Polygenic risk modeling with latent trait-related genetic components

Matthew Aguirre[1,2], Yosuke Tanigawa[1], Guhan Venkataraman[1], Johanne Marie Justesen[1,3], Rob Tibshirani[1,4], Trevor Hastie[1,4], Manuel A. Rivas[1+]

Abstract:

Polygenic risk models have led to significant advances in understanding complex diseases and their clinical presentation. While simple linear models of genetic risk like polygenic risk scores can effectively predict outcomes, they do not generally account for disease subtypes or pathways which underlie within-trait diversity. Here, we introduce a latent factor model of genetic risk based on components from matrix Decomposition of Genetic Associations (DeGAs), and highlight results for body mass index (BMI), heart attack/myocardial infarction (MI), and gout in a held out test set of 67,430 white British individuals from UK Biobank. We find the most at-risk individuals (2% percentile) have 2.5-kg/m^2 higher BMI, 3.5-fold risk of MI, and 4.8-fold risk for gout compared to the general population. We further characterize individual and population-wide genetic risk profiles for each trait, and replicate these trends in an independent set of 25,486 non-British white participants of the Biobank.

[1]Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, CA, 94305, USA.
[2]Department of Pediatrics, School of Medicine, Stanford University, Stanford, CA, 94305, USA.
[3]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
[4]Department of Statistics, Stanford University, Stanford, CA, 94305, USA.

+Correspondence to Manuel A. Rivas (mrivas@stanford.edu).

Introduction:

Common conditions like diabetes and heart disease are leading causes of death and financial burden in the developed world[1]. Polygenic risk scores (PRS), which sum the contributions of multiple risk loci, have been used with some success to identify individuals at high risk for diseases like cancer[2–4], diabetes[5,6], heart disease[7,8], and obesity[9,10]. Indeed, the performance of polygenic models has been subject to much scrutiny and methods development. Although many versions of PRS can be used to estimate risk[11–13], previous work has suggested that a "palette" model which decomposes genetic risk into its constituent pathways may more faithfully describe the clinical manifestations of complex disease[14].

Here, we present a polygenic model based on latent trait-related genetic components identified using Decomposition of Genetic Associations (DeGAs)[15]. Rather than modeling genetic risk for disease as a sum of effects from contributing genetic variants, the DeGAs polygenic risk score (dPRS) models the effects of DeGAs components. Each component has genetic determinants which can be expressed as a component PRS (cPRS). We then use these scores to estimate personalized genetic risk profiles which inform genetic subtyping of disease.

As proof of concept, we compute DeGAs using 1,905 distinct traits and 454,565 independent common variants (see Methods) in UK Biobank. We characterize components of genetic risk and build dPRS models in a subset of unrelated white British individuals (n=236,005). We then evaluate their performance in an independent sample of the same population (n=67,430), and in UK Biobank non-British Europeans (n=25,486). We highlight results for body mass index (BMI), myocardial infarction (MI/heart attack), and gout, motivated by their high polygenicity and prevalence in the general population. Weights for all models are available on the Global Biobank Engine[16] (biobankengine.stanford.edu).

Results:

Performance of the DeGAs Polygenic Risk Score (dPRS):

Trait-related genetic components were identified using DeGAs[15] on a set of input z-statistics from nominally significant ($p < 0.01$) genetic associations for a subset of traits on the Global Biobank Engine[16]. These GWAS were computed in a population of 236,005 unrelated white British individuals from UK Biobank study. We used m=454,565 independent common variants and n=1,905 traits (**Supplementary Table 1**) for DeGAs. Performance for other iterations of DeGAs with beta-values, different p-value filters, and numbers of components (see Methods for hyperparameter optimization) are presented in **Supplementary Figure 1**, and weights for each model are online on the Global Biobank Engine.

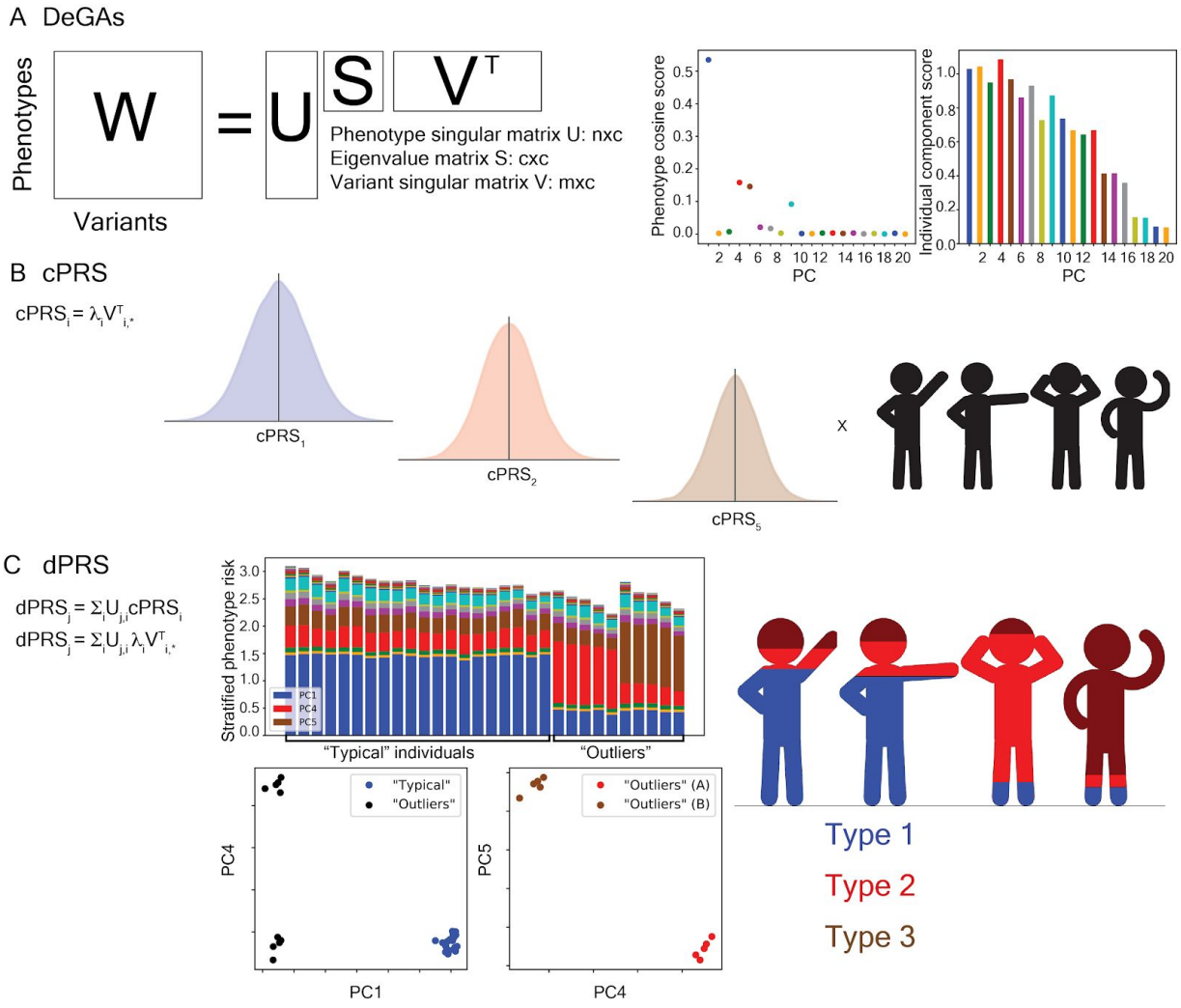


Figure 1: Study overview. (a) Matrix Decomposition of Genetic Associations (DeGAs) is performed by taking the truncated singular value decomposition (TSVD) of a matrix *W* containing summary statistics from GWAS of *n*=1,905 traits over *m*=454,565 variants from UK Biobank. The columns of the resulting singular matrices *U* (*n x c*) and *V* (*m x c*) measure the relative importance of traits (variants) to each component; the rows map traits (variants) back to components. **(b) Component polygenic risk scores (cPRS)** are computed using eigenvalue-endowed variant contribution scores $\lambda_i V_{i,*}$ (*i*th eigenvalue in *S* and *i*th row in V^T). *G* is a vector of genotypes for the *m* variants used in DeGAs. **(c) DeGAs polygenic risk scores (dPRS)** are recovered by taking the dot product of cPRS values with the phenotype cosine score vector $U_{j,*}$ (*j*th row in *U*). We also compute DeGAs risk profiles (see Methods), which measures the relative contribution of each component to the genetic risk for every individual. High risk individuals are labeled “typical” or “outliers”, then clustered on their risk profiles to identify genetic subtypes. Image credit: VectorStock.com/1143365.

As an initial assessment of dPRS, we estimated disease prevalence (or mean BMI) at several quantiles of risk in a held-out test set of self-reported white British individuals in UK Biobank (*n*=67,430). For all traits we observed increasing severity (quantitative) or prevalence (binary) at increasing quantiles of their respective dPRS (**Figure 2a-c**) adjusted for age, sex, and 4 components of genetic ancestry. This trend is most pronounced at the highest risk quantile (2%) for each trait. At this stratum we observe 2.4-kg/m^2 higher average BMI; 3.9-fold enrichment of myocardial infarction prevalence; and 4.9-fold increase in gout. However, dPRS adds comparatively little population-wide predictive value relative to factors such as age, sex, and demographic effects captured by genomic PCs (Figure 2d-f). At the population level, we find *r*=0.207 between dPRS and residual BMI adjusted for covariates, as well as an AUC of 0.560 for MI and 0.648 for gout using unadjusted dPRS as the classifying score.

Further, we find dPRS to be comparable to prune- and threshold-based PRS using the same input data (**Supplementary Table 2**). The top 2% of risk for covariate-adjusted PRS for each of these traits has 2.5-kg/m^2 higher BMI; 3.5-fold risk for MI; and 4.8-fold risk for gout. Population-wide predictive measures are also similar, with BMI residual *r*=0.205, and PRS AUC (not adjusted for covariates) 0.549 for MI and 0.603 for gout. Despite the reduced rank of the DeGAs risk models — the input matrix *W* is reduced from ~1,900 traits to a 300-dimensional representation — we achieve performance equivalent to full rank PRS for these selected traits.

Painting DeGAs Risk Profiles:

To further characterize the polygenic architecture of these traits, we “painted” the profiles of each individual in the tenth percentile of trait dPRS adjusted for age and four genetic principal components. For a given individual, the DeGAs risk profile measures the relative importance of each DeGAs component in driving their risk for some trait (see Methods). Meanwhile the phenotype/trait cosine score[15] — which is a squared row of the *U* matrix from DeGAs (**Figure 1**; **Figure 3g-i**) — describes the components which contribute to a trait based on the input GWAS summary statistics.

Body mass index is a highly polygenic trait with associated genetic variation relevant to adipogenesis, insulin secretion, energy metabolism, and synaptic function[17]. Here, the DeGAs phenotype cosine score for BMI indicates strong contribution from components related to body size and lean mass (PC1 - 7.0%), fat mass (PC3 - 13.7%), and overall health indicators like sleep duration and alcohol and water intake (PC169 - 11.2%). Genic variation in *FTO* and *DLEU1* is most represented among coding variants in PC1; both genes are strongly associated with traits affecting body size in adults, the former being a well-established obesity susceptibility gene[18] and the latter a tumor-suppressing lncRNA often deleted chronic lymphocytic leukemia[19]. *DLEU1* variants are also significant contributors to PC3 along with *TSBP1* (or *C6orf10*), an open reading frame in the human leukocyte antigen (HLA) region.

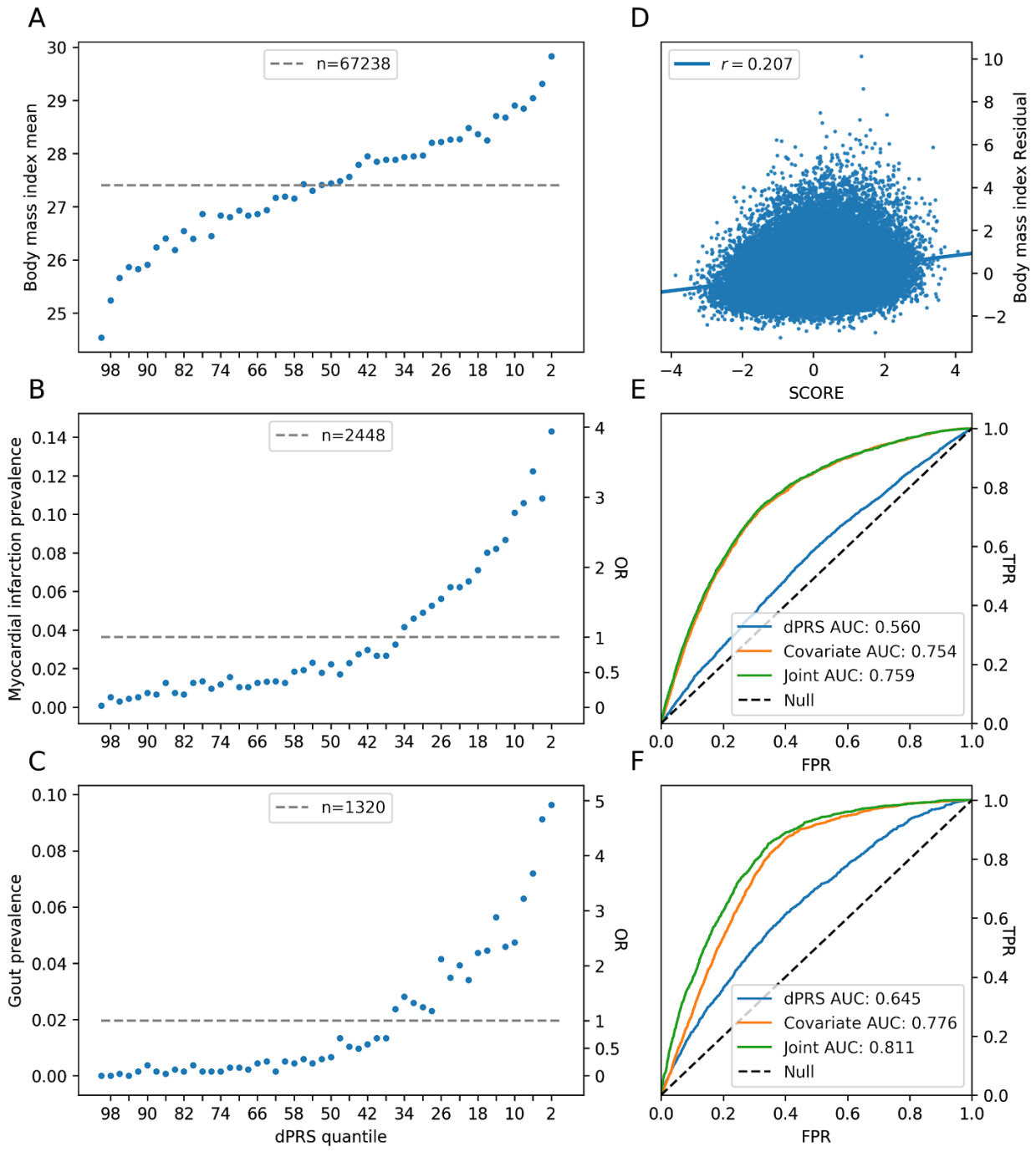


Figure 2: Performance of dPRS. (Left) Mean BMI and binary disease prevalence (MI/gout) at risk quantiles (2% increments) of dPRS in the 20% test set. dPRS is adjusted for these covariates: age, sex, and 4 genomic PCs. Dashed line is the population mean/prevalence, which is the baseline for odds-ratios calculated in each 2% quantile; *n* is the total number of quantitative values or cases for each trait. **(Right)** Correlation between dPRS and residual BMI adjusted for covariates. ROC curves with AUC (MI/gout) for dPRS, covariates alone, and joint model with dPRS and covariates. Covariate/joint models were fit in the 10% validation population, and all evaluation is conducted in the 20% test set (Methods).

Myocardial infarction is similarly polygenic with well established risk factors attributable to common and rare genetic variation[8], age, sex, and lifestyle attributes like diet and smoking. Key DeGAs components for this trait are related to an array of covariate- and statin-adjusted blood and urine biomarkers. These include urea and gamma glutamyltransferase (PC135 - 5.6%), and phosphate (PC133 and PC142 — 5.2% and 8.9% — these components also have contribution from cholesterol medications). Another relevant component (PC147 - 5.4%) has contribution from phosphate and hair color. All three components have significant contribution from variation in the lipoprotein genes *LPA* and *APOC1*, along with variants at the *9p21.3* susceptibility locus (*CDKN2B*) and brain-expressed solute carrier *SLC22A3*[20].

Gout is a heritable (h^2=17.0-35.1%) common complex form of arthritis characterized by severe sudden onset joint pain and tenderness, believed to arise due to excessive blood uric acid which crystallizes and forms deposits in the joints[21]. The top three DeGAs components for gout share strong contribution from covariate-adjusted blood urate. One component (PC94 - 7.0%) is further driven by intraocular pressure and fat percentage; another (PC108 - 8.1%) is related to visual acuity (logMAR test results); and the third (PC117 - 6.6%) is driven mainly by covariate- and statin-adjusted cystatin C and abdominal fat. Shared among all components is genetic variation in *SLC2A9*, which is involved in uric acid transport and has been associated with gout[22]. The transporter protein *ABCG2* is also key to both PC94 and PC108, and has been shown to play a role in renal urate transport[23]. PC117 is primarily driven by the cystatin gene family members *CST9*, *CST4*, and *CST1*, which are adjacent to one another on chromosome 20 and associate with renal function and chronic kidney disease[24].

Due to patterns of genetic variation within the input population, the phenotype cosine score does not always reflect the key risk components for the highest risk individuals, as seen in Figure 3a-c. This is expected behavior, as the score evenly weights all components. Meanwhile the population-wide variance of a component (and hence, its cPRS), tracks with its eigenvalue from DeGAs, resulting in stronger contributions from lower-numbered components among high risk individuals regardless of trait. For example, the phenotype cosine score for BMI suggests that PC1, PC3, and PC169 are the top 3 components; but PC6 (autoimmune traits) and PC7 (standing height and body impedance) are more significant drivers of risk at the extreme tail of population risk.

Moreover, significant polygenic diversity is present among high risk individuals via the DeGAs risk profile. To highlight this, we used the Mahalanobis criterion (distance to population centroid) to find individuals in the entire test population whose risk profiles significantly differed from average. We then intersected these outliers (z-scored Mahalanobis distance > 2) with the high risk individuals (10th percentile of covariate-adjusted dPRS) to identify “high risk outliers”. In general, this group (**Figure 3d-f**) has similar contributing components as high risk individuals (**Figure 3a-c**), but their relative importance within each of the individuals differs from population average. For example, high risk outliers for BMI tend to have strong contribution from either the fat mass (PC3) or lean mass component (PC1). On average this resembles a “typical” high risk

individual whose profile is a mixture of both contributing components, even though each individual is poorly described by that mixture.

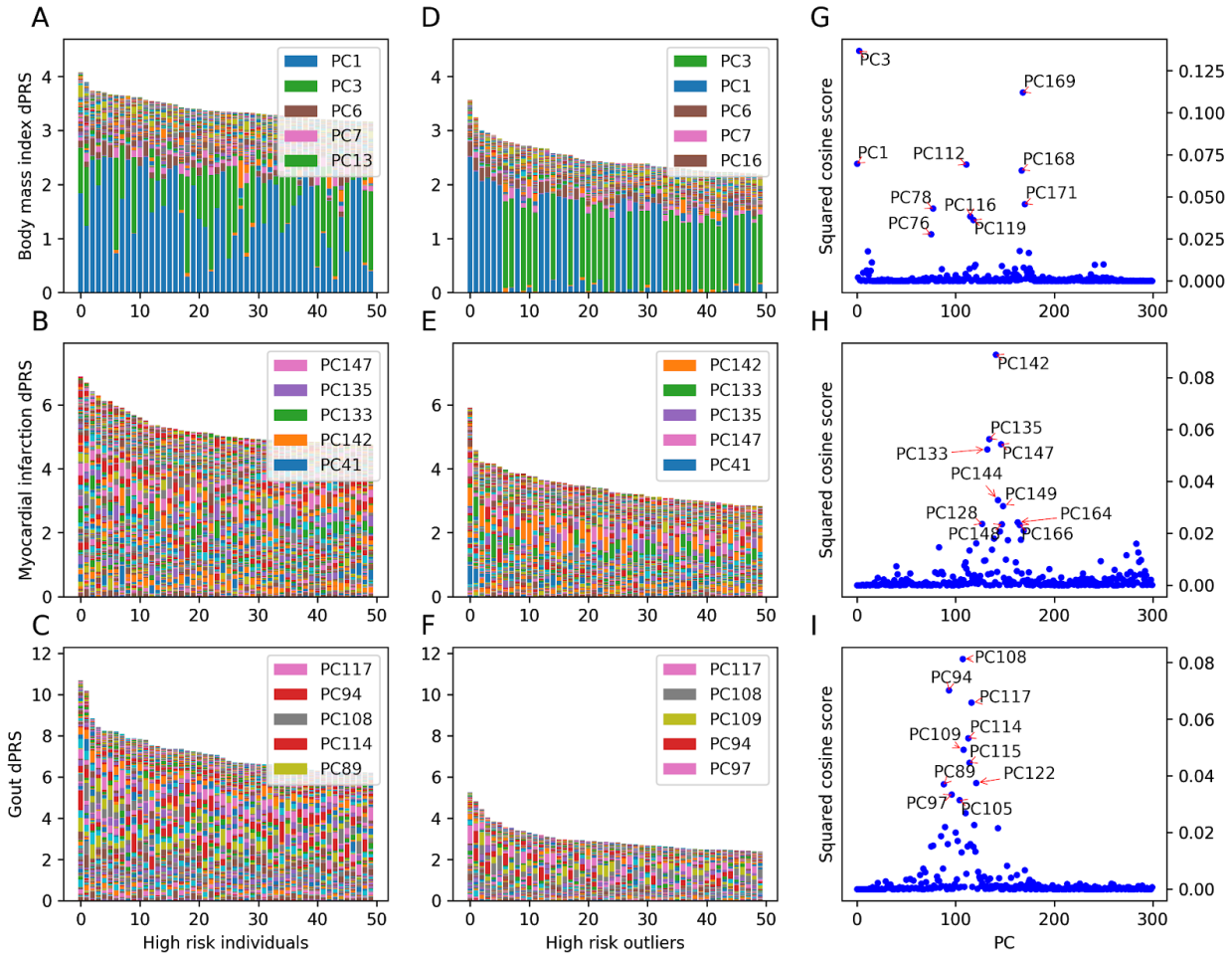


Figure 3: Painting components of genetic risk. (Left) Component-painted risk for the 50 individuals or **(Center)** outliers with highest dPRS for each trait among the 20% test set of white British unrelated individuals. Each bar represents one individual, and the colored components of the barplot are the normalized contribution made by each component on that individual's risk score. Colors for the 5 most prevalent components in each box are shown in its respective legend. **(Right)** Scatterplot of phenotype squared cosine scores denoting the relative importance of population components for each trait in the training set.

Genetic subtyping with DeGAs Risk Profiles:

To better describe the within-trait diversity of genetic risk profiles, we attempted to re-identify genetic subtypes of each trait using the high risk outlier population. We therefore performed a *k*-means clustering of this group using their DeGAs risk profiles as the distance metric; *k* was chosen using an iterative approach based on marginal increase in variance explained resulting from incrementing the number of clusters (**Methods**). We describe each cluster using its median

risk profile (**Figure 4a-c**), and further noticed that cluster membership divides individuals based on cPRS for relevant components (**Figure 4d-f**).

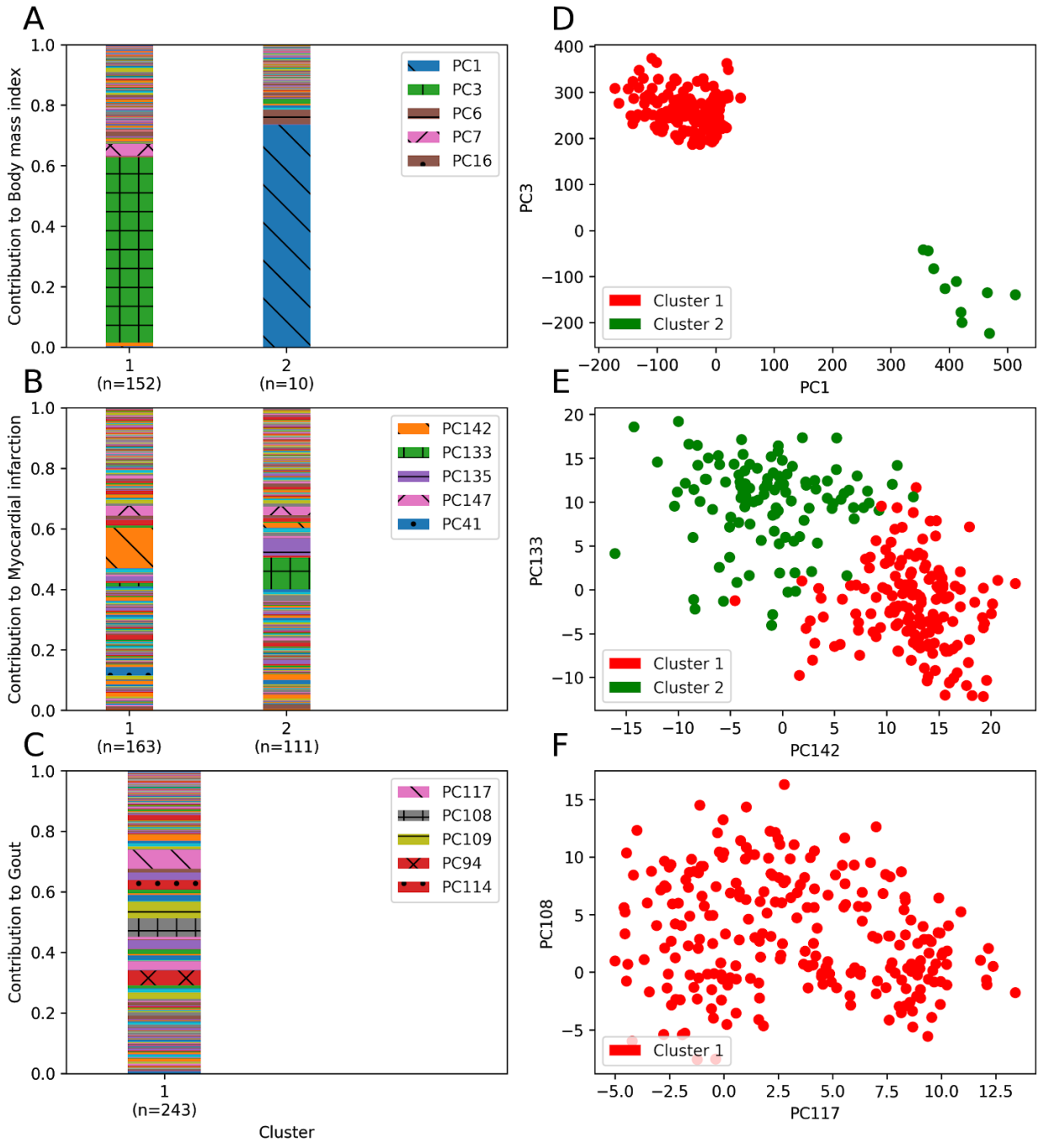


Figure 4: Clustering profiles of genetic risk. (Left) Mean risk profile for each cluster, labeled with the number of included individuals, in a *k*-means clustering of the risk profiles of "atypical" individuals with high dPRS in the validation population. **(Right)** cPRS weighted by phenotype contribution score for two representative PCs among the individuals represented in the left clusters. Each individual is represented by a dot and is colored by cluster membership.

For body mass index, we identify two risk clusters (**Figure 4a**): one driven by the fat mass component (PC3 - 61.3%, *n*=152) and the other by the lean mass component (PC1 - 73.6%, *n*=10). Most outlying individuals at risk for high BMI have genetic contribution from the near exclusively fat-related component (PC3), hence their deviation from “typical”. However, a minority of outliers display the opposite. Genetic risk from this cluster comes mainly from variant loadings related to lean-mass related traits like whole-body water and fat-free mass. While smaller in number, the existence of this cluster and its wide separation from other outliers at high risk for high BMI suggest alternative preventative and therapeutic approaches.

We find two clusters for myocardial infarction, both driven by components with strong contribution from blood phosphate and cholesterol lowering medications, which are respectively markers of cardiac calcification and atherosclerotic buildup (**Figure 4b**). One risk cluster component (PC142 - 13.3%, *n*=163) is further characterized by hair color, number of illnesses, and sitting height. In addition to genetic contribution from *LPA* and *APOC1*, this component has high loading from variation in the Fanconi anemia complement group gene *FANCA* and the melanocyte-specific transport gene *OCA2*. The other cluster’s components (PC133 and PC135 - 10.6% and 5.8%, *n*=111) has contribution from blood (gamma glutamyltransferase, urea) and urine (enzymatic creatinine, sodium) biomarkers, as well as markers of cardiac output like pulse wave stiffness and amplitude. Relevant genes for these components include *APOB* and the lipoprotein (a) associated transporter *SLC22A2*[25].

There is only one cluster of outliers for gout (**Figure 4c**), and the average risk profile closely mirrors the driving components described phenotype cosine score (PC94, PC108, PC117). Of note is the increased importance of PC109, which is driven by platelet (thrombocyte) volume and width, covariate-adjusted blood urate, and brain MRI measures (T2-star) which capture hemorrhaging. Interestingly, increased uric acid is associated with cerebral microbleeds in stroke patients[26], which indicates further shared biology associated with the biomarker. While dPRS is highly predictive for gout (**Figure 2c,f**), there appears to be insufficient diversity of genetic risk for clusters of DeGAs risk profiles to emerge. This likely owes to risk for gout being driven by a single biomarker, urate, which may act by similar mechanisms for most patients.

Discussion:

In this study we show that dPRS adds an interpretable dimension to traditional polygenic risk models by tying together disease, lifestyle, and biomarker-level data in trait-related genetic components. Predicting genetic risk with these components allowed us to infer disease pathology beyond variant-trait associations, and with minimal loss of predictive power from using a limited number of components. As validation, we presented a DeGAs risk model built on data from white British individuals in UK Biobank, which recovered known etiologies of BMI and offered further insight into the biology of diabetes and MI. These models remained predictive in an independent test set of individuals with similar ancestry as well as non-British white individuals (**Supplementary Figure 4**; all traits in **Supplementary Table 1**) from the Biobank.

Although comparable in performance using the same variants, our method has limitations compared to traditional PRS. Since TSVD does not handle gross correlation in both dimensions (i.e., across both traits and variants), we LD-pruned sites prior to analysis with DeGAs. This might cause dPRS to overfit patterns of LD in the GWAS population, as many associations in the input data will be due to correlation with causal variants. Although dPRS is predictive in independent samples with varying degrees of similarity in ethnic composition, concerns about the generalizability of PRS across populations are equally relevant here.

It is also worthwhile to consider effects due to our choices of analysis traits. Here, we avoided traits whose genetic signals may be difficult to detect (low *n*) or might be confounded (markers of socioeconomic status). This limits the patterns we can identify using DeGAs risk profiles. In a similar way, it may be the case that there is a “typical” etiology well-represented by the high risk population centroid. Taking gout as an example, we only one cluster of outliers (**Figure 4c**) whose median DeGAs risk profile mirrors the key components identified by the phenotype cosine score for the trait (**Figure 3i**).

We anticipate the application space for pathway-aware models of polygenic risk like dPRS will be broad. As DeGAs requires only summary-level data, traits with predictive underlying quantitative features or biomarkers would be good candidates for follow-up study. Examples include metabolomics, and brain and liver imaging, many of which have been collected in population-scale resources like UK Biobank. Such analysis could leverage the genetics of biomarkers and intermediate quantitative traits to further elucidate the underlying diversity of polygenic risk for complex traits.

Methods:

Study populations

UK Biobank is a large longitudinal cohort study consisting of 502,560 individuals aged 37-73 at recruitment during 2006-2010. The data acquisition and study development protocols are online (http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf). In short, participants visited a nearby center for an in-person baseline assessment where various anthropometric data, blood samples, and survey questionnaire responses were collected. Additional data were linked from registries and collected during follow-up visits.

Here, we use a subsample consisting of 337,151 unrelated individuals of self-reported white British ancestry for genetic analysis. We split this cohort at random into three groups: a 70% training population (*n*=236,005), a 10% validation population (*n*=33,716), and a 20% test population (*n*=67,430). We use the training population to conduct genome-wide association studies for DeGAs, and the validation population to evaluate model performance for selecting DeGAs hyperparameters. We also use the validation population to fit the joint dPRS model with age, sex, and 4 genomic components. We report final associations and performance measures in the test population. An additional sample of *n*=25,486 unrelated non-British whites is used as an "external" evaluation set with different ancestry. This group consists of individuals used in the UK Biobank PCA calculation who self-report white but not British ancestry in Field 21000.

Genome-wide association studies in the UK Biobank:

PLINK v2.00a[27] (multiple dated versions) was used for genome-wide associations over 805,426 directly genotyped variants, 362 HLA allelotypes, and 1,815 non-rare (AF > 0.01%) copy number variants[28] (CNV) in the UKB training population. We used the --glm Firth-fallback option to apply an additive-effect model across all sites. Quantitative trait values were inverse-transformed by rank to a normal distribution. The following covariates were used: age, sex, the first four genetic principal components, and, for sites present on both of UK Biobank's genotyping arrays, the array which was used for each sample.

Prior to public release, genotyped sites and samples were subject to rigorous quality control by UK Biobank[29]. In brief, markers were subject to outlier-based filtration on effects due to batch, plate, sex, array, as well as discordance across control replicates. Samples with excess heterozygosity (thresholded percentages varying by ancestry) and missingness (> 5%) were excluded from the data release. Prior to use in downstream methods, we performed additional variant quality control including more stringent filters on missingness (> 1%), array effects, array probe efficacy, and gross departures ($p < 10^{-7}$) from Hardy-Weinberg Equilibrium[30].

Binary disease outcomes were defined from UK Biobank resources using a previously described method which combines self-reported questionnaire data and diagnostic codes from hospital inpatient data[30]. Additional traits like biomarkers, environmental variables, and

self-reported questionnaire data like health outcomes and lifestyle measures, were collected from fields curated by UK Biobank. Multiple observations were processed by taking the median of quantitative values, or by defining an individual as a binary case if any recorded instance met the trait's defining criteria. Summary statistics and log files from all GWAS described here are publicly available on the Global Biobank Engine[16] (biobankengine.stanford.edu). In this work, we highlight results for body mass index (GBE ID: INI21001), myocardial infarction (HC326), and gout (HC328).

Risk modeling using Decomposition of Genetic Associations (DeGAs):

Given GWAS summary statistics computed using the above methods, we performed a matrix Decomposition of Genetic Associations (DeGAs), as previously described[15]. First, a sparse matrix of genetic effects (W) is assembled using effect size estimates (or z-statistics) between n=1,905 traits with at least 2 nominally significant ($p < 0.01$; see Supplement for additional cutoffs) associations with m=454,565 common variants (MAF > 0.01%) in approximate linkage disequilibrium with one another (computed with PLINK[31] v1.90b4.4 [21 May 2017] --indep 50 5 2). After filtration, input statistics were standardized to zero mean and unit variance within each traitso as to weight them equally relative to one another.

We then performed a truncated singular value decomposition (TSVD) on W using the scikit-learn TruncatedSVD package[32] to identify k=500 trait-related genetic components. This factorization results in three matrices whose product approximates W: a trait singular matrix U ($n \times k$), variant singular matrix V ($m \times k$), and eigenvalue matrix S ($k \times k$) (Figure 1a). U, S, and V were then used to compute polygenic risk scores (PRS) for each DeGAs component (cPRS), as well as DeGAs risk scores (dPRS) for each trait. For the ith DeGAs component c_i, the weights of cPRS_i can be written as

$$cPRS_i = S_{i,*}V^T$$

where $S_{i,*}$ denotes the ith row of S. Similarly, dPRS for the jth trait can be written as

$$dPRS_j = U_{j,*}SV^T$$

or in terms of the cPRS as

$$dPRS_j = \sum_i U_{j,i}\, cPRS_i$$

We further relate individuals to traits via components using a measure we call the DeGAs risk profile (dRP). The profile is a vector over the k DeGAs components, where the value for the ith component is proportional to the below:

$$dRP_{j,i} \sim \max(0,\, dPRS_j \times cPRS_i)$$

The vector is then normalized such that its nonzero entries sum to one. Since we are interested in estimating the contribution of each component to an individual's overall genetic risk, we only

consider component scores which have the same sign as overall risk score (hence the max operator). This gives risk profiles consisting of driving components for high risk individuals with positive dPRS, and protective components for low risk individuals with negative dPRS. For interpretability, the population distribution of dPRS for each trait is scaled to zero mean and unit variance.

Computing polygenic risk scores:

As baseline for dPRS, we compute polygenic risk scores with a clumping and thresholding approach using the same set of variants as in DeGAs. Since these variants are already filtered based on LD independence, this approach reduces to the same thresholding performed as in the DeGAs input filtration. In terms of the notation from that section, the weights for these PRS are rows the input matrix *W*. These PRS are computed with PLINK v1.90b4.4 [21 May 2017] using the --score flag, with the following modifiers: sum center double-dosage. These correspond to the assumptions that variants make additive contributions across sites, the mean distribution of risk is taken to be zero, and that the effect alleles have additive effects; more importantly, these are the same assumptions used by the input GWAS.

In a similar fashion, polygenic scores (cPRS) for all DeGAs components were computed with PLINK2 v2.00a2 (2 Apr 2019) --score center cols=scoresums. These modifiers correspond to the same assumptions as in the PRS: that genetic effects are additive across sites (this is the default genotype model for --score), each component is zero-centered, and alleles make additive contributions. Given population-wide estimates of cPRS for every component, we compute dPRS and DeGAs risk profiles for each trait using the above formulas.

Model validation:

To select DeGAs hyperparameters — the input *p*-value filter, and whether to use GWAS betas or z-statistics as weights — we performed a grid search over a range of *p*-values for each statistic. DeGAs performance was assessed using the average correlation of the resulting set of dPRS models and their respective traits. For all traits used in the decomposition, we computed Spearman *r* (rank correlation) between dPRS and trait residuals in the UKB validation population. These residual traits are the result of regressing out the following covariates from each trait: age, sex, and four genetic principal components. We find optimal performance in the validation (10% white British) population using centered z-statistics, a *p*-value cutoff of 0.01, and 300 components (**Supplementary Figure 1**).

For this final DeGAs instance, we present several measures of the accuracy of each polygenic score — dPRS (**Figure 2**) or traditional PRS using the same input data (**Supplementary Figure 2**) within each study population. For each score and population, we estimated disease prevalence and mean quantitative trait values at various population risk strata. We further assessed the scores' ability to predict quantitative trait values and perform binary classification on disease status. For quantitative traits, we report Pearson's *r* (rank correlation) between score and trait residuals, as defined above. For binary traits, we report the area under the receiver

operating curve (AUROC/AUC) with dPRS as the classifying score, both alone and in a joint model with the covariates used for genome-wide association above. As baseline, we also report AUC for a covariate-only model.

Classifying genetic risk profiles from DeGAs components:

To assess the utility of dPRS in identifying within-trait diversity, we attempted to cluster high risk individuals whose DeGAs risk profiles deviated from the average for the high risk group within each trait. We used the Mahalanobis criterion to select outliers in the test cohort whose *z*-scored distance from the cohort population exceeded 2, and intersected this list of individuals with the top 10% of dPRS scorers. We then performed a *k*-means clustering of these "high risk outliers" using their DeGAs risk profiles. The number of clusters *k* was determined iteratively using proportional reduction in error (PRE). This method iteratively increments *k* and recomputes the clustering, arriving at the final *k* when the reduction in within-cluster error — that is, the sum of Mahalanobis distances for all samples in the cluster, across all clusters — fails to exceed 20% on the *k*+1st clustering. We then evaluate which components drive risk in each cluster by computing the median risk profile for the group.

Acknowledgements:

This research has been conducted using the UK Biobank Resource under Application Number 24983, "Generating effective therapeutic hypotheses from genomic and hospital linkage data" (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf). Based on the information provided in Protocol 44532 the Stanford IRB has determined that the research does not involve human subjects as defined in 45 CFR 46.102(f) or 21 CFR 50.3(g). All participants of UK Biobank provided written informed consent (more information is available at https://www.ukbiobank.ac.uk/2018/02/gdpr/). We thank all the participants in the UK Biobank study. This work was supported by National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under awards R01HG010140 (M.A.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. M.A.R. is supported by Stanford University and a National Institute of Health center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080).

Author contributions:

M.A.R. conceived and designed the study. M.A. and M.A.R. carried out statistical and computational analyses. M.A., Y.T., G.V., J.M.J. and M.A.R. carried out quality control of the data. R.T. and T.H. aided in statistical design and conception. The manuscript was written by M.A. and M.A.R.; and revised by all the co-authors. All co-authors have approved of the final version of the manuscript.

Citations:

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1789–1858 (2018).
2. Fritsche, L. G. *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. doi:10.1101/205021
3. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
4. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
5. Läll, K., Mägi, R., Morris, A., Metspalu, A. & Fischer, K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet. Med.* **19**, 322 (2016).
6. Marquez-Luna, C., Loh, P.-R., Price, A. L., South Asian Type 2 Diabetes (SAT2D) Consortium & The SIGMA Type 2 Diabetes Consortium. Multi-ethnic polygenic risk scores improve risk prediction in diverse populations. doi:10.1101/051458
7. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
8. Khera, A. V. *et al.* Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation* **139**, 1593–1602 (2019).
9. Belsky, D. W. *et al.* Development and evaluation of a genetic risk score for obesity. *Biodemography Soc. Biol.* **59**, 85–100 (2013).
10. Khera, A. V. *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to

Adulthood. *Cell* **177**, 587–596.e9 (2019).

11. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
12. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
13. Qian, J. *et al.* A Fast and Flexible Algorithm for Solving the Lasso in Large-scale and Ultrahigh-dimensional Problems. *bioRxiv* 630079 (2019). doi:10.1101/630079
14. McCarthy, M. I. Painting a new picture of personalised medicine for diabetes. *Diabetologia* **60**, 793–799 (2017).
15. Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight novel adipocyte biology. doi:10.1101/442715
16. McInnes, G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty999
17. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
18. Frayling, T. M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
19. Liu, Y. *et al.* Cloning of two candidate tumor suppressor genes within a 10 kb region on chromosome 13q14, frequently deleted in chronic lymphocytic leukemia. *Oncogene* **15**, 2463–2473 (1997).
20. Paquette, M., Bernard, S. & Baass, A. SLC22A3 is associated with lipoprotein (a) concentration and cardiovascular disease in familial hypercholesterolemia. *Clin. Biochem.* **66**, 44–48 (2019).
21. Kuo, C.-F. *et al.* Familial aggregation of gout and relative genetic and environmental

contributions: a nationwide population study in Taiwan. *Ann. Rheum. Dis.* **74**, 369–374 (2015).

22. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat. Genet.* **40**, 437–442 (2008).

23. Dankers, A. C. A. *et al.* Hyperuricemia influences tryptophan metabolism via inhibition of multidrug resistance protein 4 (MRP4) and breast cancer resistance protein (BCRP). *Biochim. Biophys. Acta* **1832**, 1715–1722 (2013).

24. Köttgen, A. *et al.* Multiple loci associated with indices of renal function and chronic kidney disease. *Nat. Genet.* **41**, 712–717 (2009).

25. Mack, S. *et al.* A genome-wide association meta-analysis on lipoprotein (a) concentrations adjusted for apolipoprotein (a) isoforms. *J. Lipid Res.* **58**, 1834–1844 (2017).

26. Ryu, W.-S., Kim, C. K., Kim, B. J. & Lee, S.-H. Serum uric acid levels and cerebral microbleeds in patients with acute ischemic stroke. *PLoS One* **8**, e55210 (2013).

27. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

28. Aguirre, M., Rivas, M. A. & Priest, J. Phenome-wide Burden of Copy-Number Variation in the UK Biobank. *Am. J. Hum. Genet.* **105**, 373–383 (2019).

29. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

30. DeBoever, C. *et al.* Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).

31. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

32. Halko, N., Martinsson, P. G. & Tropp, J. A. Finding Structure with Randomness:

Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review* **53**, 217–288 (2011).

Supplementary Information:

Table S1: List of phenotypes and dPRS performance metrics across population groupings.

Table S2: Test performance of dPRS across hyperparameters.

Table S3: Test performance of prune- and threshold PRS.

Figure S1: DeGAs hyperparameter optimization and train/validation set performance.

Figure S2: Test set PRS performance for BMI, MI, and gout.

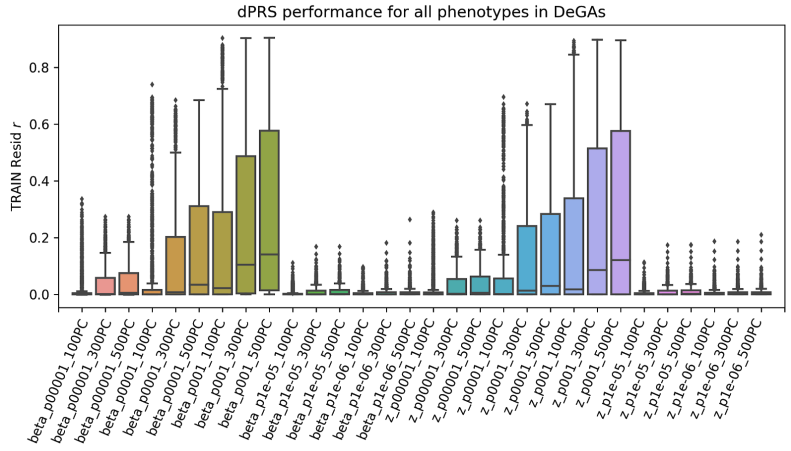
Figure S3: dPRS performance for BMI, MI, and gout in UK Biobank non-British white individuals.

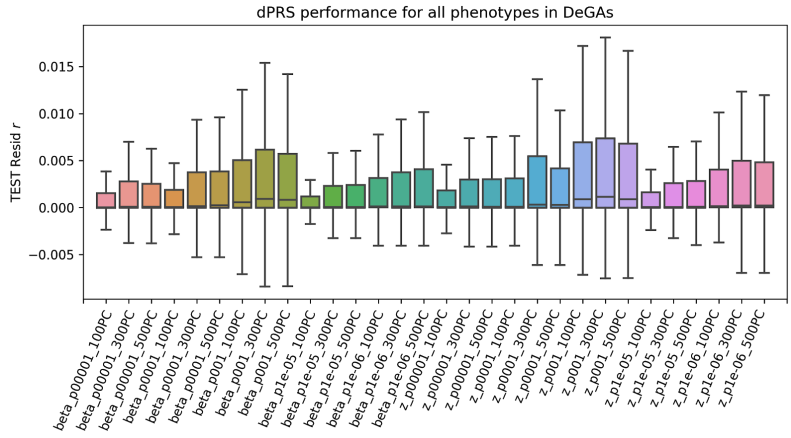
Figure S4: Phenotype contribution scores for each DeGAs component

Figure S5: Gene contribution scores for each DeGAs component

Note: All tables, and Figures S4 and S5 are available as separate documents.

Figure S1: Hyperparameter optimization





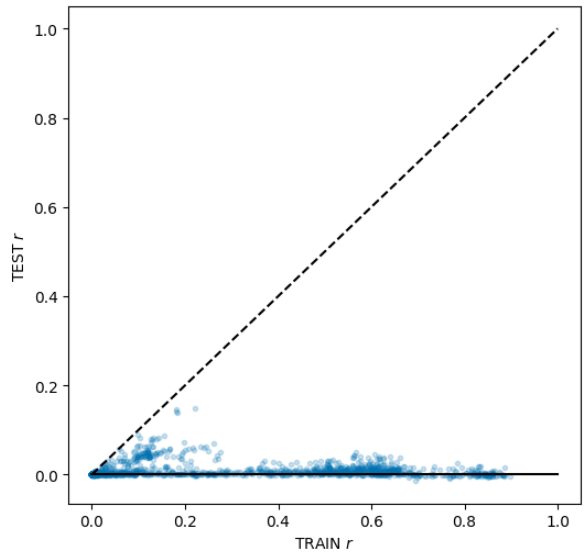


Figure S2: PRS Performance.

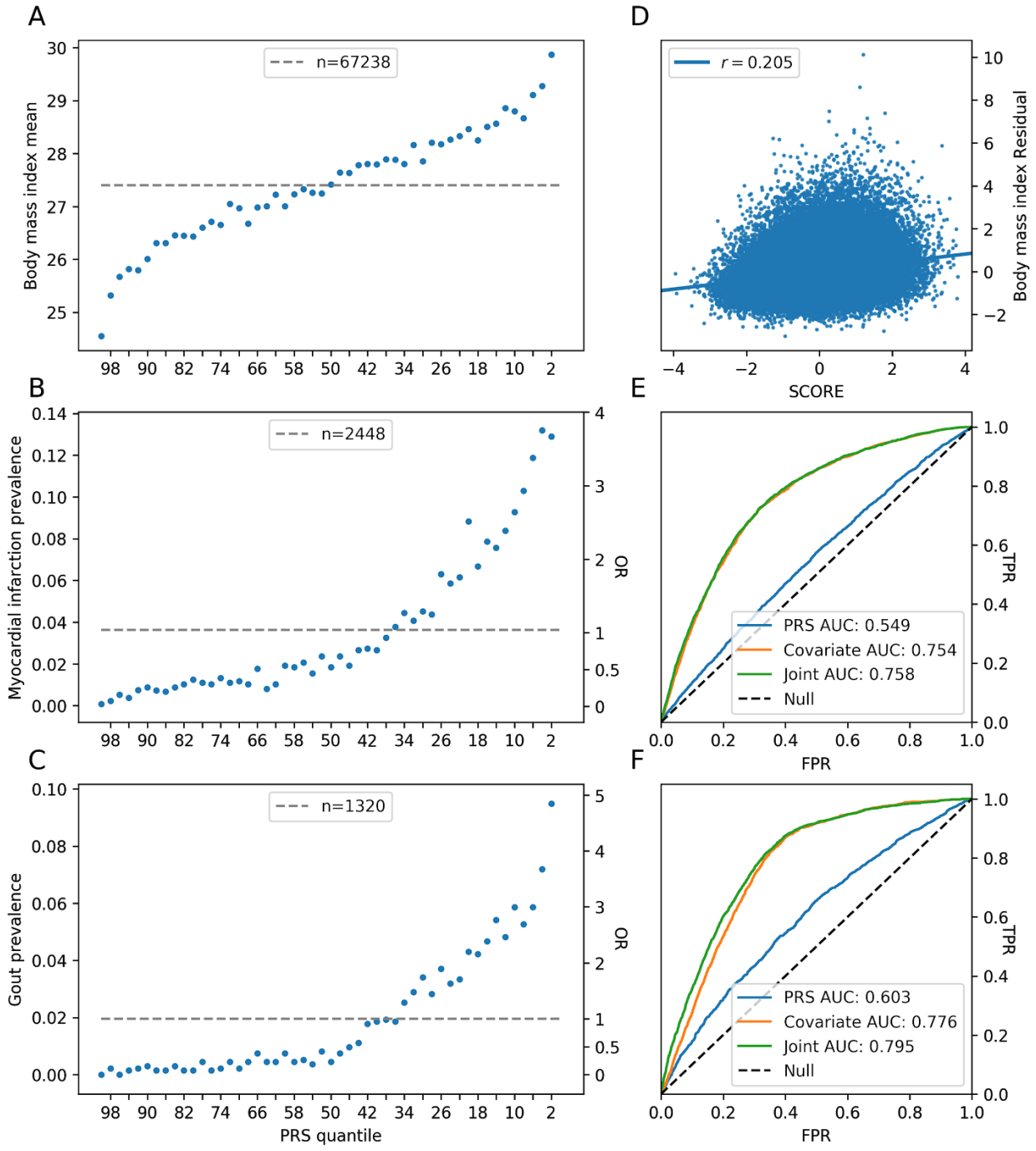


Figure S3: dPRS in NBW

Figure S4: Trait contribution scores across 300 components
Figure S3: Gene contribution scores across 300 components