versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

All participants gave informed consent for data provision and linkage. The UK Biobank project was approved by the National Research Ethics Service Committee North West-Haydock (REC reference: 11/NW/0382).

Consent for publication

The authors confirm that this manuscript does not contain any personal data or images from any individual participants.

Competing interest

The authors declare no competing financial interest.

References

1. Collaborators GBDCRD: **Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015**. *The Lancet Respiratory medicine* 2017, **5**(9):691-706.
2. Collaborators GBDCoD: **Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017**. *Lancet* 2018, **392**(10159):1736-1788.
3. Postma DS, Rabe KF: **The Asthma-COPD Overlap Syndrome**. *The New England journal of medicine* 2015, **373**(13):1241-1249.
4. Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, Pinsky MR, Fine J, Krichevsky A, Delude RL, Angus DC *et al*: **Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the Genetic and Inflammatory Markers of Sepsis (GenIMS) Study**. *Archives of internal medicine* 2007, **167**(15):1655-1663.
5. Salvi SS, Barnes PJ: **Chronic obstructive pulmonary disease in non-smokers**. *Lancet* 2009, **374**(9691):733-743.
6. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland**. *The New England journal of medicine* 2000, **343**(2):78-85.
7. Duffy DL, Martin NG, Battistutta D, Hopper JL, Mathews JD: **Genetics of asthma and hay fever in Australian twins**. *The American review of respiratory disease* 1990, **142**(6 Pt 1):1351-1358.

8. Wain LV, Shrine N, Artigas MS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, Obeidat M, Henry AP, Portelli MA, Hall RJ *et al*: **Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets**. *Nature genetics* 2017, **49**(3):416-425.
9. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson W *et al*: **A large-scale, consortium-based genomewide association study of asthma**. *The New England journal of medicine* 2010, **363**(13):1211-1221.
10. Rautanen A, Mills TC, Gordon AC, Hutton P, Steffens M, Nuamah R, Chiche JD, Parks T, Chapman SJ, Davenport EE *et al*: **Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study**. *The Lancet Respiratory medicine* 2015, **3**(1):53-60.
11. Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, Henley JR, Rocca WA, Ahlskog JE, Maraganore DM: **A genomic pathway approach to a complex disease: axon guidance and Parkinson disease**. *PLoS genetics* 2007, **3**(6):e98.
12. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE: **Integrating pathway analysis and genetics of gene expression for genome-wide association studies**. *American journal of human genetics* 2010, **86**(4):581-591.
13. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA *et al*: **Integrative approaches for large-scale transcriptome-wide association studies**. *Nature genetics* 2016, **48**(3):245-252.
14. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Investigators A, Hayward NK, Montgomery GW, Visscher PM *et al*: **A versatile gene-based test for genome-wide association studies**. *American journal of human genetics* 2010, **87**(1):139-145.
15. Mishra A, Macgregor S: **VEGAS2: Software for More Flexible Gene-Based Testing**. *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2015, **18**(1):86-91.
16. de Leeuw CA, Mooij JM, Heskes T, Posthuma D: **MAGMA: generalized gene-set analysis of GWAS data**. *PLoS computational biology* 2015, **11**(4):e1004219.
17. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S: **Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics**. *PLoS computational biology* 2016, **12**(1):e1004714.
18. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J *et al*: **The UK Biobank resource with deep phenotyping and genomic data**. *Nature* 2018, **562**(7726):203-209.
19. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M *et al*: **UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age**. *PLoS medicine* 2015, **12**(3):e1001779.
20. Canela-Xandri O, Rawlik K, Tenesa A: **An atlas of genetic associations in UK Biobank**. *Nature genetics* 2018, **50**(11):1593-1599.

21. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, Patterson N, Daly MJ, Price AL, Neale BM: **LD Score regression distinguishes confounding from polygenicity in genome-wide association studies**. *Nature genetics* 2015, **47**(3):291-295.
22. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb D, Rostamianfar A *et al*: **Association analysis identifies 65 new breast cancer risk loci**. *Nature* 2017, **551**(7678):92-94.
23. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic acids research* 2010, **38**(16):e164.
24. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic acids research* 2003, **31**(13):3812-3814.
25. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nature methods* 2010, **7**(4):248-249.
26. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B *et al*: **Reactome: a database of reactions, pathways and biological processes**. *Nucleic acids research* 2011, **39**(Database issue):D691-697.
27. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters**. *Omics : a journal of integrative biology* 2012, **16**(5):284-287.
28. Cui C, Chatterjee B, Lozito TP, Zhang Z, Francis RJ, Yagi H, Swanhart LM, Sanker S, Francis D, Yu Q *et al*: **Wdpcp, a PCP protein required for ciliogenesis, regulates directional cell migration and cell polarity by direct modulation of the actin cytoskeleton**. *PLoS biology* 2013, **11**(11):e1001720.
29. Thomas B, Rutman A, Hirst RA, Haldar P, Wardlaw AJ, Bankart J, Brightling CE, O'Callaghan C: **Ciliary dysfunction and ultrastructural abnormalities are features of severe asthma**. *The Journal of allergy and clinical immunology* 2010, **126**(4):722-729 e722.
30. Nazarenko I, Rana S, Baumann A, McAlear J, Hellwig A, Trendelenburg M, Lochnit G, Preissner KT, Zoller M: **Cell surface tetraspanin Tspan8 contributes to molecular pathways of exosome-induced endothelial cell activation**. *Cancer research* 2010, **70**(4):1668-1678.
31. Monica K, Galili N, Nourse J, Saltman D, Cleary ML: **PBX2 and PBX3, new homeobox genes with extensive homology to the human proto-oncogene PBX1**. *Molecular and cellular biology* 1991, **11**(12):6149-6157.
32. Rhee JW, Arata A, Selleri L, Jacobs Y, Arata S, Onimaru H, Cleary ML: **Pbx3 deficiency results in central hypoventilation**. *The American journal of pathology* 2004, **165**(4):1343-1350.
33. Heguy A, O'Connor TP, Luettich K, Worgall S, Cieciuch A, Harvey BG, Hackett NR, Crystal RG: **Gene expression profiling of human alveolar macrophages of phenotypically normal smokers and nonsmokers reveals a previously unrecognized subset of genes modulated by cigarette smoking**. *Journal of molecular medicine* 2006, **84**(4):318-328.

34. DeMeo DL, Mariani T, Bhattacharya S, Srisuma S, Lange C, Litonjua A, Bueno R, Pillai SG, Lomas DA, Sparrow D *et al*: **Integration of genomic and genetic approaches implicates IREB2 as a COPD susceptibility gene**. *American journal of human genetics* 2009, **85**(4):493-502.
35. Rouault TA: **The role of iron regulatory proteins in mammalian iron homeostasis and disease**. *Nature chemical biology* 2006, **2**(8):406-414.
36. O'Brien-Ladner AR, Nelson SR, Murphy WJ, Blumer BM, Wesselius LJ: **Iron is a regulatory component of human IL-1beta production. Support for regional variability in the lung**. *American journal of respiratory cell and molecular biology* 2000, **23**(1):112-119.
37. Ghio AJ, Hilborn ED, Stonehuerner JG, Dailey LA, Carter JD, Richards JH, Crissman KM, Foronjy RF, Uyeminami DL, Pinkerton KE: **Particulate matter in cigarette smoke alters iron homeostasis to produce a biological effect**. *American journal of respiratory and critical care medicine* 2008, **178**(11):1130-1138.
38. Christenson SA, Steiling K, van den Berge M, Hijazi K, Hiemstra PS, Postma DS, Lenburg ME, Spira A, Woodruff PG: **Asthma-COPD overlap. Clinical relevance of genomic signatures of type 2 inflammation in chronic obstructive pulmonary disease**. *American journal of respiratory and critical care medicine* 2015, **191**(7):758-766.
39. Barnes PJ: **Immunology of asthma and chronic obstructive pulmonary disease**. *Nature reviews Immunology* 2008, **8**(3):183-192.
40. Torres A, Blasi F, Dartois N, Akova M: **Which individuals are at increased risk of pneumococcal disease and why? Impact of COPD, asthma, smoking, diabetes, and/or chronic heart disease on community-acquired pneumonia and invasive pneumococcal disease**. *Thorax* 2015, **70**(10):984-989.

Tables

Table 1. Significant genes in novel susceptibility regions identified by gene-based analysis.

Method	Trait	Region	Gene	Chr	Locus start	Locus end	Gene-based P value	Bonferroni P value
TWAS	Asthma	8p23.1	*FAM85B*	8	8025341	8084136	2.53E-06	1.96E-02
			FAM86B3P	8	8086117	8097552	5.65E-06	4.38E-02
		11q12.2	*FADS2*	11	61588853	61634826	1.04E-06	8.07E-03
	COPD	9q33.3	*PBX3*	9	128509624	128729656	8.91E-08	6.91E-04
MAGMA	Asthma	1p36.12	*LACTBL1*	1	23279536	23291831	2.48E-06	4.45E-02
		2p15	*WDPCP*	2	63348518	64054977	1.19E-06	2.14E-02
		3q26.32	*TBL1XR1*	3	176737143	176915261	1.04E-06	1.87E-02
		6q25.1	*ZC3H12D*	6	149768794	149806197	2.18E-06	3.92E-02
		7p22.3	*ADAP1*	7	940573	995043	4.69E-08	8.44E-04
		8p23.1	*MFHAS1*	8	8640864	8751155	9.95E-07	1.79E-02
		9q33.3	*DENND1A*	9	126145934	126692431	1.09E-06	1.96E-02
		10p12.31	*MLLT10*	10	21823094	22032559	4.47E-07	8.04E-03
		11q12.2	*FADS2*	11	61588853	61634826	1.14E-09	2.06E-05
			FADS1	11	61567099	61596790	6.15E-09	1.11E-04
			TMEM258	11	61535973	61558075	4.01E-08	7.20E-04
			MYRF	11	61520121	61555990	4.10E-07	7.37E-03
			FEN1	11	61562813	61564716	4.23E-07	7.61E-03
		12q21.1	*TSPAN8*	12	71518865	71835678	4.07E-07	7.31E-03
		13q12.11	*PSPC1*	13	20248896	20357142	9.61E-07	1.73E-02
		13q32.3	*UBAC2*	13	99853028	100038688	6.18E-10	1.11E-05
			GPR183	13	99946784	99959659	4.46E-07	8.01E-03
		14q32.12	*RIN3*	14	92980118	93155339	8.82E-12	1.59E-07
		17q21.32	*ZNF652*	17	47366568	47439478	2.34E-07	4.20E-03
	COPD	8p23.1	*XKR6*	8	10753555	11058875	1.13E-06	2.03E-02
			SOX7	8	10587706	10588022	1.34E-06	2.40E-02
			PINX1	8	10622473	10691291	1.39E-06	2.50E-02
		9q33.3	*PBX3*	9	128509624	128729656	2.03E-09	3.65E-05
		9q34.13	*MED27*	9	134735494	134955295	8.46E-08	1.52E-03
	Pneumonia	2p15	*AHSA2*	2	61404553	61413216	1.06E-06	1.91E-02
		15q25.1	*IREB2*	15	78729773	78793798	1.17E-06	2.11E-02

Figures

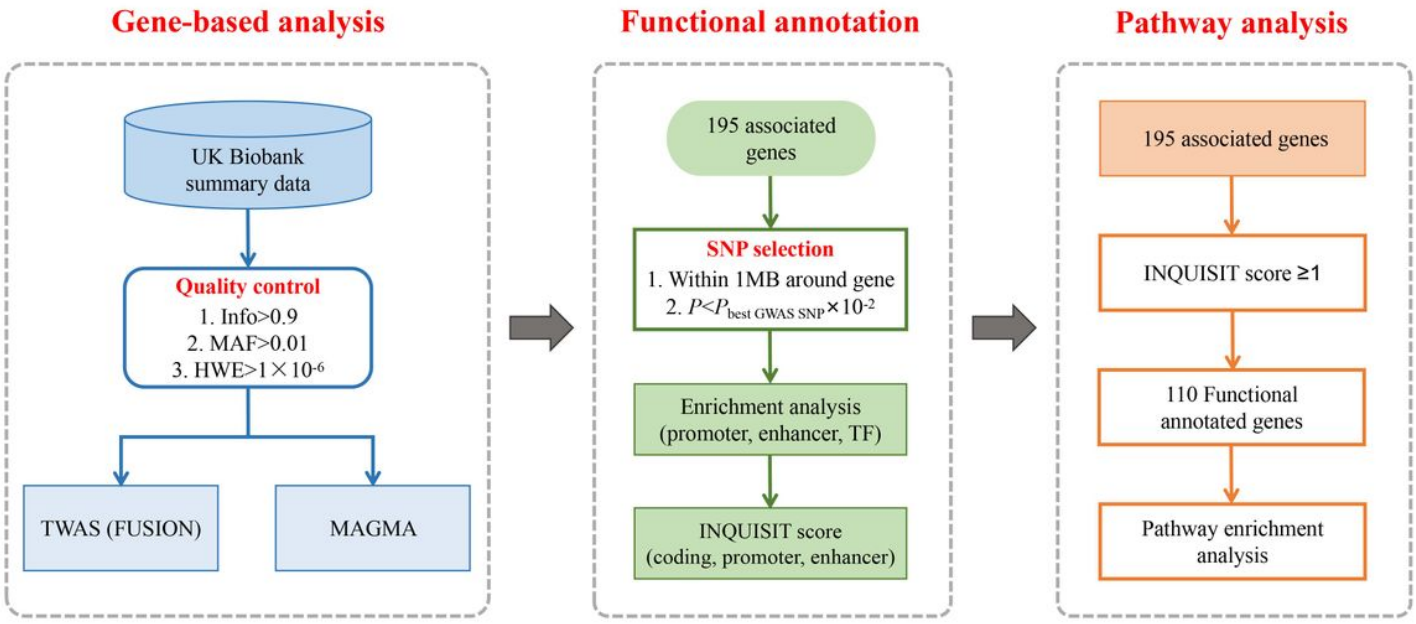

Figure 1

Flowchart for the study design. (1) Gene-based analysis. GWAS-summarized data downloaded from UKB website was used to perform gene-based analysis using FUSION and MAGMA softwares. Standard quality control was conducted for eligibility. (2) Functional annotation. A total of 195 genes were identified significantly associated with chronic respiratory disease risk (Pbonferroni<0.05). Functional annotation of these genes was performed using multiple functional data sources (promoter, enhancer, TF) with INQUISIT algorithm. (3) Pathway analysis. A total of 110 genes with INQUISIT score ≥1 were included in pathway enrichment analysis using R package (clusterProfiler).

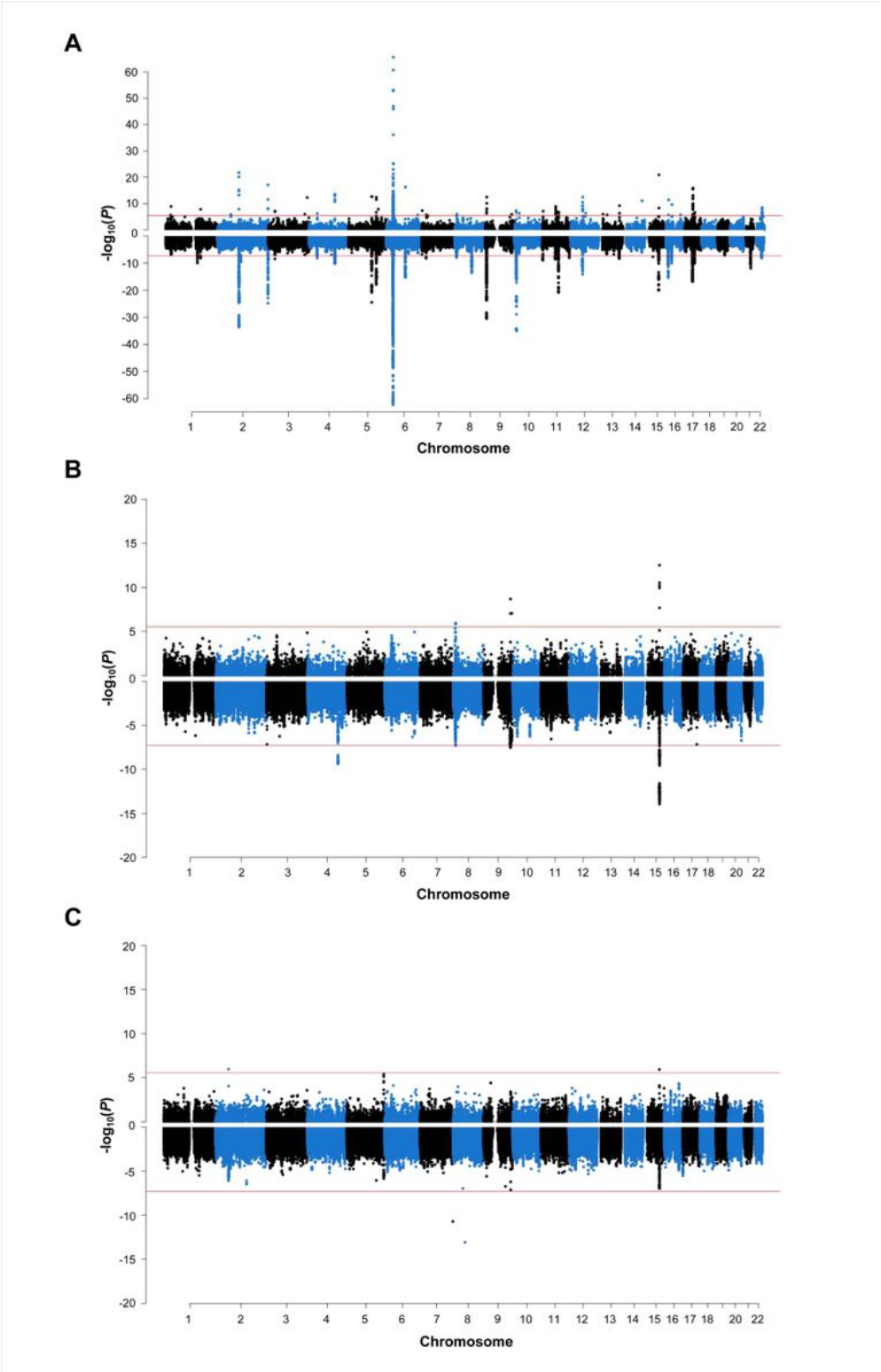

Figure 2

Manhattan plots for (A) Asthma, (B) COPD and (C) Pneumonia GWAS and gene-based associations. The top figure is Manhattan plot for gene-based associations. Each point corresponds to an association test between gene with asthma/COPD/Pneumonia risk. The red line represents the boundary for significance (2.78×10-6). The bottom figure is the GWAS Manhattan plot where each point is the result of a SNP

association test with asthma/COPD/Pneumonia risk. The red line corresponds to the traditional genome-wide significant boundary (5.00×10-8).

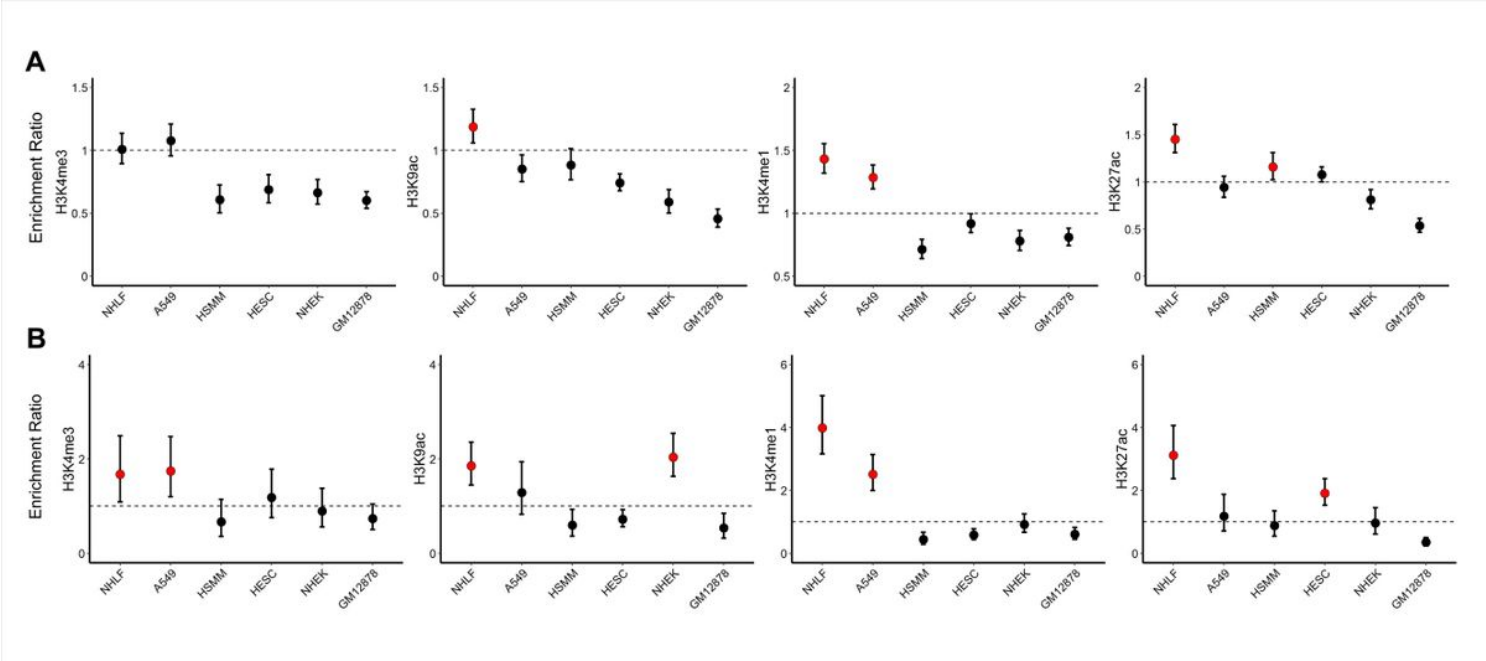

Figure 3

Functional enrichment of CRVs in histone modification regions for (A) asthma and (B) COPD. 2 cell types in human lung tissue, NHLF and A549, as well as 4 human cell types of other tissues (HSMM, HESC, NHEK, and GM12878) were included to investigate the enrichment in H3K4me3, H3K9ac, H3K4me1 and H3K27ac signals, respectively.

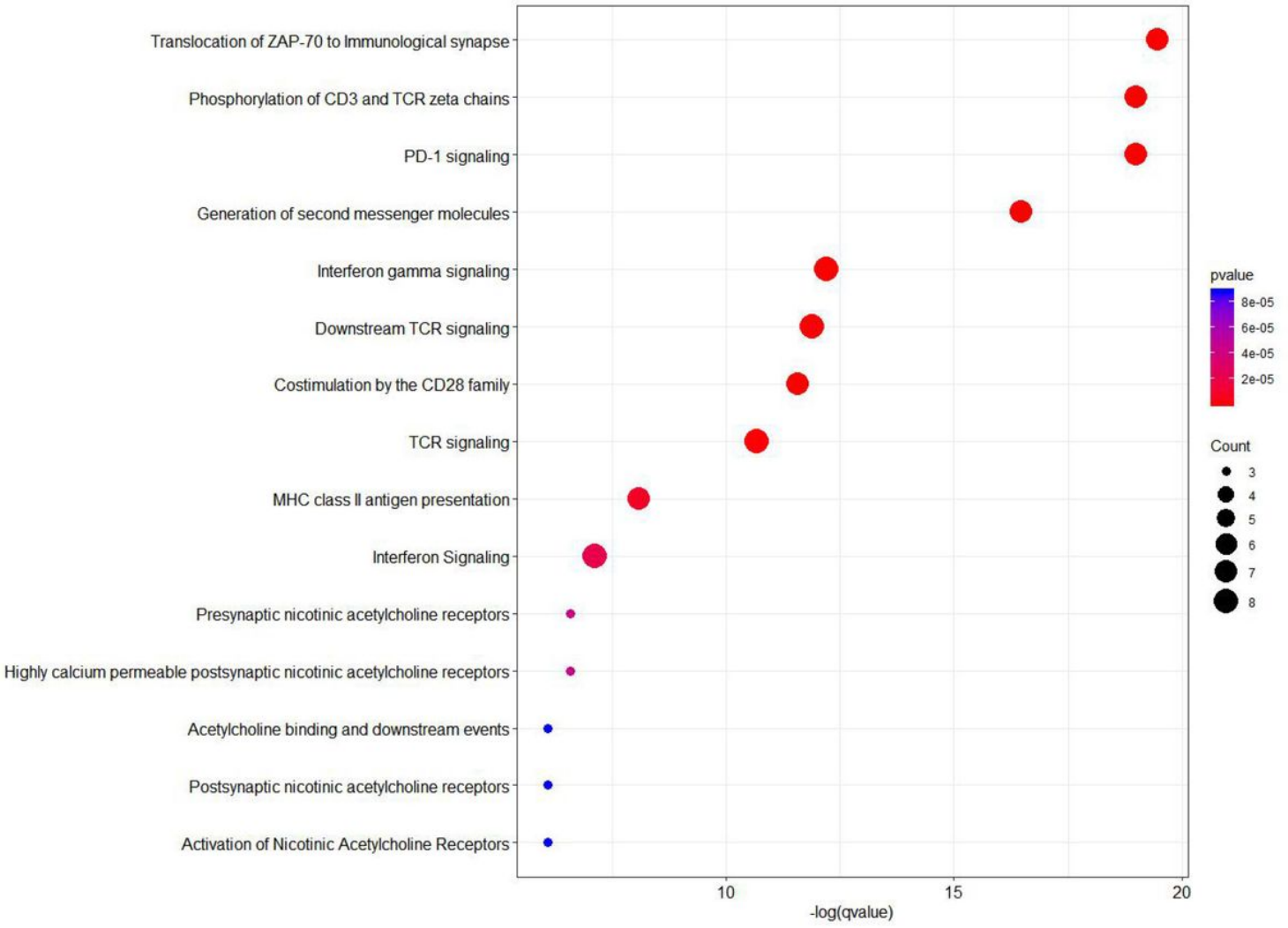

Figure 4

Pathway enrichment of genes with integrated score ≥1 defined by INQUISIT algorithm. We used the combined results from three traits as input considering the highly genetic correlations between traits. The intensity of color represents the magnitude of P value.

Supplementary Files

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

- Additionalfile2SupplementaryFigure1.doc
- Additionalfile3SupplementaryFigure2.doc
- Additionalfile4SupplementaryFigure3.doc
- Additionalfile1SupplementaryTable17.xls