Unsupervised representation learning improves genomic discovery for lung function and respiratory disease prediction

Taedong Yun[1,†], Justin Cosentino[2], Babak Behsaz[1], Zachary R. McCaw[2,◇], Davin Hill[3,4], Robert Luben[5,6], Dongbing Lai[7], John Bates[8], Howard Yang[2], Tae-Hwi Schwantes-An[7,9], Anthony P. Khawaja[5,6], Andrew Carroll[2], Brian D. Hobbs[4,10,11], Michael H. Cho[4,10,11], Cory Y. McLean[1,*,†], and Farhad Hormozdiari[1,*,†]

[*] Joint supervision.

[1] Google Research, Cambridge, MA 02142, USA.

[2] Google Research, Palo Alto, CA 94304, USA.

[3] Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 94304, USA.

[4] Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA.

[5] NIHR Biomedical Research Centre at Moorfields Eye Hospital & UCL Institute of Ophthalmology, London EC1V 9EL, UK.

[6] MRC Epidemiology Unit, University of Cambridge, Cambridge CB2 0SL, UK.

[7] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA.

[8] Verily Life Sciences, Mountain View, CA 94043, USA.

[9] Division of Cardiology, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA.

[10] Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA.

[11] Harvard Medical School, Boston, MA 02115, USA.

[◇] Present address: Insitro, South San Francisco, CA 94080, USA.

[†] Correspondence to: tedyun@google.com, cym@google.com, fhormoz@google.com

Abstract

Background: High-dimensional clinical data are becoming more accessible in biobank-scale datasets. However, accurately phenotyping high-dimensional clinical data remains a major impediment to genetic discovery.

Methods: We introduce a general deep learning framework, REpresentation learning for Genetic discovery on Low-dimensional Embeddings (REGLE), for discovering associations between genetic variants and high-dimensional clinical data. REGLE uses convolutional variational autoencoders to compute a *non-linear, low-dimensional, disentangled embedding* of the data and can also incorporate expert clinical metrics. We demonstrate the utility of REGLE by application to spirograms, which measure lung function. We generate two types of synthetic representations of

pulmonary functions we call spirogram encodings (SPINCs) and residual spirogram encodings (RSPINCs).

Findings: Genome-wide association studies on (R)SPINCs identify more genome-wide significant loci than existing methods while replicating most known lung function loci. Furthermore, (R)SPINCs are associated with overall survival and, under the latent causal variable model, they exhibit significantly high genetic causality proportion with asthma, chronic obstructive pulmonary disease (COPD), and inflammatory diseases. Finally, we construct a set of polygenic risk scores (PRS) that are generally predictive of pulmonary traits and diseases. We demonstrate superior performance predicting asthma and COPD, in multiple ancestries and across four biobanks, compared to PRSs constructed using expert-defined pulmonary function measurements.

Interpretation: REGLE is a method for generating low-dimensional, disentangled representations of high-dimensional clinical data that does not require labels, and improves upon expert-defined phenotypes for genetic discovery and disease prediction. It can flexibly incorporate expert-defined or clinical features and provides a framework to create accurate disease-specific PRS in datasets which have minimal expert phenotyping. (R)SPINCs are quantifying clinically relevant features that are not currently captured in a standardized or automated way.

Funding: Google LLC.

1 Introduction

High-dimensional clinical data (HDCD) provide a unique opportunity to reveal the genetic architecture of diseases and complex traits when coupled with biobank-scale genetic data [1, 2, 3, 4, 5, 6]. However, we lack statistical methods to fully utilize HDCD in genome-wide association studies (GWAS), as standard GWAS require the phenotype of interest to be encoded as a single scalar. Multiple methods have been developed to use HDCD in GWAS, but each has unique limitations.

A natural procedure to use HDCD in GWAS is to perform GWAS on every single data coordinate (e.g. time points or pixels). For example, prior work performed GWAS on each recorded point of electrocardiograms to identify its genetic architecture [7]. There are multiple shortcomings of this approach: 1) running GWAS on thousands of phenotypes can be prohibitively computationally expensive, 2) HDCD often have a correlation structure in which the actual number of degrees of freedom is much lower than the number of coordinates in the data, and 3) multiple hypothesis testing correction for highly correlated coordinates reduces statistical power [8, 9]. One popular approach to address these issues is to use principal component analysis (PCA) [10] on the HDCD and

then perform GWAS on a subset of the principal components (PCs) [11]. However, PCA assumes a linear relationship between the raw HDCD and the underlying biological factors of interest, and does not explicitly model temporal or spatial structure of HDCD. This incomplete representation of the data can lead to suboptimal downstream genetic analysis.

Machine learning-based (ML-based) phenotyping uses HDCD as input to a supervised machine learning model (specifically a deep learning model) to predict trait labels, and then performs GWAS using the model predictions as the target phenotype [3, 12, 6]. While ML-based phenotyping can augment standard GWAS on manual trait labels, the supervised model only learns signals related to the specific target trait and may require many labeled examples in the case of a deep learning model.

The most common method for GWAS on HDCD uses a small number of expert-defined features (EDFs) of HDCD as the target phenotypes. For example, spirograms are a graphical representation of spirometry test results, a widely-used clinical test for lung function that measures airflow and volume over time [13, 14]. Spirograms can be summarized into EDFs including forced vital capacity (FVC), forced expiratory volume in the 1st second (FEV_1), FEV_1/FVC, peak expiratory flow (PEF) and forced mid-expiratory flow ($FEF_{25\text{-}75\%}$) [15]. Spirogram EDFs are used in clinical settings to diagnose diseases such as COPD [16, 17]. EDFs are heritable, and GWAS on EDFs have helped identify the genetic architecture of lung function [18, 19, 20]. However, EDFs may not capture the entirety of biological factors encoded in spirograms and thus GWAS on these EDFs may not exploit the full potential of spirograms.

To overcome these limitations we develop a principled method, REpresentation learning for Genetic discovery on Low-dimensional Embeddings (REGLE), that is computationally efficient, requires no labels, and can incorporate information from EDFs if they are available. As a case study, we apply REGLE to understand the genetic architecture of lung function from raw spirograms. Compared to GWAS on spirogram EDFs (e.g. FEV_1), our GWAS on the learned encodings both recovers most known genetic loci linked to lung function and also detects additional loci. We computed polygenic risk scores (PRS) from GWAS on the learned encodings and evaluated their ability to discriminate asthma and COPD in multiple datasets and genetic ancestries. Several lines of evidence, including stronger lung function enrichments, genetic causality for COPD and asthma, and significantly improved polygenic risk prediction, indicate that REGLE successfully extracts a meaningful low dimensional representation of lung function from spirograms, which in turn improves genetic discovery.

2 Results

2.1 Overview of REGLE

REGLE consists of three main steps: 1) learning a non-linear, low-dimensional, disentangled representation (i.e. an encoding) of the HDCD, 2) performing GWAS on each encoding coordinate, and 3) using PRSs from the encoding coordinates as genetic scores of general biological functions, and potentially combining them to create a PRS for a disease or trait of interest (Figure 1).

In the first step, we train a variational autoencoder (VAE) [21] to compress and reconstruct HDCD (Figure 1 and Methods). Autoencoders consist of a pair of function approximators, typically called an encoder and a decoder, connected by a low-dimensional "bottleneck" layer. The encoder summarizes the input data efficiently into a small set of numbers represented at the bottleneck layer, and the decoder reconstructs the input data from the low-dimensional summary [22]. VAEs [21] are a special type of autoencoders that introduce stochasticity in the encoder. The VAE can force the learned encodings to be relatively disentangled [23], i.e. the encodings have relatively uncorrelated coordinates and separable biological factors can be better captured in each coordinate.

In addition, REGLE enables relevant EDFs to be optionally included in the input to the decoder of the model, so that the encoder is encouraged to learn only the residual signals not represented by the EDFs (Figure 1). This ability to incorporate prior knowledge of important data features (from users or clinicians) is a key advantage of REGLE.

In the second step, we perform GWAS independently on each learned encoding coordinate for all individuals (Methods). In the final step, we compute coordinate-specific PRSs that represent intermediate genetic scores of biological function, and linearly combine these coordinate-specific PRSs into a single disease-specific PRS by training on a small number of individuals with disease labels (Figure 1).

2.2 Overview of REGLE on spirograms

Spirograms are a graphical representation of clinical pulmonary function tests, typically represented by volume-time, flow-time, and flow-volume curves. Spirograms are used to diagnose respiratory diseases such as COPD and understand lung function [24, 19, 6]. To understand the genetic architecture of human lung function, we applied REGLE to obtain low-dimensional representations of spirogram curves, which we call spirogram encodings (SPINCs) (Figure 2). To construct SPINCs, we trained a convolutional VAE [21] to reconstruct spirograms (volume-time and flow-

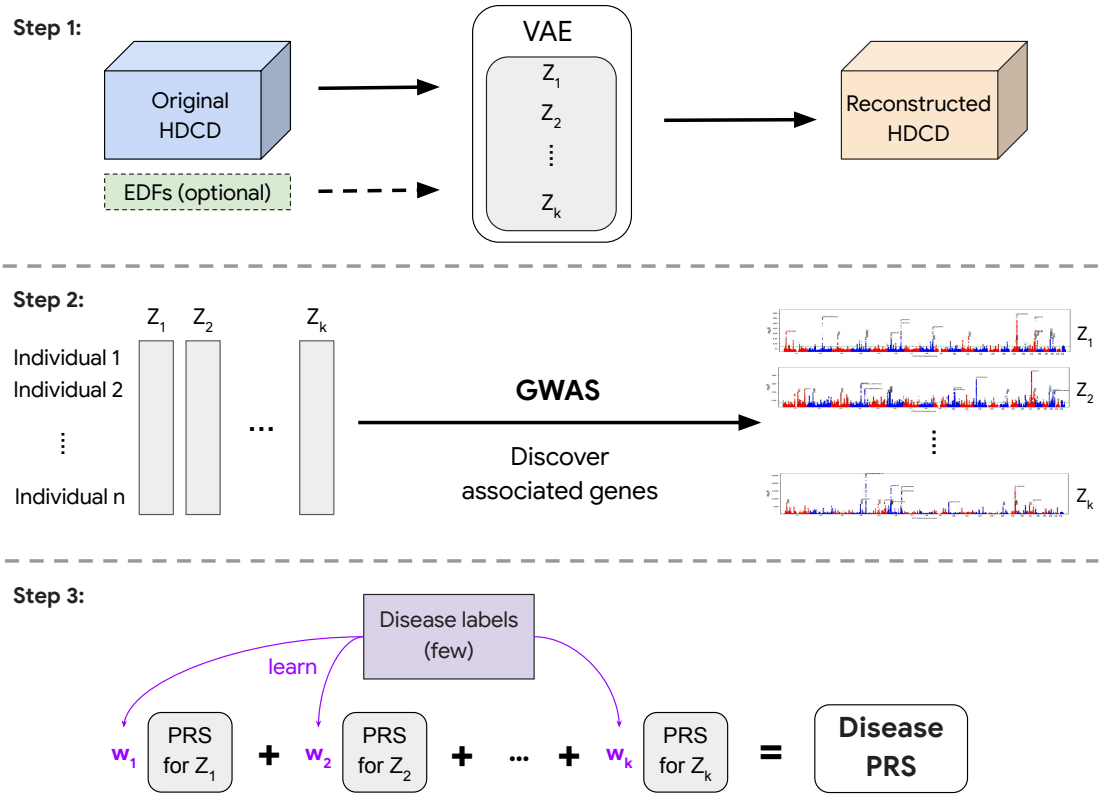

Figure 1: **Overview of representation learning for genetic discovery on low-dimensional embeddings (REGLE)**. In Step 1, we learn a low-dimensional embedding using a VAE where we optionally condition the decoder on EDFs. In Step 2, we perform GWAS on all learned variables (and EDFs if they are used). Finally in Step 3, we train a small linear model to learn weights for each latent variable PRS to obtain the final disease-specific PRS.

time; Figure 2a and Methods). In addition, we constructed another set of encodings we call *residual* spirogram encodings (RSPINCs) by injecting five EDFs (FEV_1, FVC, FEV_1/FVC, PEF, and $FEF_{25\text{-}75\%}$) into the input to the decoder of the REGLE to reconstruct flow-volume (Figure 2a). We generated SPINCs and RSPINCs for all individuals (n=351,120) in UK Biobank [25] using their first visit spirogram, excluding individuals whose spirogram failed our QC measures (Methods). We used 80% of the European-ancestry individuals (n=259,692) to train the (R)SPINCs models and 20% (n=65,266) to evaluate the reconstruction performance and choose hyperparameters (Supplementary Figure 1, Supplementary Table 1, and Methods). Using just 5 SPINCs (the same as the number of common spirogram EDFs), we observed highly accurate reconstruction of the input spirograms (Figure 2b) based on mean squared error across time. SPINCs consistently outperformed an equivalent number of PCs in terms of reconstruction accuracy at small latent dimensions (Figure 2c,

Supplementary Notes). We observed similarly accurate reconstructions for EDFs+RSPINCs as well with RSPINCs (Supplementary Figure 2); we used 2 RSPINCs to balance the number of additional coordinates and the reconstruction accuracy. Importantly, the learned representations are highly consistent when trained with multiple different initializations (Supplementary Figure 3, Supplementary Notes).

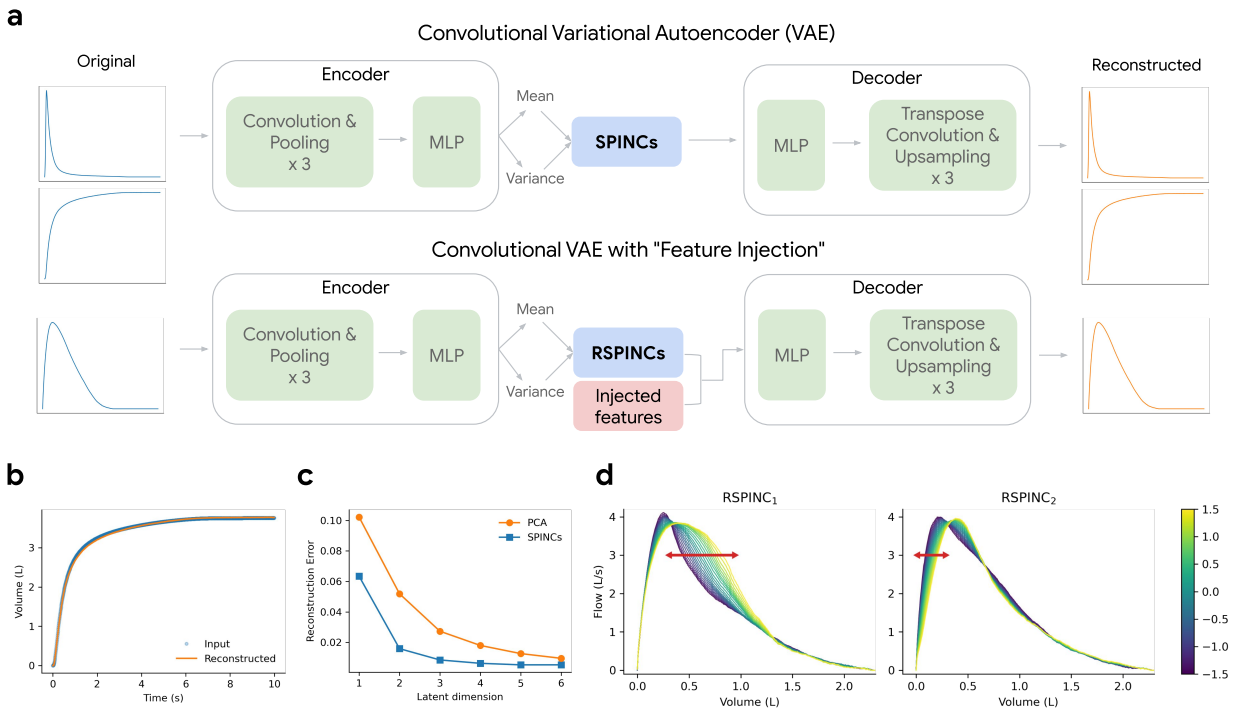

Figure 2: **Overview of REGLE on spirograms**. a) Learning spirogram encodings (SPINCs) using a convolutional variational autoencoder (VAE) and *residual* spirogram encodings (RSPINCs) using a convolutional VAE with "features injection", using expert-defined features (EDFs) for example. b) Reconstructing a spirogram (volume-time curve) from SPINCs (dim=5). c) Reconstruction errors (mean squared error across time points) for reconstructed spirograms using the SPINCs model and PCA with a varying latent dimension. Both the SPINCs model and PCA are trained (or "fitted") on a training set and the reconstruction error was evaluated in a separate validation set. d) Spirograms created by RSPINCs (dim=2) decoder using a fixed set of injected features (i.e. EDFs) and varying one RSPINC coordinate while fixing the other one to be zero. Line color indicates the varying RSPINC coordinate value.

2.3 (R)SPINCs are partially interpretable

To interpret the influence of RSPINC coordinates on spirogram shape, we fixed the values of EDFs (obtained from a randomly selected individual in the validation set) and varied one RSPINC co-

ordinate while keeping the other one fixed at zero, and generated the corresponding flow-volume spirograms using only the decoder portion of the RSPINCs model (Figure 2d). A typical flow-volume spirogram consists of two distinct parts: the first part, a relatively brief part to reach peak flow where the flow increases monotonically as the volume increases, and the second part, the main part of the spirogram where the flow is monotonically decreasing. In Figure 2d, we clearly observed that varying the first coordinate of RSPINCs amounts to widening or narrowing of the second part (negative slope) while keeping the first part relatively fixed. Similarly, varying the second coordinate of RSPINCs widens or narrows the first part (positive slope) while keeping the second part relatively fixed. Notably, when varying either coordinate of RSPINCs, the maximum flow value (PEF) and the final volume value (FVC) stay roughly the same, as expected since all EDFs were fixed.

2.4 (R)SPINCs encode information beyond EDFs

Some coordinates of SPINCs are highly correlated with known EDFs. For example, the 3rd coordinate of SPINCs is 96% correlated with FVC and 94% correlated with FEV_1, while the 2nd coordinate is 73% correlated with FEV_1/FVC (after flipping the signs) (Supplementary Figure 4). Both RSPINCs coordinates have low correlation ($|R| < 0.3$) with all EDFs, which is expected since they were encouraged to learn only residual signals not captured by the EDFs (Supplementary Figure 4). Both SPINCs and RSPINCs are correlated with other predictors of lung function ("covariates"), such as age, sex, height, body mass index, and smoking status (Supplementary Figure 4). To investigate if (R)SPINCs include information beyond EDFs and covariates, we residualized both the EDFs and the covariates from (R)SPINCs and computed correlation with tabular UK Biobank features. Multiple groups of fields strongly and significantly correlated with the (R)SPINCS even after residualizing the EDFs and the covariates, including asthma, breathing issues, cognitive function, and allergies (Supplementary Tables 3 and 4). Using the Cox proportional hazards model, we also observed (R)SPINCs are associated with overall survival (Supplementary Notes, Supplementary Figures 5 and 6, Supplementary Table 5, Methods).

2.5 (R)SPINCs detect additional novel loci for lung function

We generated SPINCs (dim=5) and RSPINCs (dim=2, in addition to 5 EDFs) for all individuals with valid first-visit spirograms in UK Biobank (Supplementary Figures 1, 7 and 8; Methods). Then, we

performed GWAS on all European-ancestry individuals(n=324,702) on all encoding coordinates and 5 EDFs using BOLT-LMM [26, 27], adjusting for age, sex, age^2, age $\times$ sex, height, height^2, body mass index, smoking status, pack-years of smoking, the type of genotyping array, and the top 15 genetic principal components (Methods). The Manhattan plots of 5 SPINCs and 2 RSPINCs GWAS are illustrated in Supplementary Figures 9 to 13 and Supplementary Figures 14 and 15, respectively. The intercept term from the stratified linkage disequilibrium score regression (S-LDSC) [28] was close to 1 (Supplementary Table 6) for the GWAS of SPINCs and RSPINCs, indicating minimal confounding bias. The SNP-heritability estimated from S-LDSC for SPINCs and RSPINCs showed strong genetic components (Supplementary Table 6). For comparison, we also performed GWAS on the first 5 PCs of the raw spirograms following the same steps.

We observed that GWAS on 5 SPINCs detected 575 independent genome-wide significant (GWS) loci ($R^2 \leq 0.1$ and $P \leq 5 \times 10^{-8}$) after merging hits within 250kb together (Table 1, Methods). To compare our results to known lung function loci from previous literature, we combined the largest published GWAS on lung function (using FEV_1, FVC, PEF, and FEV_1/FVC) from Shrine et al. [20] (580,869 individuals, compared to our 324,702 individuals in UK Biobank) with all lung function-related loci in the NHGRI-EBI GWAS Catalog [29] (Methods). This resulted in 1104 independent loci after merging loci by distance (250kb), hereafter referred to as "previously known loci". Most GWS loci from SPINCs and EDFs+RSPINCs recover previously known loci (89% for SPINCs, 90% for EDFs+RSPINCs). Out of 575 genome-wide significant (GWS) SPINCs loci, 65 (11%) were not previously known, compared to 32 from EDFs and 15 from PCA. Of 659 EDFs+RSPINCs GWS loci, 63 (10%) were not previously known (Table 1). Functional enrichment analysis with GARFIELD [30] shows that these loci are enriched for lung tissue DNase I hypersensitive sites (Supplementary Figures 16 to 20; Supplementary Figures 21 and 22; Supplementary Notes) and the EDFs+RSPINCs loci show stronger ontology term enrichments than EDFs loci (Supplementary Figure 23) using GREAT [31]. Notably, we found a strong enrichment for $RSPINC_2$ in blood (Supplementary Figure 22).

Among the GWS loci discovered by SPINCs and EDFs+RSPINCs GWAS, we applied a stricter P-value threshold of 1.0×10^{-8} (to account for testing multiple coordinates) and further removed all loci discovered by our GWAS on EDFs, in addition to the previously known loci. We found 25 loci remain from SPINCs and 30 loci from RSPINCs (Supplementary Tables 7 and 8). Lastly, to further validate these novel loci, we used GCTA-COJO [32] to compute the association statistics conditioned on previously known loci. Nearly all novel loci remain significant after the conditional

analyses (Supplementary Tables 7 and 8). Thus, we conclude that these are potentially novel loci associated with lung function that were not discovered by previous methods and may warrant further biological investigation.

We note that in recent GWAS literature, some researchers transform the target phenotype values at the cohort level using rank-based "inverse-normal transformation (INT)", which can improve statistical power in certain cases [33]. After applying INT on all traits (EDFs, PCs, SPINCs, and RSPINCs), we observed increased numbers of hits from all methods (Supplementary Table 9), but the overall relative trends remained consistent (Table 1).

Method (# traits)	Sample size	Total	Known (%)	Novel (%)
Shrine 2023 + GWAS Catalog	> 581K*	1104	–	–
Shrine 2023	581K	754	–	–
EDFs (5)	325K	613	581 (95%)	32 (5%)
PCA (5)	325K	412	397 (96%)	15 (4%)
SPINCs (5)	325K	575	510 (89%)	65 (11%)
EDFs+RSPINCs (7)	325K	659	596 (90%)	63 (10%)

Table 1: **Comparison of (R)SPINCs loci with previous GWAS.** Expert-defined features (EDFs) are FEV_1, FVC, FEV_1/FVC, PEF, and $FEF_{25\text{-}75\%}$. "Known" and "novel" is in reference to lung function loci in Shrine et al. [20] and GWAS catalog.
* GWAS in Shrine et al. [20] has 580,869 individuals and other previous GWAS in the GWAS catalog may have more individuals.

2.6 (R)SPINCs improves asthma and COPD PRS over EDFs in UK Biobank

We computed PRSs using BOLT-LMM [26, 27] for 5 SPINCs and 2 RSPINCs coordinates, in addition to 5 EDFs. We treat these sets of PRSs as intermediate genetic scores for lung function. Given a specific trait, a set of such intermediate PRSs, and a (small) set of individuals for whom the trait status is available, one can combine the intermediate PRSs into a single trait-specific PRS as a linear weighted sum of the intermediate PRSs by learning the weights using the individuals with trait status (Figure 1). We created disease-specific PRSs for asthma and COPD from three sets of intermediate PRSs: 1) 5 EDFs, 2) 5 SPINCs, and 3) 5 EDFs plus 2 RSPINCs. We trained the disease-specific PRSs within the modeling set (n=324,958) of European ancestry individuals in UK Biobank using medical-record-based asthma and COPD statuses. To evaluate the performance of each disease-specific PRS, we computed the accuracy of the PRS in a completely separate set of individuals of European ancestry (n=110,722) not previously used for model training or GWAS.

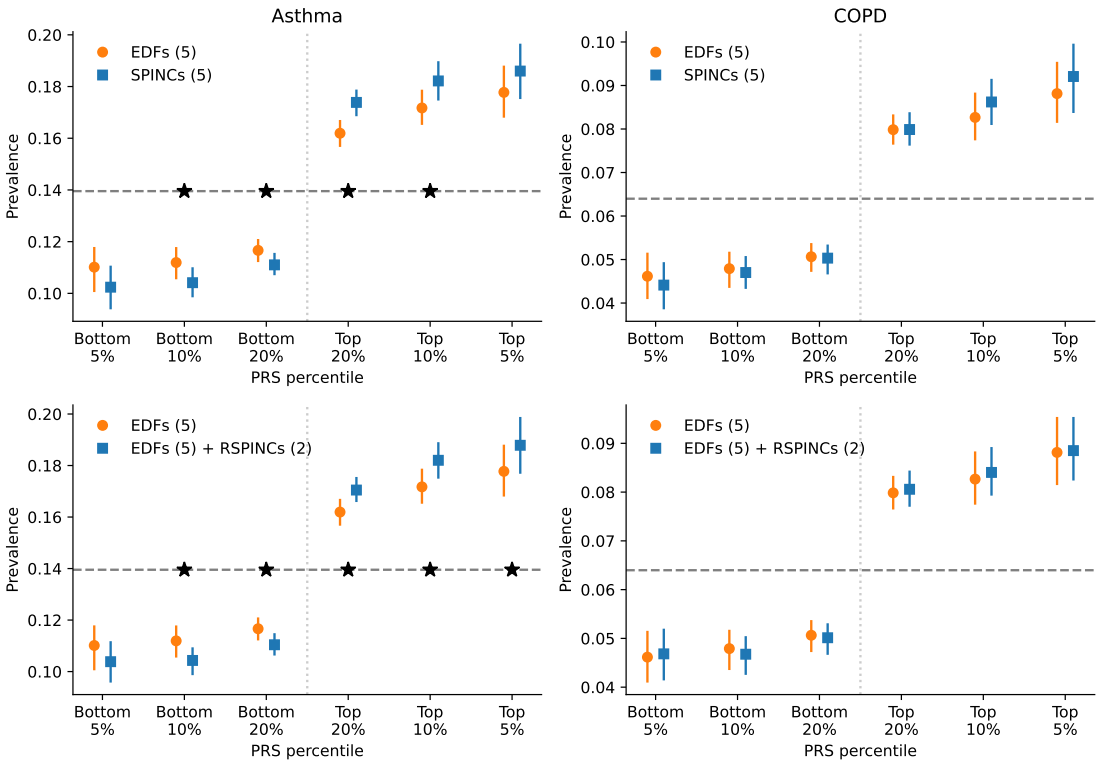

Figure 3: **PRS using SPINCs and RSPINCs in UK Biobank.** Combined PRS for medical-record-based asthma and COPD using three sets of intermediate PRS: 5 EDFs, 5 SPINCs, and 5 EDFs + 2 RSPINCs. Each set of PRS is combined by a linear model trained using the target phenotype labels and the prevalence of the phenotypes in the top and bottom 5%, 10%, and 20% PRS individuals is evaluated in a separate evaluation set. Vertical line segments indicate 95% confidence interval generated by bootstrapping (300 repetitions). The horizontal dashed line shows the total prevalence. Star (*) sign indicates a statistically significant difference between the two methods using *paired* bootstrapping (300 repetitions) with 95% confidence. Lower is better for the bottom percentiles; higher is better for the top percentiles.

We observed that the top-decile high-risk individuals based on the SPINCs asthma PRS have an asthma prevalence of 18.2%, while the top-decile individuals in EDFs asthma PRS have a prevalence of 17.2% (5.8% overall improvement; Figure 3, Supplementary Table 10). In fact, when considering the top and bottom 5%, 10%, 20% PRS individuals based on SPINCs and EDFs asthma PRSs, we observed that all top percentiles of SPINCs asthma PRS show higher asthma prevalence, and all bottom percentiles of SPINCs asthma PRS show lower asthma prevalence, than the EDFs asthma PRS, and four of the six differences are statistically significant (95% confidence with paired bootstrapping) (Figure 3). Thus, SPINCs asthma PRS more effectively stratifies the risk groups than EDFs asthma PRS on both ends of the risk spectrum. In addition, we observed statistically significant improvements in AUC-ROC, AUC-PR, and Pearson correlation by using the SPINCs asthma

PRS (Supplementary Table 10). We observed the same trend for COPD, in which the top-decile prevalence from SPINCs COPD PRS is 8.6% compared to 8.3% for the EDFs COPD PRS (Figure 3, Supplementary Table 11). We also observed the same trend of improvement in other metrics including AUC-ROC, AUC-PR, and Pearson correlation (Supplementary Tables 10 and 11). Lastly, we observed that the EDFs+RSPINCs PRS outperforms the EDFs PRS on all metrics for both asthma and COPD, with all differences except for COPD top-decile prevalence reaching statistical significance (Figure 3, Supplementary Tables 10 and 11).

We observed that the SPINCs COPD PRS outperforms the FEV_1/FVC PRS (Supplementary Table 11) for predicting medical-record-based COPD, despite FEV_1/FVC having previously been shown to be one of the best phenotypes for generating a COPD PRS, even outperforming the binary COPD PRS [6]. These results provide further evidence that SPINCs capture more genetic determinants of lung function related to asthma and COPD than the same number of EDFs, and that RSPINCs capture additional genetic factors not captured by the EDFs.

Finally, since each disease-specific PRS is constructed from intermediate (R)SPINCs coordinate PRS using just 5–7 learned weights, we explored whether an accurate disease-specific PRS could be learned from a subset of the training data. For both COPD and asthma, (R)SPINCs-based PRS fit with as few as 100 disease cases performed indistinguishably from those trained on the full training data (Figure 1 "Step 3", Supplementary Figure 25). We also evaluated PRSs generated by GWAS with a cohort-level phenotype adjustment using inverse-normal transformation. We observed fewer significant differences with this adjustment, though SPINCs and EDFs+RSPINCs still maintain some statistically significant differences for asthma PRS (Supplementary Figure 24).

2.7 (R)SPINCs PRS transferred to independent datasets and ancestry groups

The performance of PRS computed in one dataset can degrade significantly when transferring the variant weights directly to another dataset or a different ancestry group [34]. In addition, the quality of COPD and asthma labels in different datasets can vary widely (e.g. using physician diagnoses, self report, or medical records). To test the generalizability of our (R)SPINCs PRSs to individuals outside of UK Biobank and non-European ancestry individuals, we transferred the variant weights from the asthma and COPD PRSs in UK Biobank directly to the Genetic Epidemiology of COPD (COPDGene) dataset [35], the eMERGE III dataset, the EPIC-Norfolk dataset [36], and Indiana Biobank dataset [37].

For the COPDGene dataset, we computed PRS of all individuals using the same variant effect

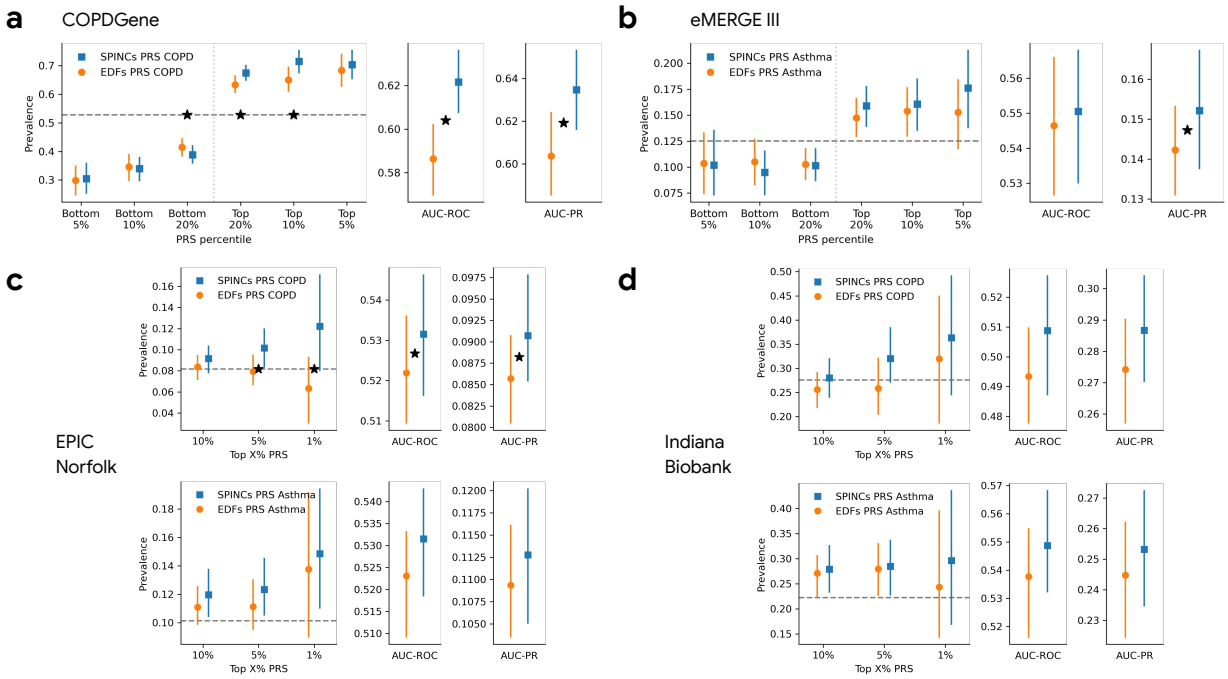

Figure 4: **SPINCs PRS transferred to multiple independent datasets**. SPINCs PRS for COPD and asthma generated on UK Biobank are transferred to four independent datasets: COPDGene, eMERGE III, EPIC Norfolk, and Indiana Biobank. a) PRS evaluation in COPDGene dataset on COPD. b) PRS evaluation in eMERGE III dataset on asthma. c) PRS evaluation in EPIC Norfolk study on COPD and asthma. d) PRS evaluation in Indiana Biobank on COPD and asthma. In all figures, solid vertical intervals represent 95% confidence interval generated by bootstrapping (300 repetitions). The horizontal dashed lines show the total prevalence in the evaluation set. Star symbols indicate a statistically significant difference between the two methods using paired bootstrapping (300 repetitions) with 95% confidence.

sizes obtained by BOLT-LMM in UK Biobank and the same linear weights to combine the EDFs, EDFs+RSPINCs, and SPINCs PRS as before after matching variants. We used the "race" field in COPDGene as a proxy for genetic ancestry and computed PRS performance in the two available subsets separately: "Non-Hispanic White" and "African American". We observed that for COPD, the SPINCs COPD PRS outperforms the EDFs COPD PRS for both subset of individuals in COPDGene for all four evaluation metrics (AUC-ROC, AUC-PR, top-decile prevalence, and Pearson correlation). In the "Non-Hispanic White" subset (n=6,576), which matches the UK Biobank ancestry group on which the PRS was trained, all four metrics are statistically significant (paired bootstrapping; Figure 4a, Supplementary Table 12). In the "African American" subset (n=3,140), differences were statistically significant for AUC-ROC and Pearson correlation (Supplementary Table 12). The EDFs+RSPINCs COPD PRS significantly outperformed the EDFs COPD PRS in "Non-Hispanic White" in AUC-ROC and Pearson correlation, but did not in the "African American" subset (Supple-

mentary Table 12).

We also transferred the UK Biobank PRSs to eMERGE III ("White" subset, n=8,288), EPIC Norfolk (self-reported "White", n=21,010), and the Indiana Biobank (mostly European, see Methods, n=5,254), to evaluate asthma, asthma and COPD, and asthma and COPD, respectively. We observed consistent improvement from using SPINCs PRSs over EDFs PRSs for both COPD and asthma phenotypes for top-percentile prevalences, AUC-ROC, and AUC-PR. The improvement was statistically significant for AUC-PR and the top 1% and 5% prevalence in eMERGE III, and for AUC-ROC and AUC-PR in EPIC Norfolk (Figure 4b-d).

2.8 High association between (R)SPINCs and UKB phenotypes PRSs

To assess the influence of (R)SPINCs on traits and health outcomes, we performed phenome-wide association studies (PheWAS). We compared pruning+thresholding PRSs of all (R)SPINCs coordinates to PRSs of 7,145 phenotypes computed by the Pan-UK Biobank consortium (Methods, URLs). For SPINCs, significant associations with strong correlation magnitude are driven by the 2nd and 3rd coordinates, which as mentioned above are strongly correlated with EDFs, and include expected phenotypes such as FEV_1, FEV_1/FVC, and PEF (Supplementary Tables 13 and 14). These coordinates also show strong correlation with other diseases and traits previously hypothesized to relate to COPD, including systemic lupus erythematosus [38, 39], thyroid dysfunction [40], and gluten-free diet [41]. For RSPINCs, significant associations with strong correlation magnitude are nearly all driven by $RSPINC_1$ and also include the same diseases (Supplementary Tables 15 and 16).

Finally, we performed latent causal variable (LCV) analysis [42] to identify potentially causal relationships between (R)SPINCs and EDFs with COPD, asthma, and lung disease phenotypes (sarcoidosis, systemic lupus erythematosus, thyroid dysfunction, and gluten-free diet) where we observed significant correlation between their PRSs in PheWAS (Supplementary Tables 17 and 18). We observed a significant positive LCV output between the 2nd SPINCs coordinate and COPD, suggesting a genetic causal link from SPINCs to COPD. We observed a significant negative LCV output between the 5th SPINCs coordinate and asthma, suggesting a genetic causal link from asthma to SPINCs. Interestingly, for all other lung disease phenotypes, we observed a direction of effect from these phenotypes to SPINCs and RSPINCs (Supplementary Tables 17 and 18; Supplementary Notes).

3 Discussion

Large biobanks provide unique opportunities to identify the genetic factors underlying complex traits and diseases, but accurately phenotyping [43] the cohorts remains a core challenge. We proposed a general unsupervised deep representation learning method, REGLE (representation learning for genetic discovery on low-dimensional embedding), to discover the full genetic component of a high dimensional clinical data (HDCD). We showcased the effectiveness of REGLE in spirograms, where it produced latent variables ("encodings") that are both partially interpretable and effective for identifying genetic variants associated with lung function. Multiple lines of evidence show the relevance of the model representations for quantifying general lung function.

Unsupervised representation learning of HDCD for genomic discovery is attractive owing to the difficulty of acquiring (or defining) manual phenotypes at scale. Prior work has explored applying transfer learning [44] and contrastive learning [45] to retinal fundus images, or multimodal autoencoders to cardiac data modalities [46]. A key strength of our method is the use of a VAE to generate the non-linear low-dimensional representations of HDCD. First, by construction, the coordinates of the latent representation are minimally correlated, which strengthens the combined power of the downstream GWASs. As a result, the PRSs of the learned encodings are also minimally correlated and contain relatively orthogonal genetic signals, which may contribute to the superior accuracy of the disease/trait specific PRS created by the REGLE pipeline. Second, the learned representations are stable up to changes in sign or order as we observed empirically (Section 2.3), potentially due to a grounding effect of a VAE prior in the probabilistic model. As changes in sign or order do not affect the outcomes of GWASs, the results of the REGLE pipeline are stable and replicable. Regular autoencoders without a prior do not have this stability property as they can learn any invertible linear transformation of a specific learned representation.

The architecture modification we introduce in REGLE to support expert-defined features (EDFs) enables a principled use of expert human knowledge and encourages the remaining latent coordinates to encode biological function explicitly not captured by the EDFs. This provides the opportunity to build upon and improve the existing clinical practices with the extra information provided from the residual features. For example, clinical review of pulmonary function tests includes visual inspection of the curves for variation in shape. Coving of the flow volume loop is an indicator of obstruction. The ability of (R)SPINCs to identify these differences while holding EDFs constant suggest that (R)SPINCs are quantifying clinically relevant parameters that are not currently cap-

tured in a standardized or automated way. While we demonstrated the value of the method on spirograms to model lung function, it can be generalized to other HDCD modalities such as images.

The improved performance of SPINCs and EDFs+RSPINCs PRS over EDFs PRS provides evidence for the existence of such residual genetic information. Importantly, the (R)SPINCs PRS for asthma and COPD reduce each (R)SPINC coordinate GWAS into a PRS based solely on the effect size estimates from the GWAS on the learned coordinates, and the disease-specific PRS is simply a learned weighted sum of the five or seven constituent coordinate PRSs. This has important implications for disease risk prediction: it demonstrates that given a dataset with widespread unlabeled quantification of lung system function (i.e. spirograms), genetic predictors for specific lung diseases can be accurately learned with very few disease labels (enough to learn 5-7 features). We hypothesize that unsupervised quantification of other organ systems may be similarly beneficial for improving polygenic prediction across a wealth of diseases. Finally, we note that the PRS performance reported here likely represents a lower bound achievable by the method; jointly estimating disease-specific variant effect size estimates on the set of variants identified by the (R)SPINCs GWAS may further improve performance.

There are several limitations to this work. First, while unsupervised representation learning of HDCD for biological function is likely beneficial across data modalities, the generalizability of the specific VAE method introduced here to even higher dimensional modalities like imaging and video has not been assessed. In particular, VAEs tend to produce blurry image reconstructions [47, 48]; whether and how that affects the ability of a VAE to encode representations meaningful for genomic discovery is important future work. Second, we did not directly optimize multiple GWASs for novel genomic discovery, but used a straightforward (conservative) method to define and merge independent associated loci. A possible extension would be to combine the signals from multiple (R)SPINC coordinate GWAS [11] or apply methods that maximize heritability (e.g. [49]). Third, we did not fully explore model architecture and training strategies specifically for genomic discovery. Some ideas which may warrant further investigation include: 1) using previously proposed modifications to the VAE loss function and the training procedure to maximize the degree of disentanglement of coordinates while balancing the reconstruction error [50, 51, 52], 2) incorporating an additional loss term to explicitly discourage correlation between RSPINCs and EDFs, and 3) introducing (semi-) supervision in model training to overcome the limitations of purely unsupervised training [53]. Fourth, none of the spirograms in UK Biobank were obtained after bronchodilation, and the asthma and COPD labels defined using clinical records are known to be noisy [6]. Fifth, we generated

individual-level spirogram representations from a single blow, despite some individuals having up to three acceptable blows in UK Biobank. Integrating all acceptable blows from an individual could produce a more comprehensive representation of their lung function [54]. Sixth, model training was performed exclusively in individuals of European ancestry. While PRS evaluation was performed in multiple datasets and ancestries, the impact of ancestry-specific model training was not explored.

Despite these limitations, our method provides a mechanism for identifying genetic influences on organ function in the absence of labeled data, and naturally admits incorporating expert features in the model. It also provides a method to create disease/trait specific PRS with very few labels (i.e. in the order of hundreds). As biobanks with rich imaging, activity monitoring, medical records, and paired genetic data continue to grow, we anticipate that this or similar methods will be increasingly used to further elucidate the genetic underpinnings of human traits and diseases.

4 Methods

4.1 UK Biobank data preparation

Spirograms from UK Biobank were sourced from the data field 3066, which contains the volume in milliliters of exhalation at 10 millisecond intervals (volume-time curve) and were preprocessed closely following the procedures in Cosentino et al. [6]. To generate flow-time curves, we approximated the first derivative of volume with respect to time by taking a finite difference in the volume-time curves. We normalized the volume-time and flow-time curves to 1000 time points, by either truncating longer curves or by right-padding shorter curves with zero (for flow-time curves) or the final value (for volume-time curves), and removed FEV_1, FVC, PEF values in the extreme tail (top/bottom 0.5%) of the observed values and all blows that fail the acceptability provided by UK Biobank data field 3061. We used the first acceptable blow of an individual when there is more than one. In addition, we dropped all flow curves whose values don't fall in [-10, 20], all volume curves whose values are not in [-5, 10], and all flow curves in which the proportion of nonzero values are less than 20%. Finally, we generated flow-volume curves from volume-time and flow-time curves by interpolating 1000 evenly spaced volume values between 0 and 6.58 liters (the maximum observed volume in the dataset).

We then subdivide all European ancestry individuals processed this way into a 80% training set and a 20% validation set. After additionally removing related individuals, there are 259,692 indi-

viduals in the training set and 65,266 individuals in the validation set (Supplementary Figure 1).

Asthma and COPD status were determined by medical records using self report, ICD9, and ICD10 codes as defined in Cosentino et al. [6].

4.2 Convolutional VAE model architecture and training

To generate SPINCs, we encode the flow-time and volume-time curves. In our VAE, we use one-dimensional convolutional layers to utilize the temporal context of this time series, encoding the two curves in two channels. In the encoder, we first apply three 1D convolutional layers, each followed by max pooling, and use three fully-connected layers to generate the mean and variance of the bottleneck layer. We use 5 latent dimensions, identical to the number of EDFs, and each latent coordinate is sampled from the Gaussian distribution with the learned means and variances. The decoder architecture is a mirror image of the encoder. We start with three fully-connected layers followed by transpose convolutions layers, each prepended by an upsampling layer. See "SPINCs model architecture" in Supplementary Notes for full details.

For RSPINCs, we encode the flow-volume curve alone, and we apply the same sequences of convolutional and fully-connected layers as we did for SPINCs, while using only 2 latent dimensions in this case. Importantly, we use a novel VAE architecture to concatenate the 5 EDFs directly to the sampled output of the bottleneck layer (the layer right before the decoder) to learn only the residual signals not represented by the EDFs, as previously discussed in Figure 2a. As a result, the encoder output dimension is 2 while the decoder input has dimension $5 + 2 = 7$. See "RSPINCs model architecture" in Supplementary Notes for full details.

Both models are trained using the standard VAE loss function consisting of the reconstruction loss and the (rescaled) Kullback–Leibler (KL) divergence loss. For RSPINCs the KL divergence loss is only applied to the learned encodings, not to the injected EDFs. For optimization, the Adam optimizer [55] is used with varying learning rates and batch sizes. The final learning rate and batch size values ("hyperparameters") for SPINCs and RSPINCs were chosen to minimize the VAE loss in the validation set (Supplementary Notes, Supplementary Table 1).

After training SPINCs and RSPINCs models, we use the encoders of the trained models to generate the encodings for each individuals, using the mean value of the learned Gaussian distribution of the encodings.

All models were implemented in TensorFlow V2 [56] and XManager (URLs) was used to manage multiple machine learning experiments.

4.3 Phenotypic Correlation Analysis

To residualize EDFs and/or covariates from (R)SPINCs, we used ordinary least squares linear regression. To compute the correlation of the EDFs-and-covariates-residualized (R)SPINCs with the tabular fields in UK Biobank, we first preprocessed the tabular fields to remove special codes, normalize, impute and aggregate the values, and finally transformed the categorical fields into one-hot encodings. For each individual correlation analysis between a feature and one of the (R)SPINCs, we computed the Pearson correlation and the P-value with two-sided alternative hypothesis.

4.4 Survival analysis

We performed analysis of overall survival for European individuals in the validation set (n=65,266) using the time from first assessment (field 53) to death from any cause (field 40000). Subjects who were not known to have died were right-censored at the date of UKB data ingestion (Feb 12, 2018). We quantified the association between overall survival and each SPINC, RSPINC, and EDF per standard deviation using the hazard ratio, which was estimated from a Cox proportional hazards model adjusting for age and sex. The proportional hazards assumption, with respect to each SPINC, RSPINC, and EDF, was assessed using the Schoenfeld residual test. After stratifying patients into quartiles using each SPINC, RSPINC, or EDF coordinate, the overall survival curves in Supplementary Figures 5 and 6 were constructed using the standard Kaplan-Meier estimator with bootstrapped 95% confidence intervals.

4.5 Genome-wide association studies and polygenic risk scores

GWAS on all EDFs, SPINCs, and RSPINCs were performed using BOLT-LMM v2.3.6 [26, 27], adjusting for age, sex, age^2, age $\times$ sex, height, height^2, body mass index, smoking status, the number of packs of cigarettes smoked per year, the type of genotyping array, and the top 15 genetic principal components as covariates. GWAS was restricted to European ancestry individuals to minimize confounding. For quality control we kept variants with minor allele frequency ≥ 0.001, imputation INFO score ≥ 0.8, missing call fraction ≤ 0.05, and Hardy-Weinberg equilibrium P-value $\geq 10^{-10}$, among all genotyped and imputed variants provided by UK Biobank. After GWAS, we performed S-LDSC [28] to estimate SNP-heritability and detect potential confounding.

Genome-wide significant "hits" were defined as the most significant variants with $p \leq 5 \times 10^{-8}$ and independent at $R^2 < 0.1$ using the PLINK `--clump` command. A reference panel for

linkage disequilibrium (LD) calculation contained 10,000 unrelated European samples from the UKB. Significant "loci" were created based on the span of reference panel SNPs in LD ($R^2 \geq 0.1$) with the hits. Loci separated by fewer than 250 kb were subsequently merged.

While performing GWAS, PRSs for all traits (EDFs, SPINCs, RSPINCs, etc.) were computed using the `-predBetasFile` option of BOLT-LMM, which generates PRS coefficients using a Bayesian linear mixed model. While GWAS was performed on individuals with valid spirometry measurements, we evaluated the performance of the PRS in a separate set of individuals not used for GWAS.

To determine the known lung function loci from previous literature, we extracted all significant loci from Shrine et al. [20] and searched for lung function-related keywords in the NHGRI-EBI GWAS Catalog (version `v1.0.2-associations_e106_r2022-07-09`) [29]. We used the following keywords for the Catalog search (case insensitive): "asthma", "chronic obstructive pulmonary disease", "copd", "expiratory flow", "fev1", "forced expiratory", "forced vital capacity", and "lung function".

4.6 Summary statistics conditional and joint analysis

We applied conditional and joint analysis (COJO) on set of previously known loci using GCTA (genome-wide complex trait analysis) software (version `1.93.3beta`) and we set `-cojo-cond` to the set of known loci. We provided 10,000 unrelated European samples randomly chosen from UKB as the reference samples, which is the same reference used to perform LD clumping to define hits, as part of GCTA-COJO input parameters.

4.7 PRS evaluation on respiratory diseases on multiple datasets

COPDGene dataset: COPDGene is a study of 10,300 current and former smokers with and without COPD, self-reported non-Hispanic White and African-American, without known lung diseases other than COPD and asthma (dbGaP accession phs000179.v6.p2). Additional study details, the study protocol and details of genotyping have been previously published [35, 57], and additionally detailed at copdgene.org. We used the provided variant calls in VCF files and imputed the variants to the Haplotype Reference Consortium (HRC) reference panel using Michigan Imputation Server [58], resulting in 39,127,678 total variants. Among 6,576 non-Hispanic White individuals, we had access to 1,131 (17%) asthma cases and 2,781 (42%) COPD cases, and the rest were used as controls. Meanwhile, among 3,140 African-American individuals, 760 (24%) were asthma cases and

802 (26%) were COPD cases.

EPIC-Norfolk dataset: The European Prospective Investigation into Cancer in Norfolk (EPIC-Norfolk) is a general population-based cohort study of men and women aged 40–79 years living in Norfolk, UK and recruited from general practices between 1993 to 1997. EPIC-Norfolk cohort participants were linked annually to nationally held hospital records and death certificates from 1999 to 2019 using UK National Health Service numbers. COPD was defined as any hospital admission or cause of death coded 490–492, 494–496 (ICD-9) or J40–J44, J47 (ICD-10). Asthma was similarly defined using codes 493 (ICD-9) or J45, J46 (ICD-10).

eMERGE III dataset: We utilize five consent groups that does not require IRB approval: General Research Use (GRU), Health/Medical/Biomedical-Genetic Studies, (HBM-GSO), Health/Medical/Biomedical (HMB), Health/Medical/Biomedical (MDS) HMB-MDS, and Health/Medical/Biomedical (PUB, GSO) (HMB-PUB-GSO) (dbGaP accession phs001584.v2.p2). In eMERGE III, we have access to 1,038 asthma cases and 7,250 controls for European ancestry while in the case of African ancestry we have access to 649 asthma cases and 1,367 controls. We used the 39,131,578 variants that are imputed to the HRC reference provided by dbGaP [59].

Indiana Biobank dataset: The Indiana Biobank is a state-wide collaboration that provides centralized processing and storage of specimens that are linked to participants' electronic medical information via Regenstrief Institute at Indiana University. COPD was diagnosed by using ICD9: 491, 492, and 496, and ICD10: J41, J42, J43, and J44. Asthma was diagnosed by using ICD9: 493, and ICD10: J45 and J46. Cases were defined as having at least one in-patient diagnosis or two out-patient diagnoses. Those participants not having any diagnosis were defined as controls. Thus, we have 1,445 COPD cases and 3,808 controls while we have 1,171 asthma cases and 4,083 controls. Among 5,253 individuals for COPD evaluation, 3797 were of European ancestry, 1,371 of African ancestry, and 85 Hispanic ancestry. Among 5,254 individuals for asthma evaluation, 3805 of European ancestry, 1,362 of African ancestry, and 87 of Hispanic ancestry. Indiana Biobank samples used in this study were genotyped using Illumina Infinium Global Screening Array (GSA, Illumina, San Diego, CA) by Regeneron (Tarrytown, NY). SNPs with missing rate $> 5\%$, MAF $\leq 1\%$, HWE P-value $< 1 \times 10^{-10}$ among cases and $< 1 \times 10^{-6}$ in controls were excluded as reported previously [37]. Genotyping data were imputed to 1000 Genomes using the Michigan Imputation Server [58]. Imputed variants with $R^2 < 0.30$ and MAF$< 1\%$ were excluded. PLINK [60, 61] was used to calculate PRS by using imputation dosages.

4.8 Functional significance of discovered loci

We ran GREAT v4.0.4 [31] on the human GRCh37 assembly to perform functional enrichment analysis of SPINCs, RSPINCs, and EDFs loci. We used the default "basal+extension" region-gene association rule with 5 kb upstream, 1 kb downstream, 1000 kb extension, and curated regulatory domains included. Furthermore, we ran GARFIELD [30] with default parameters to perform tissue-specific analysis where we utilized 424 DNase I hypersensitive site hotspot annotations provided by the GARFIELD authors [30].

4.9 Latent causal variable analysis

We applied LCV [42] (URLs) on genome-wide summary statistics for each pair of phenotypes. To create the right input for LCV, we used the munge script provided by S-LDSC software (URLs) to restrict the variants to HapMap3 SNPs with MAF > 0.05 and outside the MHC region. We utilized the baseline LD scores for HapMap3 SNPs. A two-sided test was used for the estimated GCP and to compute the significant level.

4.10 Genetic phenome-wide association study

To compute PheWAS, we downloaded GWAS summary statistics for 7,221 phenotypes from the Pan-UKB consortium 20200615 release (URLs). After restricting to phenotypes that contained European statistics and did not persistently fail in LD clumping, we were left with 7,145 pruning+thresholding PRSs generated by PLINK (URLs) using the `--clump` command with an index variant significance threshold of 5×10^{-8} and LD threshold of 0.1, with LD computed from a random subset of 10,000 European individuals in UK Biobank.

SPINCs and RSPINCs pruning+thresholding PRSs were computed analogously to the Pan-UKB PRSs. We computed the Pearson correlations between the PRSs derived from latent dimensions with the PRSs derived from Pan-UKB phenotypes and the P-values with two-sided alternative hypothesis.

URLs

Baseline and BaselineLD annotations: `https://data.broadinstitute.org/alkesgroup/ldscore`
BOLT-LMM software: `https://data.broadinstitute.org/alkesgroup/bolt-lmm`

COPDGene study: `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000179.v6.p2`

GCTA software: `https://github.com/jianyangqt/gcta`

GREAT software: `http://great.stanford.edu`

GWAS Catalog: `https://www.ebi.ac.uk/gwas/`

Indiana Biobank study: `https://indianabiobank.org/`

Pan-UK Biobank GWAS: `https://pan.ukbb.broadinstitute.org`

PLINK software: `https://www.cog-genomics.org/plink1.9`

TensorFlow: `https://www.tensorflow.org`

UCSC LiftOver: `https://genome.ucsc.edu/cgi-bin/hgLiftOver`

UK Biobank study: `https://www.ukbiobank.ac.uk`

XManager: `https://github.com/deepmind/xmanager`

Michigan Imputation Server `https://imputationserver.sph.umich.edu/index.html#!pages/home`

Data availability

Open-source code and trained model weights are available at `https://github.com/Google-Health/genomics-research` under the `spirogram-encodings` directory. SPINCs and RSPINCs values of UK Biobank individuals will be returned to UK Biobank and will be made available by UK Biobank.

Acknowledgments

We thank all participants, dataset creators, and maintainers of UK Biobank, COPDGene, eMERGE III, EPIC Norfolk, and Indiana Biobank. Full acknowledgement for each dataset can be found in "Dataset acknowledgment" in Supplementary Notes. We also thank Nick Furlotte for helpful discussions.

M.H.C. was supported by NHLBI R01HL153248, R01HL149861, and R01HL147148. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funding body has no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

References

[1] Lloyd T Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M Smith. Genome-wide association studies of brain imaging phenotypes in UK biobank. *Nature*, 562(7726):210–216, October 2018.

[2] Wenjia Bai, Hideaki Suzuki, Jian Huang, Catherine Francis, Shuo Wang, Giacomo Tarroni, Florian Guitton, Nay Aung, Kenneth Fung, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, Evangelos Evangelou, Abbas Dehghan, Declan P O'Regan, Martin R Wilkins, Yike Guo, Paul M Matthews, and Daniel Rueckert. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat. Med.*, 26(10):1654–1662, October 2020.

[3] Babak Alipanahi, Farhad Hormozdiari, Babak Behsaz, Justin Cosentino, Zachary R McCaw, Emanuel Schorsch, D Sculley, Elizabeth H Dorfman, Paul J Foster, Lily H Peng, Sonia Phene, Naama Hammel, Andrew Carroll, Anthony P Khawaja, and Cory Y McLean. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *Am. J. Hum. Genet.*, 108(7):1217–1230, July 2021.

[4] Nay Aung, Jose D Vargas, Chaojie Yang, Kenneth Fung, Mihir M Sanghvi, Stefan K Piechnik, Stefan Neubauer, Ani Manichaikul, Jerome I Rotter, Kent D Taylor, Joao A C Lima, David A Bluemke, Steven M Kawut, Steffen E Petersen, and Patricia B Munroe. Genome-wide association analysis reveals insights into the genetic architecture of right ventricular structure and function. *Nat. Genet.*, pages 1–9, June 2022.

[5] James P Pirruccello, Paolo Di Achille, Victor Nauffal, Mahan Nekoui, Samuel F Friedman, Marcus D R Klarqvist, Mark D Chaffin, Lu-Chen Weng, Jonathan W Cunningham, Shaan Khurshid, Carolina Roselli, Honghuang Lin, Satoshi Koyama, Kaoru Ito, Yoichiro Kamatani, Issei Komuro, Sean J Jurgens, Emelia J Benjamin, Puneet Batra, Pradeep Natarajan, Kenney Ng, Udo Hoffmann, Steven A Lubitz, Jennifer E Ho, Mark E Lindsay, Anthony A Philippakis, and Patrick T Ellinor. Genetic analysis of right heart structure and function in 40,000 people. *Nat. Genet.*, 54(6):792–803, June 2022.

[6] Justin Cosentino, Babak Behsaz, Babak Alipanahi, Zachary R McCaw, Davin Hill, Tae-Hwi Schwantes-An, Dongbing Lai, Andrew Carroll, Brian D Hobbs, Michael H Cho, Cory Y McLean,

and Farhad Hormozdiari. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. *Nat. Genet.*, April 2023.

[7] Niek Verweij, Jan-Walter Benjamins, Michael P Morley, Yordi J van de Vegte, Alexander Teumer, Teresa Trenkwalder, Wibke Reinhard, Thomas P Cappola, and Pim van der Harst. The genetic makeup of the electrocardiogram. *Cell Syst*, 11(3):229–238.e5, September 2020.

[8] Yoav Benjamini, Dan Drai, Greg Elmer, Neri Kafkafi, and Ilan Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research*, 125(1-2):279–284, nov 2001. doi: 10.1016/s0166-4328(01)00297-2. URL https://doi.org/10.1016%2Fs0166-4328%2801%2900297-2.

[9] David L. Streiner and Geoffrey R. Norman. Correction for multiple testing. *Chest*, 140(1):16–18, jul 2011. doi: 10.1378/chest.11-0523. URL https://doi.org/10.1378%2Fchest.11-0523.

[10] Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, November 1901.

[11] Hugues Aschard, Bjarni J Vilhjálmsson, Nicolas Greliche, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and Peter Kraft. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am. J. Hum. Genet.*, 94(5):662–676, May 2014.

[12] Xikun Han, Kaiah Steven, Ayub Qassim, Henry N. Marshall, Cameron Bean, Michael Tremeer, Jiyuan An, Owen M. Siggs, Puya Gharahkhani, Jamie E. Craig, Alex W. Hewitt, Maciej Trzaskowski, and Stuart MacGregor. Automated AI labeling of optic nerve head enables insights into cross-ancestry glaucoma risk and genetic discovery in >280, 000 images from UKB and CLSA. *The American Journal of Human Genetics*, 108(7):1204–1216, July 2021. doi: 10.1016/j.ajhg.2021.05.005. URL https://doi.org/10.1016/j.ajhg.2021.05.005.

[13] David P. Johns, Julia A.E. Walters, and E. Haydn Walters. Diagnosis and early detection of copd using spirometry. *Journal of Thoracic Disease*, 6(11), 2014. ISSN 2077-6624. URL https://jtd.amegroups.com/article/view/3088.

[14] Bartolome R. Celli. The importance of spirometry in copd and asthma: Effect on approach to management. *Chest*, 117(2, Supplement):15S–19S, 2000. ISSN 0012-3692. doi: https://doi.org/10.1378/chest.117.2_suppl.15S. URL https://www.sciencedirect.com/science/article/pii/S0012369215527505.

[15] M R Miller, J Hankinson, V Brusasco, F Burgos, R Casaburi, A Coates, R Crapo, P Enright, C P M van der Grinten, P Gustafsson, R Jensen, D C Johnson, N MacIntyre, R McKay, D Navajas, O F Pedersen, R Pellegrino, G Viegi, J Wanger, and ATS/ERS Task Force. Standardisation of spirometry. *Eur. Respir. J.*, 26(2):319–338, August 2005.

[16] David M Mannino and A Sonia Buist. Global burden of COPD: risk factors, prevalence, and future trends. *The Lancet*, 370(9589):765–773, September 2007. doi: 10.1016/s0140-6736(07)61380-4. URL https://doi.org/10.1016/s0140-6736(07)61380-4.

[17] Jørgen Vestbo, Suzanne S. Hurd, Alvar G. Agustí, Paul W. Jones, Claus Vogelmeier, Antonio Anzueto, Peter J. Barnes, Leonardo M. Fabbri, Fernando J. Martinez, Masaharu Nishimura, Robert A. Stockley, Don D. Sin, and Roberto Rodriguez-Roisin. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 187(4):347–365, feb 2013. doi: 10.1164/rccm.201204-0596pp.

[18] Edwin Silverman, Scott Weiss, Steven Shapiro, and David Lomas. *Respiratory genetics*. CRC Press, 2005.

[19] Nick Shrine, Anna L Guyatt, A Mesut Erzurumluoglu, Victoria E Jackson, Brian D Hobbs, Carl A Melbourne, Chiara Batini, Katherine A Fawcett, Kijoung Song, Phuwanat Sakornsakolpat, Xingnan Li, Ruth Boxall, Nicola F Reeve, Ma'en Obeidat, Jing Hua Zhao, Matthias Wielscher, Stefan Weiss, Katherine A Kentistou, James P Cook, Benjamin B Sun, Jian Zhou, Jennie Hui, Stefan Karrasch, Medea Imboden, Sarah E Harris, Jonathan Marten, Stefan Enroth, Shona M Kerr, Ida Surakka, Veronique Vitart, Terho Lehtimäki, Richard J Allen, Per S Bakke, Terri H Beaty, Eugene R Bleecker, Yohan Bossé, Corry-Anke Brandsma, Zhengming Chen, James D Crapo, John Danesh, Dawn L DeMeo, Frank Dudbridge, Ralf Ewert, Christian Gieger, Amund Gulsvik, Anna L Hansell, Ke Hao, Joshua D Hoffman, John E Hokanson, Georg Homuth, Peter K Joshi, Philippe Joubert, Claudia Langenberg, Xuan Li, Liming Li, Kuang Lin, Lars Lind, Nicholas Locantore, Jian'an Luan, Anubha Mahajan, Joseph C Maranville, Alison

Murray, David C Nickle, Richard Packer, Margaret M Parker, Megan L Paynton, David J Porteous, Dmitry Prokopenko, Dandi Qiao, Rajesh Rawal, Heiko Runz, Ian Sayers, Don D Sin, Blair H Smith, María Soler Artigas, David Sparrow, Ruth Tal-Singer, Paul R H J Timmers, Maarten Van den Berge, John C Whittaker, Prescott G Woodruff, Laura M Yerges-Armstrong, Olga G Troyanskaya, Olli T Raitakari, Mika Kähönen, Ozren Polašek, Ulf Gyllensten, Igor Rudan, Ian J Deary, Nicole M Probst-Hensch, Holger Schulz, Alan L James, James F Wilson, Beate Stubbe, Eleftheria Zeggini, Marjo-Riitta Jarvelin, Nick Wareham, Edwin K Silverman, Caroline Hayward, Andrew P Morris, Adam S Butterworth, Robert A Scott, Robin G Walters, Deborah A Meyers, Michael H Cho, David P Strachan, Ian P Hall, Martin D Tobin, Louise V Wain, and Understanding Society Scientific Group. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.*, 51(3):481–493, March 2019.

[20] Nick Shrine, Abril G Izquierdo, Jing Chen, Richard Packer, Robert J Hall, Anna L Guyatt, Chiara Batini, Rebecca J Thompson, Chandan Pavuluri, Vidhi Malik, Brian D Hobbs, Matthew Moll, Wonji Kim, Ruth Tal-Singer, Per Bakke, Katherine A Fawcett, Catherine John, Kayesha Coley, Noemi Nicole Piga, Alfred Pozarickij, Kuang Lin, Iona Y Millwood, Zhengming Chen, Liming Li, China Kadoorie Biobank Collaborative Group, Sara R A Wijnant, Lies Lahousse, Guy Brusselle, Andre G Uitterlinden, Ani Manichaikul, Elizabeth C Oelsner, Stephen S Rich, R Graham Barr, Shona M Kerr, Veronique Vitart, Michael R Brown, Matthias Wielscher, Medea Imboden, Ayoung Jeong, Traci M Bartz, Sina A Gharib, Claudia Flexeder, Stefan Karrasch, Christian Gieger, Annette Peters, Beate Stubbe, Xiaowei Hu, Victor E Ortega, Deborah A Meyers, Eugene R Bleecker, Stacey B Gabriel, Namrata Gupta, Albert Vernon Smith, Jian'an Luan, Jing-Hua Zhao, Ailin F Hansen, Arnulf Langhammer, Cristen Willer, Laxmi Bhatta, David Porteous, Blair H Smith, Archie Campbell, Tamar Sofer, Jiwon Lee, Martha L Daviglus, Bing Yu, Elise Lim, Hanfei Xu, George T O'Connor, Gaurav Thareja, Omar M E Albagha, Qatar Genome Program Research (QGPR) Consortium, Karsten Suhre, Raquel Granell, Tariq O Faquih, Pieter S Hiemstra, Annelies M Slats, Benjamin H Mullin, Jennie Hui, Alan James, John Beilby, Karina Patasova, Pirro Hysi, Jukka T Koskela, Annah B Wyss, Jianping Jin, Sinjini Sikdar, Mikyeong Lee, Sebastian May-Wilson, Nicola Pirastu, Katherine A Kentistou, Peter K Joshi, Paul R H J Timmers, Alexander T Williams, Robert C Free, Xueyang Wang, John L Morrison, Frank D Gilliland, Zhanghua Chen, Carol A Wang, Rachel E Foong, Sarah E Harris, Adele Taylor, Paul Redmond, James P Cook, Anubha Mahajan, Lars Lind, Teemu Palviainen,

Terho Lehtimäki, Olli T Raitakari, Jaakko Kaprio, Taina Rantanen, Kirsi H Pietiläinen, Simon R Cox, Craig E Pennell, Graham L Hall, W James Gauderman, Chris Brightling, James F Wilson, Tuula Vasankari, Tarja Laitinen, Veikko Salomaa, Dennis O Mook-Kanamori, Nicholas J Timpson, Eleftheria Zeggini, Josée Dupuis, Caroline Hayward, Ben Brumpton, Claudia Langenberg, Stefan Weiss, Georg Homuth, Carsten Oliver Schmidt, Nicole Probst-Hensch, Marjo-Riitta Jarvelin, Alanna C Morrison, Ozren Polasek, Igor Rudan, Joo-Hyeon Lee, Ian Sayers, Emma L Rawlins, Frank Dudbridge, Edwin K Silverman, David P Strachan, Robin G Walters, Andrew P Morris, Stephanie J London, Michael H Cho, Louise V Wain, Ian P Hall, and Martin D Tobin. Multi-ancestry genome-wide association analyses improve resolution of genes and pathways influencing lung function and chronic obstructive pulmonary disease risk. *Nat. Genet.*, 55(3):410–422, March 2023.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013. doi: 10.48550/ARXIV.1312.6114. URL `https://arxiv.org/abs/1312.6114`.

[22] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[24] Brian D Hobbs, Kim de Jong, Maxime Lamontagne, Yohan Bossé, Nick Shrine, María Soler Artigas, Louise V Wain, Ian P Hall, Victoria E Jackson, Annah B Wyss, Stephanie J London, Kari E North, Nora Franceschini, David P Strachan, Terri H Beaty, John E Hokanson, James D Crapo, Peter J Castaldi, Robert P Chase, Traci M Bartz, Susan R Heckbert, Bruce M Psaty, Sina A Gharib, Pieter Zanen, Jan W Lammers, Matthijs Oudkerk, H J Groen, Nicholas Locantore, Ruth Tal-Singer, Stephen I Rennard, Jørgen Vestbo, Wim Timens, Peter D Paré, Jeanne C Latourelle, Josée Dupuis, George T O'Connor, Jemma B Wilk, Woo Jin Kim, Mi Kyeong Lee, Yeon-Mok Oh, Judith M Vonk, Harry J de Koning, Shuguang Leng, Steven A Belinsky, Yohannes Tesfaigzi, Ani Manichaikul, Xin-Qun Wang, Stephen S Rich, R Graham Barr, David Sparrow, Augusto A Litonjua, Per Bakke, Amund Gulsvik, Lies Lahousse,

Guy G Brusselle, Bruno H Stricker, André G Uitterlinden, Elizabeth J Ampleford, Eugene R Bleecker, Prescott G Woodruff, Deborah A Meyers, Dandi Qiao, David A Lomas, Jae-Joon Yim, Deog Kyeom Kim, Iwona Hawrylkiewicz, Pawel Sliwinski, Megan Hardin, Tasha E Fingerlin, David A Schwartz, Dirkje S Postma, William MacNee, Martin D Tobin, Edwin K Silverman, H Marike Boezen, Michael H Cho, COPDGene Investigators, ECLIPSE Investigators, LifeLines Investigators, SPIROMICS Research Group, International COPD Genetics Network Investigators, UK BiLEVE Investigators, and International COPD Genetics Consortium. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.*, 49(3):426–432, March 2017.

[25] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779, March 2015.

[26] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, 47(3):284–290, March 2015.

[27] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P Schoech, and Alkes L Price. Mixed-model association for biobank-scale datasets. *Nat. Genet.*, 50(7):906–908, July 2018.

[28] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, March 2015.

[29] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, Arwa Ibrahim, Yue Ji, Sajo John, Elizabeth Lewis, Jacqueline A L MacArthur, Aoife McMahon, David Osumi-Sutherland, Kalliope Panoutsopoulou, Zoë Pendlington, Santhi Ramachandran, Ray Stefancsik, Jonathan Stewart, Patricia Whetzel, Robert Wilson, Lucia Hindorff, Fiona Cunningham, Samuel A Lambert, Michael In-

ouye, Helen Parkinson, and Laura W Harris. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res.*, 51(D1):D977–D985, January 2023.

[30] Valentina Iotchkova, Graham R S Ritchie, Matthias Geihs, Sandro Morganella, Josine L Min, Klaudia Walter, Nicholas John Timpson, UK10K Consortium, Ian Dunham, Ewan Birney, and Nicole Soranzo. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.*, 51(2):343–353, February 2019.

[31] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, 28(5):495–501, May 2010.

[32] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Pamela A F Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael N Weedon, Ruth J Loos, Timothy M Frayling, Mark I McCarthy, Joel N Hirschhorn, Michael E Goddard, and Peter M Visscher. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, 44(4):369–75, S1–3, March 2012.

[33] Zachary R McCaw, Jacqueline M Lane, Richa Saxena, Susan Redline, and Xihong Lin. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, 76(4):1262–1272, December 2020.

[34] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, 51(4):584–591, April 2019.

[35] Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of COPD (COPDGene) study design. *COPD*, 7(1):32–43, February 2010.

[36] N Day, S Oakes, R Luben, K T Khaw, S Bingham, A Welch, and N Wareham. EPIC-Norfolk: study design and characteristics of the cohort. european prospective investigation of cancer. *Br. J. Cancer*, 80 Suppl 1:95–103, July 1999.

[37] Dongbing Lai, Tae-Hwi Schwantes-An, Marco Abreu, Grace Chan, Victor Hesselbrock, Chella Kamarajan, Yunlong Liu, Jacquelyn L. Meyers, John I. Nurnberger, Martin H. Plawecki, Leah Wetherill, Marc Schuckit, Pengyue Zhang, Howard J. Edenberg, Bernice Porjesz, Arpana Agrawal, and Tatiana Foroud. Gene-based polygenic risk scores analysis of alcohol use disorder in african americans. *Translational Psychiatry*, 12(1), July 2022. doi: 10.1038/s41398-022-02029-2. URL `https://doi.org/10.1038/s41398-022-02029-2`.

[38] Kari Hemminki, Xiangdong Liu, Jianguang Ji, Kristina Sundquist, and Jan Sundquist. Subsequent COPD and lung cancer in patients with autoimmune disease. *European Respiratory Journal*, 37(2):463–465, 2011.

[39] Te-Chun Shen, Cheng-Li Lin, Chia-Hung Chen, Chih-Yen Tu, Te-Chun Hsia, Chuen-Ming Shih, Wu-Huei Hsu, and Yen-Jung Chang. Increased risk of chronic obstructive pulmonary disease in patients with systemic lupus erythematosus: a population-based cohort study. *PLOS ONE*, 9(3):e91821, 2014.

[40] Dan Huang, Dong Wu, Jinhong He, Min Chen, Xuanna Zhao, Dongming Li, and Bin Wu. Association between thyroid function and acute exacerbation of chronic obstructive pulmonary disease. *International Journal of Chronic Obstructive Pulmonary Disease*, 16:333, 2021.

[41] Jonas F Ludvigsson, Malin Inghammar, Marie Ekberg, and Arne Egesten. A nationwide cohort study of the risk of chronic obstructive pulmonary disease in coeliac disease. *Journal of Internal Medicine*, 271(5):481–489, 2012.

[42] Luke J O'Connor and Alkes L Price. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.*, 50(12):1728–1734, December 2018.

[43] Cathryn M Delude. Deep phenotyping: the details of disease. *Nature*, 527(7576):S14–S15, 2015.

[44] Matthias Kirchler, Stefan Konigorski, Matthias Norden, Christian Meltendorf, Marius Kloft, Claudia Schurmann, and Christoph Lippert. transferGWAS: GWAS of images using deep transfer learning. *Bioinformatics*, 38(14):3621–3628, July 2022.

[45] Ziqian Xie, Tao Zhang, Sangbae Kim, Jiaxiong Lu, Wanheng Zhang, Cheng-Hui Lin, Man-Ru Wu, Alexander Davis, Roomasa Channa, Luca Giancardo, Han Chen, Sui Wang, Rui Chen,

and Degui Zhi. iGWAS: image-based genome-wide association of self-supervised deep phenotyping of human medical images. *medRxiv*, 2022. doi: 10.1101/2022.05.26.22275626. URL https://doi.org/10.1101/2022.05.26.22275626.

[46] Adityanarayanan Radhakrishnan, Sam Freesun Friedman, Shaan Khurshid, Kenney Ng, Puneet Batra, Steven Lubitz, Anthony Philippakis, and Caroline Uhler. A cross-modal autoencoder framework learns holistic representations of cardiovascular state. *bioRxiv*, 2022. doi: 10.1101/2022.05.26.493497. URL https://doi.org/10.1101/2022.05.26.493497.

[47] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[48] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/larsen16.html.

[49] Jin J Zhou, Michael H Cho, Christoph Lange, Sharon Lutz, Edwin K Silverman, and Nan M Laird. Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. *Hum. Hered.*, 79(2):93–104, June 2015.

[50] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

[51] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

[52] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of dis-

entangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1kG7GZAW.

[53] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[54] Davin Hill, Max Torop, Aria Masoomi, Peter J Castaldi, Edwin K Silverman, Sandeep Bodduluri, Surya P Bhatt, Taedong Yun, Farhad Hormozdiari, Cory Y McLean, Jennifer Dy, Michael H Cho, and Brian D Hobbs. Deep learning utilizing suboptimal spirometry data to improve lung function and mortality prediction in the UK Biobank. *Preprint*, 2023.

[55] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[56] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale machine learning on heterogeneous distributed systems. March 2016.

[57] Michael H Cho, Merry-Lynn N McDonald, Xiaobo Zhou, Manuel Mattheisen, Peter J Castaldi, Craig P Hersh, Dawn L Demeo, Jody S Sylvia, John Ziniti, Nan M Laird, Christoph Lange, Augusto A Litonjua, David Sparrow, Richard Casaburi, R Graham Barr, Elizabeth A Regan, Barry J Make, John E Hokanson, Sharon Lutz, Tanda Murray Dudenkov, Homayoon Farzadegan, Jacqueline B Hetmanski, Ruth Tal-Singer, David A Lomas, Per Bakke, Amund Gulsvik, James D Crapo, Edwin K Silverman, Terri H Beaty, and NETT Genetics, ICGN, ECLIPSE and COPDGene Investigators. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med*, 2(3):214–225, March 2014.

[58] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stam-

bolian, Po-Ru Loh, William G Iacono, Anand Swaroop, Laura J Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonçalo R Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287, August 2016. doi: 10.1038/ng.3656. URL https://doi.org/10.1038/ng.3656.

[59] Ian B Stanaway, Taryn O Hall, Elisabeth A Rosenthal, Melody Palmer, Vivek Naranbhai, Rachel Knevel, Bahram Namjou-Khales, Robert J Carroll, Krzysztof Kiryluk, Adam S Gordon, Jodell Linder, Kayla Marie Howell, Brandy M Mapes, Frederick T J Lin, Yoonjung Yoonie Joo, M Geoffrey Hayes, Ali G Gharavi, Sarah A Pendergrass, Marylyn D Ritchie, Mariza de Andrade, Damien C Croteau-Chonka, Soumya Raychaudhuri, Scott T Weiss, Matt Lebo, Sami S Amr, David Carrell, Eric B Larson, Christopher G Chute, Laura Jarmila Rasmussen-Torvik, Megan J Roy-Puckelwartz, Patrick Sleiman, Hakon Hakonarson, Rongling Li, Elizabeth W Karlson, Josh F Peterson, Iftikhar J Kullo, Rex Chisholm, Joshua Charles Denny, Gail P Jarvik, eMERGE Network, and David R Crosslin. The eMERGE genotype set of 83,717 subjects imputed to 40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet. Epidemiol.*, 43(1):63–81, February 2019.

[60] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, September 2007. doi: 10.1086/519795. URL https://doi.org/10.1086/519795.

[61] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), February 2015. doi: 10.1186/s13742-015-0047-8. URL https://doi.org/10.1186/s13742-015-0047-8.